Vom Promotionsausschuss des Fachbereichs Psychologie der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau (Campus Landau) zur Verleihung des akademischen Grades Doktor der Philosophie (Dr. phil.) genehmigte Dissertation

The Effects of Design Choices in Ambulatory Assessment Studies on Participant Burden, Data Quantity, and Data Quality

vorgelegt von

Kilian Hasselhorn, M.Sc.

Vorsitzende des Promotionsausschusses:

Prof. Dr. Tanja Lischetzke
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

Berichterstatterinnen:

Prof. Dr. Tanja Lischetzke
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

Prof. Dr. Tanja Könen
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

Vorsitzender der Promotionskommission:

Prof. Dr. Benjamin Hilbig
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

Tag der Disputation: 13. Juli 2023

Table of Contents

List of Figures

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 participants' perceived burden, data quantity (i.e., compliance with the AA protocol), and data quality is still surprisingly restricted. The aims of this dissertation were to experimentally manipulate aspects of an AA study's sampling strategy - sampling frequency (Study 1) and questionnaire length (Study 2) - and to investigate their impact on perceived burden, data quantity, and aspects of data quality in three papers. In Study 1, students (n = 313) received either 3 or 9 questionnaires per day for the first 7 days of the study. In Study 2, students (n = 282) received either a 33- or 82-item questionnaire 3 times a day for 14 days.

Paper 1 described that a higher sampling frequency (Study 1) led to a higher perceived participant burden, but did not affect other aspects of data quantity and quality. Furthermore, a longer questionnaire (Study 2) did not affect perceived participant burden or data quantity, but did lead to a lower within-person variability, and a lower within-person relationship between time-varying variables. Paper 2 investigated the effects of the sampling frequency (Study 1) on careless responding by identifying careless responding indices that could be applied to AA data and by extending the multilevel latent class analysis model to a multigroup multilevel latent class analysis model. Results indicated that a higher sampling frequency did not affect careless responding. Paper 3 investigated the effects of questionnaire length (Study 2) on (the relative impact of) response styles by extending the item response tree (IRTree) modeling approach to a multilevel data structure. Results indicated that a longer questionnaire led to a greater relative impact of RS.

Although further validation of the results is essential, I hope that future researchers will integrate the results of this dissertation when designing an AA study.

Zusammenfassung [Abstract]

Ambulatory Assessment (AA) wird zu einer immer beliebteren Forschungsmethode in der Psychologie und in den Lebenswissenschaften. Trotzdem ist das Wissen inwiefern Designentscheidungen, wie z.B. die Länge des Fragebogens (d.h. die Anzahl der Items pro Fragebogen), sich auf die von den Studienteilnehmer*Innen wahrgenommene Belastung, die Datenmenge (d.h. die Einhaltung des AA-Protokolls) und die Datenqualität auswirken, erstaunlich limitiert. Das Ziel dieser Dissertation war es einzelne Aspekte des Studiendesigns von AA-Studien – die Anzahl an Fragebögen pro Tag (Studie 1) und die Länge des Fragebogens (Studie 2) – experimentell zu manipulieren und ihre Auswirkungen auf die wahrgenommene Belastung, die und die Datenqualität in drei Artikeln zu untersuchen. In Studie 1 erhielten Studierende (n = 313) entweder 3 oder 9 Fragebögen pro Tag in den ersten 7 Tagen der Studie. In Studie 2 erhielten Studierende (n = 282) entweder einen 33- oder 82-Items pro Fragebogen drei Mal am Tag für 14 Tage.

Artikel 1 beschreibt, dass eine höhere Anzahl an Fragebögen pro Tag (Studie 1) nur die von den Studienteilnehmer*Innen wahrgenommene Belastung erhöht, aber nicht die Datenmenge, die intraindividuellen Variabilität von Zustands-Extraversion und momentaner angenehmer-unangenehmer Stimmung, oder die intraindividuelle Beziehung zwischen Zustands-Extraversion und momentaner angenehmer-unangenehmer Stimmung beeinflusst. Artikel 2 untersuchte die Auswirkungen von einer höheren Anzahl an Fragebögen pro Tag auf unachtsames Antwortverhalten, indem Indizes für unachtsames Antwortverhalten identifiziert wurden, die auf AA-Daten angewandt werden konnten, und das Modell der Multi-Level Latenten Klassenanalyse auf ein Modell der Multigruppen Multi-Level Latenten Klassenanalyse erweitert wurde. Die Ergebnisse zeigten, dass eine höhere Anzahl an Fragebögen pro Tag keinen Einfluss auf unachtsames Antwortverhalten hatte. Artikel 3 untersuchte den Einfluss von der Länge des Fragebogens auf Antworttendenzen beim

Beantworten der Fragebögen (Studie 2), indem der Ansatz der Item-Response-Tree-Modellierung (IRTree) auf eine Multi-Level Datenstruktur erweitert wurde. Die Ergebnisse zeigten, dass ein längerer Fragebogen zu einem größeren relativen Einfluss von Antworttendenzen führte.

Obwohl eine weitere Validierung der Ergebnisse unerlässlich ist, hoffe ich, dass zukünftige Forscher*Innen die Ergebnisse dieser Dissertation bei der Konzeption weiterer AA-Studien berücksichtigen werden.

Danksagung [Acknowledgments]

Es gibt eine Reihe von Menschen, die mich bei der Arbeit an dieser Dissertation unterstützt haben. Zuallererst bedanke ich mich bei Tanja Lischetzke. Danke für deine Bereitschaft meine Arbeit zu betreuen, deine Freundlichkeit, für die Hilfestellungen, deine Führung, und für dein ehrliches Feedback, wenn ich es brauchte. Danke auch vor allem für deine Großzügigkeit bezüglich der guten finanziellen Unterstützung. Ein weiterer Dank geht an Thorsten Meiser, für die Hilfestellungen und Ratschläge im Laufe meiner Dissertation. Ebenfalls danke ich Tanja Könen und Benjamin Hilbig herzlich dafür, dass ihr euch die Zeit nehmt, meine Dissertation zu lesen und zu begutachten.

Die Arbeit wurde ermöglicht durch ein Promotionsstipendium von der Research Training Group „Statistical Modeling in Psychology“ (SMiP), welches durch die Deutsche Forschungsgemeinschaft (DFG) finanziert wurde.

Ferner danke ich meinen Kollegen*Innen: Danke Charlotte Ottenstein, für deine offenen Ohren und deine hilfsbereite Art. Danke Christine Reither und Katharina Reich für die Unterstützung im Hintergrund. Danke Nikoletta Symeonidou, für deine bereichernde Art, und deine ehrlichen und aufmunternden Worte. Danke für deine Einblicke und deine Bewusstseinsschaffung für viele andere Bereiche außerhalb der Wissenschaft.

Allen meinen lieben Freunden danke ich für die Aufmerksamkeit, Ruhe und Ablenkung, womit sie mir stets zur Seite standen und mich immer wieder aufgemuntert haben.

Schließlich möchte ich meinen Eltern und meinen Geschwistern danken: Danke Mama und Papa, dass ihr mir das Studium der Psychologie ermöglicht habt und für eure anhaltende Unterstützung und Liebe aus der Ferne. Danke liebe Geschwister, für eure Freundschaft, euer Mitgefühl und eure Ermutigungen in meinem Leben.

An letzter, doch eigentlich an erste Stelle möchte ich meiner Verlobten Anna-Lena danken. Deine Geduld, deine Ehrlichkeit, deine Aufmunterungen, und deine Liebe bereichern und verschönern das Leben jeden Tag für mich. Danke, dass du Teil meines Lebens bist.

1 Introduction

Psychologists have long been interested in the assessment of dynamic processes within individuals. Wanting to understand these dynamic processes, researchers' interest in conducting ambulatory assessment (AA) studies has grown rapidly over the past decade (Hamaker & Wichers, 2017; Jaso et al., 2021). AA (also called experience sampling, ecological momentary assessment, or daily diary) is a data collection method that can be used to assess individuals' daily life experiences such as ongoing behaviors, experiences, physiology, and environmental aspects of people in naturalistic and unconstrained settings (Bolger & Laurenceau, 2013; Fahrenberg, 2006; Larson & Csikszentmihalyi, 1983). By using AA, researchers can study within-person dynamic processes (e.g., within-person relationships between time-varying variables) as well as individual differences in these within-person dynamics (Hamaker & Wichers, 2017). Furthermore, AA provides reduced recall bias and high ecological validity through real-time (or near real-time) assessments in individuals' natural environment (Mehl & Conner, 2012; Trull & Ebner-Priemer, 2014).

When designing an AA study, after deciding on the research question and the target sample, researchers must make multiple decisions about the design of the sampling strategy to strike a balance between being able to collect rich information, not overburdening participants (Carpenter et al., 2016), and not compromising aspects of AA data (e.g., data quantity and data quality; Arslan et al., 2020; May et al., 2018). These decisions include the types of reports to be included (e.g., event-triggered, time-based), the number of days to survey individuals, the number of assessments to be administered each day (sampling frequency), and the number of items to administer per questionnaire (questionnaire length). For more detailed information on design considerations and data collection methods, see, for example Mehl and Conner (2012).

In light of the very large number of studies that have used AA in the last decade in a variety of research areas (Jaso et al., 2021), the questions arise about the extent to which, and

under what conditions, different design features (of the sampling strategy) affect participants' perceived burden and aspects of AA data (e.g., data quantity and data quality). However, the methodological research on the effects of design features is surprisingly restricted, meaning that researchers cannot base their design of the sampling strategy on empirical evidence (Eisele et al., 2020; Himmelstein et al., 2019). Researchers have already begun to investigate the effects of design features in meta-analytic or pooled data analyses (e.g., Ottenstein & Werner, 2021; Podsakoff et al., 2019; Vachon et al., 2019), but these analyses of between-study differences cannot rule out the effects of third variables. To increase the internal validity of causal inferences, experimental designs are well suited to investigate the effects of different design features on participants' perceived burden and aspects of AA data (e.g., data quantity and data quality). However, the relatively scarce methodological research has failed to examine several outcome variables of interest (e.g., within-person variability or response styles) or has produced inconsistent results (e.g., Conner & Reid, 2012; Eisele et al., 2020; Stone et al., 2003). Clearly, more research is needed to elucidate the effects of design features on participants' perceived burden and aspects of AA data (e.g., data quantity and data quality).

The present dissertation seeks to enhance the understanding of the extent to which, and under what conditions two central design features – the sampling frequency and the questionnaire length – may affect participants' perceived burden and aspects of AA data (e.g., data quantity and data quality). To quantify the effects of these design features, this dissertation investigates the effects of design features on participants' perceived burden, compliance, within-person variance (also called within-person variability), within-person relationship (between time-varying variables), careless responding, and response styles as outcome variables.

In the following, I will briefly describe the importance of each of the outcome variables and review the research that has been conducted on the effects of sampling frequency and questionnaire length (as design choices) on each of the outcome variables, in order to deduce relevant research questions about the effects of these design choices. Subsequently, I will outline the three manuscripts that form this dissertation and specify how they address the outcome variables.

1.1 Outcome Variables

1.1.1 Participant Burden

For participants in an AA study, it is assumed that a higher sampling frequency or a longer questionnaire (vs. a lower sampling frequency or a shorter questionnaire) will increase the participants' perceived burden (e.g., Moskowitz et al., 2009; Ono et al., 2019; Piasecki et al., 2007; Roekel et al., 2019; Santangelo et al., 2013; Wen et al., 2017). In addition, many researchers have conceptualized increased perceived burden as a psychological process that is expected to result in a reduction in the quantity and quality of AA data (e.g., Eisele et al., 2020; Fuller-Tyszkiewicz et al., 2013; Moskowitz et al., 2009; Piasecki et al., 2007; Santangelo et al., 2013; Wrzus & Neubauer, 2022). However, there are only two studies that have experimentally manipulated the sampling frequency or the questionnaire length and analyzed perceived burden as an outcome variable (Eisele et al., 2020; Stone et al., 2003). The study by Stone et al. (2003) experimentally manipulated sampling frequency over a two-week period and found that the groups with a higher sampling frequency perceived higher burden. By contrast, Eisele et al. (2020), analyzed the effects of sampling frequency and questionnaire length over a two-week period, and found that the higher sampling frequency groups did not perceive a higher burden, but that the longer questionnaire group did perceive a higher burden (vs. the short questionnaire group). Given these different and limited amount of results on perceived burden as an outcome variable, it remains unclear whether a higher sampling

frequency or a longer questionnaire (vs. a lower sampling frequency or a shorter questionnaire) increases participants' perceived burden. Taken together, my first research question was to test whether a higher sampling frequency (RQ1A) or a longer questionnaire (RQ1B) leads to a higher perceived burden.

1.1.2 Compliance

In AA studies, compliance is considered a particularly important outcome variable. Specifically, high compliance is necessary to obtain a representative picture of individuals' day-to-day experiences and behaviors (Stone et al., 2003), and low compliance can lead to lower statistical power (Maxwell et al., 2008) and biased inferences about aggregated person-level data (Courvoisier et al., 2012). A higher sampling frequency or a longer questionnaire (vs. a lower sampling frequency or a shorter questionnaire) can be assumed to compromise compliance (i.e., data quantity) because participants may intentionally reduce the burden by not completing a particular measurement occasion (Stone et al., 2003; Vachon et al., 2019).

Surprisingly, studies that experimentally manipulated sampling frequency (with sampling frequencies ranging between 1 and 20 questionnaires per day across studies) found no difference in compliance between experimental groups (Conner & Reid, 2012; Eisele et al., 2020; McCarthy et al., 2015; Stone et al., 2003; Walsh & Brinker, 2016). Results from meta-analytic or pooled data analyses that analyzed the effect of sampling frequency on compliance were somewhat inconsistent: The majority of studies found no support for the idea that higher sampling frequencies are related to lower compliance rates (Jones et al., 2019; Morren et al., 2009; Ono et al., 2019; Ottenstein & Werner, 2021; Soyster et al., 2019; Wrzus & Neubauer, 2022), with the exception of the study by Vachon et al. (2019), which found lower compliance rates in studies that administered a greater number of questionnaires per day.

Regarding questionnaire length, the only study I know of that experimentally manipulated questionnaire length found that longer questionnaires (vs. shorter questionnaires) led to lower compliance rates (Eisele et al., 2020). Meta-analytic or pooled data analyses found no support for the idea that longer questionnaires are related to a lower compliance rate (Jones et al., 2019; Ono et al., 2019; Rintala et al., 2019; Soyster et al., 2019; Vachon et al., 2019), with the exception of Morren et al. (2009), who found that a longer questionnaire leads to lower compliance.

Given these somewhat inconsistent results on the effects of sampling frequency and the limited amount of experimental research on the effects of questionnaire length on compliance, more research is needed to further scrutinize these effects. Therefore, my second research question was to investigate the effects of a higher sampling frequency (RQ2A) and a longer questionnaire (RQ2B) on compliance.

1.1.3 Within-Person Variability

In AA studies, within-person variability is a prerequisite for studying dynamic processes within individuals (Heck & Thomas, 2015; Hox, 2002; Raudenbush & Bryk, 2002), and researchers have warned that when within-person variability is low, research on within-person relationships between time-varying variables should not be conducted (Podsakoff et al., 2019; Rosen et al., 2016; Sonnentag et al., 2008; Trougakos et al., 2008). With respect to the psychological processes, Podsakoff et al. (2019) argued that a higher sampling frequency (vs. a lower sampling frequency) might lead to a higher within-person variability in time-varying variables, because participants may become more aware of differences between the current state and previous states. However, other researchers argued that higher sampling frequencies or longer questionnaires (vs. lower sampling frequencies or shorter questionnaires) increase participant fatigue (e.g., Beal, 2015), which leads to more heuristic

and less nuanced responses, thereby reducing the degree of within-person variability (Fuller-Tyszkiewicz et al., 2013; Podsakoff et al., 2019).

The only empirical evidence that analyzed the effects of sampling frequency on within-person variability comes from a meta-analysis by Podsakoff et al. (2019), which found that higher sampling frequency, but not the study duration (i.e., the number of days participants were surveyed), predicted larger within-person variability. However, there is some indirect evidence from three AA studies that analyzed the change of the degree of within-person variability over the course of an AA study: These studies found that within-person variability declined over the course of an AA study (Eisele et al., 2023; Fuller-Tyszkiewicz et al., 2013; Vachon et al., 2016). Moreover, two of these studies (Fuller-Tyszkiewicz et al., 2013; Vachon et al., 2016) did not use an experimental manipulation, thus third variables could have driven the effects. The only study that used an experimental manipulation was the study by Eisele et al. (2023), who additionally analyzed the effects of sampling frequency and questionnaire length on the change in the degree of within-person variability over the course of an AA study. However, they found that the effects of sampling frequency and questionnaire length were not consistent across different substantive constructs. Taken together, the results of these studies were somewhat inconsistent. Therefore, my third research question was to investigate the effects of sampling frequency (RQ3A) and questionnaire length (RQ3B) on within-person variability.

1.1.4 Within-Person Relationship Between Time-Varying Variables

A large body of research on dynamic processes within individuals tends to focus on within-person relationships between time-varying variables (Liu et al., 2019; May et al., 2018; Sitzmann & Yeo, 2013). Fuller-Tyszkiewicz et al. (2013) argued that when within-person variability is reduced, the strength of within-person relationships between time-varying variables may also be reduced. Empirically, however they did not find that a reduction in the

within-person variability translated into a smaller within-person relationship between time-varying variables as a function of the number of days in the study (Fuller-Tyszkiewicz et al., 2013).

To my knowledge, no empirical study has investigated the effects of sampling frequency or questionnaire length on within-person relationships between time-varying variables. However, there is a need for (more) experimental research on the effects of sampling frequency or questionnaire length on within-person relationships between time-varying variables. Therefore, my fourth research question was to investigate the effects of sampling frequency (RQ4A) and a questionnaire length (RQ4B) on within-person relationships between time-varying variables.

1.1.5 Careless Responding

Careless responding (also called insufficient effort responding) is defined as a response behavior that is characterized by responding to items without sufficient regard to the item content (Huang et al., 2012; Meade & Craig, 2012). Careless responding is one potential threat to the data quality in AA studies, because careless responding can compromise the psychometric properties of measurement instruments and potentially bias the correlations between substantive measures (Goldammer et al., 2020; Huang et al., 2015; McGrath et al., 2010). Therefore, it is crucial to identify careless responses (or careless responders) in order to maximize the reliability, power, and validity of the results obtained using AA data (Jaso et al., 2021).

With respect to the psychological processes, Jones et al. (2019) hypothesized that participants may provide data that is not of sufficient quality, such as responding carelessly, in order to reduce the burden of participating in an AA study. Since sampling frequency is assumed to increase perceived burden (Stone et al., 2003), a higher sampling frequency in an AA study may result in more careless responding. To my knowledge, the study by Eisele et

al. (2020) is the only empirical study that has investigated the effects of design features (sampling frequency and questionnaire length) on careless responding in AA. They found that sampling frequency was not associated with careless responding, while a longer (vs. a shorter) questionnaire increased careless responding. However, the authors relied exclusively on self-report measures of careless responding, which are dependent on the ability and willingness of participants to expend effort responding in the study. Therefore, it remains an open question whether design features (e.g., sampling frequency) influence careless responding, which has been identified through the use of unobtrusive indices that have been used in cross-sectional research (e.g., Goldammer et al., 2020; Meade & Craig, 2012).

In cross-sectional research, careless responding is typically identified by using either a latent class analysis (LCA) approach (e.g., Meade & Craig, 2012) or a multiple hurdle approach (e.g., Curran, 2016). Both approaches use a variety of (obtrusive and unobtrusive) indices for detecting and removing careless responders, consequently improving data quality (Curran, 2016; Meade & Craig, 2012). In the multiple hurdle approach, a cut-off score is defined for each index, and a participant is classified as a careless responder if they fail to pass all of the "hurdles". In the LCA approach, the careless responding indices serve as individually observed indicators that are grouped into latent classes of individuals with different patterns of careless responding. Advantages of the LCA approach (compared to the multiple hurdle approach) are that researchers can identify careless responders without having to define cut-off scores for careless responding indices and potentially distinguish between different types of careless responders (e.g., invariable responding vs. inconsistent responding). However, research on careless responding has focused almost exclusively on cross-sectional data, and no guidelines or best practices exist for identifying careless responses or careless responders in AA (Jaso et al., 2021; van Berkel et al., 2018). To my knowledge, no study to date has used the LCA approach to examine careless responding in AA data. Thus, this

dissertation proposes that careless responding indices can be operationalized on the occasions level (by translating previously defined indices into AA data) and that multilevel LCA (ML-LCA) can be used to model careless responding in AA data (measurement occasions nested in persons). Specifically, I expected to identify latent profiles of momentary careless responding at the occasion level (Level 1) and latent classes of individuals at the person level (Level 2) who differed in their use of careless responding over time (in each group separately). Additionally, I extended the ML-LCA approach to a multigroup ML-LCA approach in order to investigate the effects of sampling frequency on careless responding (across two experimental groups) in AA.

Taken together, my fifth research question (RQ5) was to identify latent profiles of momentary careless responding at the occasion level (Level 1) and to differentiate between latent classes of individuals at the person level (Level 2) who differ in their use of careless responding over time using (multigroup) ML-LCA in order to test whether a higher sampling frequency leads to more careless responding. Specifically, I expected to identify the same number of latent profiles of momentary careless responding at the occasion level (Level 1) and latent classes of individuals at the person level (Level 2) who differed in their use of careless responding over time using both sampling frequency groups. Furthermore, I expected that a high sampling frequency (vs. a low sampling frequency) would lead to a higher proportion of participants that are assigned to a careless responding class. As a starting point, this dissertation focuses on the effects of sampling frequency on careless responding and does not focus on the effects of questionnaire length on careless responding.

1.1.6 Response Styles

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). Similar to

careless responding, RS are a potential threat to the data quality in AA studies, because they can introduce systematic measurement error. Hence, RS can, for instance, distort relationships between measured variables (Böckenholt & Meiser, 2017; Bolt & Newton, 2011; Park & Wu, 2019) and threaten construct and predictive validity (Baumgartner & Steenkamp, 2001; van Herk et al., 2004). More detailed information on the adverse effects of RS has been discussed elsewhere (e.g., Ames & Myers, 2021; Moors, 2012; Weijters et al., 2010). It is therefore important to account for RS. Doing so can debias estimates of the substantive trait and reduce the bias that is associated with RS (Adams et al., 2019; Henninger & Meiser, 2020b). With respect to the psychological processes, RS may be influenced by the questionnaire length as a design feature. Specifically, a longer questionnaire (vs. a shorter) may impose a higher cognitive load or a higher perceived burden on participants when attempting to complete such a questionnaire. As a consequence, participants may be more driven by heuristic processes like RS in an attempt to reduce cognitive load or perceived burden (Bolt & Johnson, 2009; Jones et al., 2019; Knowles & Condon, 1999).

There is a large variety of different methods for modeling and accounting for RS, such as count procedures, latent class analytic approaches, or Item Response Theory (IRT) models (Van Vaerenbergh & Thomas, 2013). In the last few decades, IRT models have become increasingly popular in the literature (Henninger & Meiser, 2020a). These models can be divided into two groups: extensions of traditional IRT models for ordinal responses and IRTree models. For an overview of the extensions of traditional IRT models, see Henninger and Meiser (2020a). 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 enable researchers to differentiate between processes that are based on the trait of interest and processes that are based on (a priori

specified) RS, such as extreme RS (ERS; Plieninger & Meiser, 2014; Zettler et al., 2016). ERS refers to an individual's tendency to endorse the extreme ends of the rating scale (e.g., Ames & Myers, 2021; Baumgartner & Steenkamp, 2001). Importantly, IRTree models are well suited to analyze and control RS effects in a confirmatory way (Böckenholt & Meiser, 2017). However, to my knowledge, no study to date has used IRTree models to examine RS in the context of AA. Therefore, this dissertation focused on extending the IRTree approach to model RS in the multilevel data structure (measurement occasions nested within persons) which is obtained by using AA, in order to investigate the effect of a design feature (questionnaire length) on RS.

Because no empirical study has investigated the effects of questionnaire length on the potential effects on RS in the context of AA my sixth research question (RQ6) is to investigate the effects of questionnaire length on RS in an AA study. Note that the chosen modeling approach in this dissertation does not analyze the effects of questionnaire length on RS in an AA study directly, but rather the effects of questionnaire length on the relative impact of RS in an AA study. As a starting point, this dissertation focuses on the effects of questionnaire length on RS and does not focus on the effects of sampling frequency on careless responding.

1.2 The Present Dissertation

This present dissertation examines the effects of design features, particularly sampling frequency and the questionnaire length, on several outcome variables. The outcome variables that have been identified as especially important in the context of AA are the participant burden (RQ1A and RQ1B), compliance (RQ2A and RQ2B), within-person variability (RQ3A and RQ3B), and the within-person relationship between time-varying variables (RQ4A and RQ4B), careless responding (RQ5) and RS (RQ6). Because this dissertation seeks to improve the understanding of the extent to which, and under what conditions these two central design

features might affect the identified outcome variables, this dissertation used experimental designs within AA studies. In doing so, it was possible to increase the internal validity of causal inferences compared to meta-analytic or pooled data analyses. Specifically, two AA studies were conducted, each manipulating one of the chosen design features (sampling frequency in Study 1 and questionnaire length in Study 2). All of the research questions within this dissertation will be addressed in three papers that build upon these two AA studies. The AA studies will be described in detail in the following papers.

Paper 1 (Hasselhorn et al., 2021) addresses the outcome variables participants' perceived burden, compliance, within-person variance, and the within-person relationship between time-varying variables (RQ1A to RQ4B). First, the relevance of the proposed outcome variables was described (in more detail than outlines so far). Second, multilevel regression analyses, *t* tests, and multigroup multilevel models were applied in order to assess whether these outcome variables were influenced by the design features sampling frequency and questionnaire length.

Paper 2 (Hasselhorn et al., in press) addresses the effects of the sampling frequency on careless responding (RQ5). First, indirect careless responding indices that could be applied to AA data are identified and entered into multilevel latent class analyses models. Second, the association between the sampling frequency and careless responding was investigated by extending the multilevel latent class analyses model to a multigroup multilevel latent class analyses model. Third, additional exploratory analyses were conducted to investigate the effects of situational factors (time-varying variables), and respondent-level factors (time-invariant variables) on careless responding in an AA study.

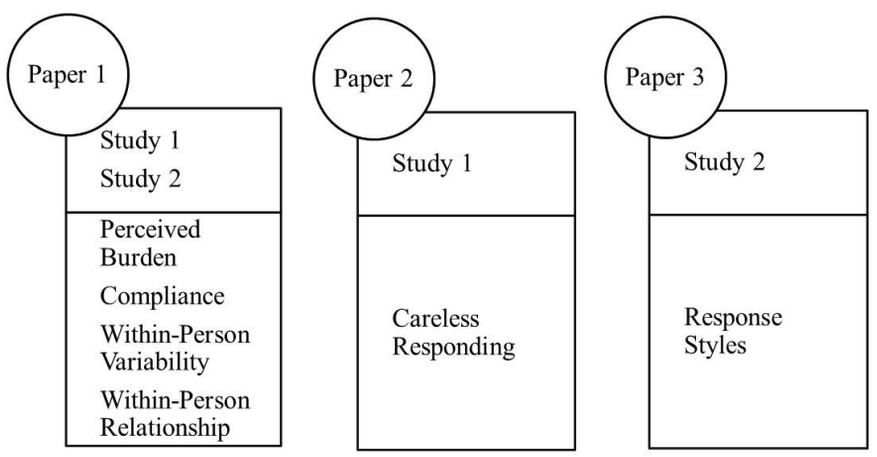
Paper 3 (Hasselhorn et al., 2023) addresses the effects of questionnaire length on (the relative impact of) RS (RQ6). First, the existing literature on the effects of questionnaire length in AA was described. Second, the IRTree modeling approach and its extension to a

multilevel data structure were described. Third, the association between the effects of questionnaire length on (the relative impact of) RS was investigated by using the extension of the IRTree models. Finally, supplemental exploratory analyses on the impact of RS and questionnaire length on regression coefficients with external criteria were reported.

Taken together, the papers reported in the three empirical chapters differed in their focus on the effects of the design features (sampling frequency and questionnaire length) and examined different outcome variables potentially affected by these design features. Figure 1 summarizes the three papers and their associated empirical studies and research questions. The General Discussion (Chapter 5) summarizes and critically reflects on the findings from the empirical chapters, places them in a broader research context, and discusses the methodological and practical implications of the present dissertation.

Figure 1

Overview of Papers



Note. Sampling Frequency was experimentally manipulated in Study 1. Questionnaire Length was experimentally manipulated in Study 2. The upper panel describes which data sets are used to investigate the outcome variables of the lower panel.

2 Paper 1: Participant Burden, Compliance, Within-Person Variance, and Within-Person Relationships

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2021). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01683-6

The Effects of Assessment Intensity on Participant Burden, Compliance, Within-Person Variance, and Within-Person Relationships in Ambulatory Assessment

Kilian Hasselhorn, Charlotte Ottenstein, and Tanja Lischetzke

University of Koblenz-Landau

Author Note

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant GRK 2277 (Research Training Group "Statistical Modeling in Psychology").

The analyses were preregistered on the OSF (Study 1: https://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2; Study 2: https://osf.io/xt3rf/?view_only=6e6bf509c6374759b56faac680c55825)

Correspondence concerning this article should be addressed to Kilian Hasselhorn, Department of Psychology, University of Koblenz-Landau, Fortstr. 7, 76829 Landau, Germany. E-mail: hasselhorn@uni-landau.de, Tel: +49 6341 280-31298

Abstract

Considering the very large number of studies that have applied ambulatory assessment (AA) in the last decade across diverse fields of research, knowledge about the effects that these design choices have on participants' perceived burden, data quantity (i.e., compliance with the AA protocol), and data quality (e.g., within-person relationships between time-varying variables) is surprisingly restricted. The aim of the current research was to experimentally manipulate aspects of an AA study's assessment intensity—sampling frequency (Study 1) and questionnaire length (Study 2)—and to investigate their impact on perceived burden, compliance, within-person variability, and within-person relationships between time-varying variables. In Study 1, students (n = 313) received either 3 or 9 questionnaires per day for the first 7 days of the study. In Study 2, students (n = 282) received either a 33- or 82-item questionnaire 3 times a day for 14 days. Within-person variability and within-person relationships were investigated with respect to momentary pleasant-unpleasant mood and state extraversion. The results of Study 1 showed that a higher sampling frequency increased perceived burden but did not affect the other aspects we investigated. In Study 2, longer questionnaire length did not affect perceived burden or compliance but yielded a smaller degree of within-person variability in momentary mood (but not in state extraversion) and a smaller within-person relationship between state extraversion and mood. Differences between Studies 1 and 2 with respect to the type of manipulation of assessment intensity are discussed.

Keywords: Ambulatory assessment, sampling frequency, questionnaire length, perceived burden, compliance, within-person variability, within-person relationships

The Effects of Assessment Intensity on Participant Burden, Compliance, Within-Person Variance, and Within-Person Relationships in Ambulatory Assessment

A growing body of literature is using ambulatory assessment (AA) in the fields of psychology and life science (Hamaker & Wichers, 2017). AA (also called daily diary, experience sampling, or ecological momentary assessment) is a method for assessing daily life experiences, for example, the ongoing behavior, experience, physiology, and environmental aspects of people in naturalistic and unconstrained settings (Fahrenberg, 2006). One of the main advantages of AA is that it allows researchers to study within-person dynamics (e.g., within-person relationships between time-varying variables) as well as individual differences in these within-person dynamics (Hamaker & Wichers, 2017). Furthermore, AA provides reduced recall bias and high ecological validity (Mehl & Conner, 2014; Trull & Ebner-Priemer, 2014).

When researchers plan an AA study, they have to make decisions about multiple design features in order to strike a balance between being able to gather rich information and ensuring that they do not overburden their participants (Carpenter, Wycoff, & Trull, 2016). Some of these features consist of 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), but many other design features could also be considered (e.g., the population of interest, item content, item difficulty, item order, the instructions given to the participants, the software used to signal the participants, financial compensation). For more detailed information about study design considerations and methods of data collection, see, for example, Mehl and Conner (2014). In the present research, we focused on sampling frequency and questionnaire length as important aspects of assessment intensity.

Considering the very large number of studies that have applied AA in the last decade across diverse fields of research, knowledge about the effects that these design choices have on participants' perceived burden, data quantity (i.e., compliance with the AA protocol), and data quality (e.g., careless responding or psychometric properties of measures) is surprisingly restricted (cf. Eisele et al., 2020). The (relatively scarce) previous methodological research on the effects of design-related characteristics on aspects of AA data (e.g., data quantity and data quality) can be divided into two groups: On the one hand, meta-analytic research has analyzed whether between-study differences in assessment intensity are related to between-study differences in compliance (e.g., Jones et al., 2019; Ottenstein & Werner, 2020; Vachon, Viechtbauer, Rintala, & Myin-Germeys, 2019) and in the proportion of within-person variance in time-varying constructs (Podsakoff, Spoelma, Chawla, & Gabriel, 2019)—the latter being a characteristic of AA data that should be of particular interest to researchers who want to investigate within-person dynamics (Hamaker & Wichers, 2017). On the other hand, experimental research has manipulated assessment intensity in an AA study and analyzed its effects on participant burden (e.g., Eisele et al., 2020; Stone et al., 2003), compliance (e.g., Conner & Reid, 2012; Stone et al., 2003), and careless responding (Eisele et al., 2020). Whereas third variables cannot be ruled out in correlational analyses of between-study differences, experimental AA studies have the advantage that the internal validity of causal conclusions is higher. In the present research, we aimed to contribute to the literature in the following ways: First, with respect to assessment intensity as the independent variable, in Study 1, we manipulated sampling frequency to allow comparisons with previous research, whereas in Study 2, we manipulated questionnaire length, which (to our knowledge) has been experimentally manipulated in only one study (Eisele et al., 2020) to date. Second, with respect to characteristics of the AA data as dependent variables, our aim was to investigate both previously studied variables (perceived burden, compliance) and understudied variables

(within-person variability, within-person relationships between time-varying variables). See Table 1 for an overview of and additional information (study design, studied design features, dependent variables, and results) about previous research in this area. In the following, we address each of these dependent variables in turn to derive our hypotheses (which were preregistered on the OSF, view-only link for review: https://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2).

Perceived Burden

For participants in an AA study, a higher (vs. lower) assessment intensity (e.g., more questionnaires per day or more items per questionnaire) means that participants have to invest more time and energy in participating in the study if they aim to be thorough (Santangelo, Ebner-Priemer, & Trull, 2013). A higher perceived burden has been conceptualized as a process that can result in a reduction in the quantity and quality of AA data (Eisele et al., 2020; Fuller-Tyszkiewicz et al., 2013; Santangelo et al., 2013). Although several researchers have stated that increases in sampling frequency or questionnaire length increase perceived burden (Moskowitz, Russell, Sadikaj, & Sutton, 2009; Napa Scollon, Prieto, & Diener, 2009; Santangelo et al., 2013), there are only a few empirical studies that have experimentally manipulated assessment intensity and analyzed perceived burden as an outcome (Eisele et al., 2020; Stone et al., 2003). In the study by Stone et al. (2003), participants with pain syndromes were randomly assigned to either no AA phase or sampling densities of 3, 6, and 12 questionnaires per day over 2 weeks. Perceived burden, which was assessed retrospectively after the AA phase, was higher in groups with a higher sampling frequency. In a sample of students, Eisele et al. (2020) analyzed the effects of sampling frequency and questionnaire length on perceived burden over 2 weeks. They operationalized perceived burden as momentary perceived burden, which was measured with each questionnaire, and as retrospective perceived burden, which was measured after the AA phase. Their results

revealed no increase in perceived burden with a higher sampling frequency, but perceived burden did increase with longer questionnaires (Eisele et al., 2020). One reason for the difference in results between these experimental AA studies with respect to the effect of sampling frequency on perceived burden might be the population of interest (clinical vs. nonclinical). In the present research, we decided to target a nonclinical population (students), as Eisele et al. (2020) did. Moreover, given that Eisele et al.'s study is the only previous study that experimentally manipulated questionnaire length, and given that we know of no correlational papers or meta-analyses on this topic, more experimental research on the effect of this aspect of assessment intensity on perceived burden is needed.

Compliance

In AA studies, high compliance rates per person are considered particularly important for obtaining a representative picture of individuals' everyday experiences and behaviors (Stone et al., 2003) because missing data can lead to biased inferences about person-level (aggregated) data (Courvoisier, Eid, & Lischetzke, 2012). A higher (vs. a lower) sampling frequency can be assumed to potentially compromise compliance. In the case of more frequent or longer questionnaires, participants might try to intentionally reduce the burden by not responding to prompts or by completing only a portion of the items on a particular measurement occasion (Vachon et al., 2019).

Surprisingly, previous experimental AA studies that manipulated sampling frequency in clinical samples (Stone et al., 2003; 3 vs. 6 vs. 12 questionnaires per day for 14 days) or in nonclinical samples (Conner & Reid, 2012; 1, 3, or 6 daily questionnaires for 13 days, McCarthy et al., 2015; 1 vs. 6 daily questionnaires for 14 days; Eisele et al., 2020; 3 vs. 6 vs. 9 questionnaires for 14 days; Walsh & Brinker, 2016; 20 questionnaires for either 1 or 2 days) found that the experimental groups did not differ in their compliance. Meta-analyses and pooled data analyses have provided a somewhat mixed picture: Some studies found no

support for the idea that a higher sampling frequency is related to lower compliance rates in clinical samples (Jones et al., 2019; Ono, Schneider, Junghaenel, & Stone, 2019; Soyster, Bosley, Reeves, Altman, & Fisher, 2019) or in other (clinical samples mixed with nonclinical) samples (Morren, Dulmen, Ouwerkerk, & Bensing, 2009), whereas a recent meta-analysis that focused on AA studies in mental health research (Vachon et al., 2019) found lower compliance rates in studies that administered larger numbers of questionnaires per day.

With respect to questionnaire length, Eisele et al. (2020) found that longer questionnaires (60 items) led to lower compliance than shorter questionnaires (30 items). To our knowledge, this study is the only one to ever experimentally manipulate questionnaire length between conditions. Most meta-analyses and pooled data analyses have found no support for the idea that a longer questionnaire leads to lower compliance (Jones et al., 2019; Ono et al., 2019; Rintala, Wampers, Myin-Germeys, & Viechtbauer, 2019; Soyster et al., 2019; Vachon et al., 2019) with the exception of Morren et al. (2009), who found that compliance was positively associated with shorter questionnaires.

Taken together, given that evidence of an effect of sampling frequency on compliance has been mixed, there is a need to scrutinize this effect further. Also, given that only one previous study manipulated questionnaire length (Eisele et al., 2020), there is also a need for more experimental research on the effect of questionnaire length on compliance.

Within-Person Variability

Within-person variability is a prerequisite for investigating within-person dynamics (Heck & Thomas, 2015; Hox, 2002; Raudenbush & Bryk, 2002), and researchers have warned that research on within-person relationships between time-varying variables should not be conducted when within-person variability is low (Podsakoff et al., 2019; Rosen, Koopman, Gabriel, & Johnson, 2016; Sonnentag, Binnewies, & Mojza, 2008; Trougakos, Beal, Green, & Weiss, 2008).

In AA studies with a higher sampling frequency or longer questionnaire length, participants might become more fatigued over time (e.g., Beal, 2015) and might consequently respond in a more heuristic, less nuanced way to repeated prompts, thereby reducing the degree of within-person variability in time-varying variables (Fuller-Tyszkiewicz et al., 2013; Podsakoff et al., 2019). With respect to sampling frequency, Podsakoff et al. (2019) added that another process might work in the opposite direction: More frequent prompts might give participants the opportunity to become more aware of differences between a current state and previous states, thereby potentially increasing within-person variability in time-varying variables.

Empirical evidence on whether higher (vs. lower) assessment intensity has an effect on the degree of within-person variability in time-varying variables is scarce. Two AA studies that analyzed whether the degree of within-person variability changed over the course of an AA study have provided indirect evidence: In a community sample of young women, Fuller-Tyszkiewicz et al. (2013) found that the within-person variability in state body dissatisfaction scores declined as a function of the number of days into the study when they were measured. In a sample of depressed patients, Vachon et al. (2016) analyzed the trajectory of within-person variability in psychological states across a twice-daily AA study that spanned a total of 5 months. Their results revealed a decrease in the variability of cognitive (e.g., rumination) and affective (e.g., depressed mood) states across the course of the study. Podsakoff et al. (2019) conducted a meta-analysis on AA studies that had obtained their data from working employees and analyzed whether between-study differences in sampling frequency (number of questionnaires per day) and study duration (number of days) were related to between-study differences in the proportion of within-person variance. Among the time-varying constructs that were included, momentary affect and stressors were studied the most. Podsakoff et al. (2019) found that higher sampling frequency (but not study duration) predicted larger within-

person variability. Taken together, the findings of these studies provide a mixed picture. Moreover, the evidence is not fully conclusive for methodological reasons: When analyzing within-person variability as a function of the amount of time into the study, variables that were confounded with the day of the week (e.g., change in compliance on weekdays compared with weekend days; Phillips, Phillips, Lalonde, & Dykema, 2014), the day (e.g., changes in daily activities), or the month (e.g., seasonal effects) might provide alternative explanations. Likewise, in correlational analyses at the level of studies, between-study differences in third variables could have driven the effect. Hence, in the present research, we followed Podsakoff et al.'s (2019) suggestion that "scholars may find experimental designs to be particularly well-suited to addressing research questions about the effect of study design on variance in measures" (p. 15). To our knowledge, no study has yet analyzed the effect of manipulations of assessment intensity on within-person variability.

Within-Person Relationships between Time-Varying Variables

A great deal of research on within-person dynamics tends to focus on within-person relationships between time-varying variables (Liu, Xie, & Lou, 2019; May, Junghaenel, Ono, Stone, & Schneider, 2018; Sitzmann & Yeo, 2013). Fuller-Tyszkiewicz et al. (2013) argued that a decrease in within-person variability can lead to a decline in the strength of within-person relationships between time-varying variables. Empirically, however, the decrease in within-person variability that was found did not translate into smaller within-person relationships as a function of the number of days into the study (Fuller-Tyszkiewicz et al., 2013).

To our knowledge, no AA studies to date have investigated the effect of experimentally manipulated assessment intensity on within-person relationships between time-varying variables. Therefore, more experimental research on the effect of assessment intensity on within-person relationships between time-varying variables is needed.

The Current Research

The aim of the current research was to experimentally manipulate sampling frequency (Study 1) and questionnaire length (Study 2) as aspects of an AA study's assessment intensity and to investigate their impact on perceived burden, compliance, within-person variability, and within-person relationships between time-varying variables. We expected that higher assessment intensity would increase perceived burden (Hypothesis 1) and would decrease compliance (H2), within-person variability (H3), and within-person relationships between time-varying variables (H4). Note that we preregistered these hypotheses in February 2019 (Study 1) and in January 2020 (Study 2) and that some of the previous research we cited had not been published at that time.[1]

To test Hypothesis 1 on perceived burden, we decided to assess perceived burden as both a daily and a retrospective burden, similar to Eisele et al.'s (2020) study. To test Hypothesis 3 on within-person variability and Hypothesis 4 on within-person relationships, we selected momentary mood and state extraversion as two time-varying constructs that (a) have frequently been assessed in previous research, (b) have been shown to possess adequate within-person variability (e.g., McCabe & Fleeson, 2016; Podsakoff et al., 2019), and (c) have been shown to be related within persons across multiple studies (e.g., Fleeson, Malanos, & Achille, 2002; Lischetzke, Pfeifer, Crayen, & Eid, 2012; McNiel & Fleeson, 2006; McNiel, Lowman, & Fleeson, 2010).

Study 1

Previous AA studies on state personality have typically administered four to five questionnaires per day (e.g., Fleeson & Gallagher, 2009). AA studies on momentary mood have shown a larger range of sampling frequencies: Studies assessing mood in the field of applied psychology have typically included one to three questionnaires per day (Podsakoff et al., 2019), whereas studies in the field of affect dynamics research have typically included seven to 10 questionnaires per day (Dejonckheere et al., 2019). For our experimental manipulation of sampling frequency in Study 1, we therefore selected a low and a high number within this observed range (three vs. nine questionnaires per day).

Method

Study design. The study consisted of an initial online survey, an AA phase across 7 days with either three or nine questionnaires per day (depending on the experimental condition that participants had been randomly assigned to), and a retrospective online survey. For the AA phase, participants chose a specific time schedule that best fit their waking hours (9:00-21:00 or 10:30-22:30). In the low sampling frequency group, the three questionnaires per day were distributed evenly across the day. In the high sampling frequency group, the first, the fifth, and the ninth questionnaire were scheduled at the same time of day as the three questionnaires from the low sampling frequency group, and the six additional questionnaires were distributed between these questionnaires (see Table 2 for more detailed information).

After the 7-day AA phase, a second 7-day AA phase followed immediately. This time, the sampling frequency was switched between the groups. This was done to ensure that each participant invested a comparable amount of time participating in the study so that the financial compensation, which was the same for both groups, was fair. Given that our focus was on the between-group comparison (high vs. low sampling frequency), and not on the

effect of switching sampling frequency within persons, the analyses in the present paper are based on the data from the first 7-day AA phase.

During the initial online survey, participants completed a demographic questionnaire and trait self-report measures. In each AA questionnaire, participants rated their momentary motivation, time pressure, mood, clarity of mood, and state personality (extraversion and conscientiousness). In the last AA questionnaire each day, participants additionally rated their daily stress and perceived burden. In the retrospective online survey, participants rated their perceived burden and careless responding with respect to the past 7 days (as well as other constructs that were not relevant for the present analyses: retrospective mood, clarity of mood, and attention to feelings).

Participants. Participants were required to be currently enrolled as a student and to be in possession of an android smartphone. Participants were recruited via flyers, posters, e-mails, and posts on Facebook during students' semester breaks.

As most hypotheses in this study focused on group differences, we based our sample size considerations on the power to detect a small to moderate (d = 0.30) mean difference (independent-samples t test). We needed 278 participants to achieve a power of .80.

A total of 474 individuals filled out the initial online survey. Due to technical problems with the software for the AA phase, various participants could not synchronize their smartphone with our study and withdrew their participation. One of the reasons for dropout was that participants with an iOS smartphone realized only at this stage that participation required an Android smartphone, as had been indicated in the information we gave them about the study. A total of 318 individuals took part in the AA phase that followed (155 individuals in the low sampling frequency condition), and 200 individuals responded in time to the retrospective online survey after the first 7 days (within the prespecified time frame of 12 hr). Participants who did not respond to the retrospective online survey were not excluded

from the data analyses. Data from five participants were excluded from the analyses due to careless responding (see the data cleaning section). The final sample consisted of 313 students (low sampling frequency group: 86% women; age Range: 18 to 34 years, $M = 23.18$, $SD = 3.23$; high sampling frequency group: 83% women; age Range: 18 to 40 years, $M = 23.98$, $SD = 4.12$).

Procedure. All study procedures were approved by the psychological Ethics Committee at the University Koblenz-Landau, Germany. After informed consent was obtained, the study began with the initial online survey. Subsequently, participants were randomly assigned to one of two experimental conditions (low sampling frequency or high sampling frequency) and randomly assigned to a starting day of the week. Prior to their AA phase, participants received a manual that explained how to install and run the smartphone application movisensXS, Versions 1.4.5, 1.4.6, 1.4.8, 1.5.0, and 1.5.1 (movisens GmbH, Karlsruhe, Germany) and connect to our study as this was required for participation. Participants were told that the number of questionnaires administered per day would range from three to nine times over the 2 weeks. At each measurement occasion, participants could either respond to the questionnaire or delay their response for up to 15 min. Participants who missed the first alarm were signaled 5 min later. If the questionnaire was not started by 15 min after the first signal, the questionnaire became unavailable. At the end of the 7th day of each AA phase (21:00 for participants with the early time schedule or 22:30 for participants with the later time schedule), participants were sent a link to the retrospective online survey via e-mail. This online survey had to be completed within a 12-hr time frame. Students were given 15€ in exchange for their participation if they had answered at least 50% of the AA questionnaires and had the chance to win 25€ extra if they had answered at least 80% of the AA questionnaires. Furthermore, at the end of the second retrospective online survey, they could indicate whether they wished to receive personal feedback regarding the constructs

measured in the study after their participation was complete. In the low sampling frequency group (high sampling frequency group), 98 (92) participants requested feedback, 10 (10) participants did not want feedback, and 45 (58) participants did not answer the item.

Data cleaning. To screen for careless responding, we analyzed inconsistent responding across reverse-poled items (see below) and a "Use Me" item (Meade & Craig, 2012). First, data from five participants (three in the high sampling frequency group) who indicated in the retrospective online survey that their data should not be used in our analyses were excluded from the analyses. The remaining 313 participants had completed 9,158 AA questionnaires.

Subsequently, we removed 332 AA questionnaires (149 AA questionnaires in the low sampling frequency group) due to inconsistent responding (Meade & Craig, 2012) across the reverse-poled (mood) items.[2] Because these questionnaires had been completed by the participants, compliance was unaffected by the AA questionnaires that were removed (see the Measures section). Hence, our analyses were based on 8,528 AA questionnaires nested in 313 participants (with the exception of the compliance analysis, which was based on 9,158 AA questionnaires nested in 313 participants).

Measures[3]. ***Male.*** A factor was used to indicate gender, with a value of 0 for female participants and 1 for male participants.

Feedback. A factor was used to indicate whether participants wanted feedback, with a value of 0 for participants who did not want to receive feedback and 1 for participants who wanted to receive feedback.

Sampling frequency. A factor was used to indicate the sampling frequency, with a value of 0 for the low sampling frequency group and 1 for the high sampling frequency group.

Momentary mood. We measured momentary (pleasant-unpleasant) mood with an adapted short version of the Multidimensional Mood Questionnaire (Steyer, Schwenkmezger,

Notz, & Eid, 1997) that has been used in previous AA studies (Lischetzke et al., 2012; Ottenstein & Lischetzke, 2020). Participants indicated how they *felt at the moment* on four items (bad-good [reverse-scored], unwell-well, unhappy-happy [reverse-scored], and unpleased-pleased). The response format was a 7-point Likert scale with each pole labeled (e.g., 1 = *very unwell* to 7 = *very well*). We calculated a mean score across the items so that a higher score indicated more pleasant mood. The within-person ω (Geldhof, Preacher, & Zyphur, 2014) was .91, and the between-person ω was .99.

State extraversion. We measured state extraversion by taking the adjectives that McCabe and Fleeson (2016) had introduced for each subcomponent (sociability, assertiveness, and talkativeness) and modifying them so that they formed three bipolar items (one for each subcomponent). Participants indicated how they *behaved in the last half hour* on the items (outgoing-unsociable [reverse-scored], unassertive-assertive, and talkative-quiet [reverse-scored]. The response format was a 7-point Likert scale with each pole labeled (e.g., 1 = *very quiet* to 7 = *very talkative*) plus an extra category (i.e., *not applicable*) if the respondent wanted to skip the question (displayed below the Likert scale). We used a bipolar (instead of a unipolar) response format to have a common response format for the time-varying dependent variables (momentary mood, state extraversion). We calculated a mean score across items so that a higher score indicated more extraverted behavior. The within-person ω (Geldhof et al., 2014) was .80, and the between-person ω was .91.

Daily perceived burden. Daily perceived burden was measured using three items from Stone et al. (2003). Participants were asked on a 7-point Likert scale (1 = *not at all* to 7 = *very much so*): "How much of a burden was it to participate in the study during the day?" "How much did participating in the study interfere with your usual activities?" and "How much were you annoyed with the number of times you were signaled per day?" We calculated a

mean score across items so that a higher score indicated more perceived burden. The within-person ω (Geldhof et al., 2014) was .78, and the between-person ω was .94.

Retrospectively perceived burden. Retrospectively perceived burden was measured with the same three items as daily perceived burden with the modification that they referred to the previous 7-day AA phase. We calculated a mean score across items so that a higher score indicated more retrospectively perceived burden. Revelle's omega total (McNeish, 2018) was .79.

Compliance. Compliance at the questionnaire level was defined as having responded to the last item on the AA questionnaire (coded 1 = yes and 0 = no). We calculated the relative compliance across all questionnaires for each person so that a higher score indicated more compliance.

Data analytic methods. Hypothesis 1 (on the between-group difference in perceived burden) and Hypothesis 2 (on the between-group difference in compliance) were tested with two-level regression models with daily perceived burden (and daily compliance) at Level 1 and persons at Level 2. To test group differences in the respective variables, we included sampling frequency at the person level.[4] Effects on retrospective perceived burden (Hypothesis 1) were tested with an independent-samples *t* test using R. Afterwards, we corrected for multiple testing using Benjamini and Hochberg's (1995) procedure for controlling the false discovery rate (FDR) in H1 and H3. When testing for differences between groups, we corrected for two multiple tests (daily and retrospective perceived burden). Hypothesis 3 (on the between-group difference in within-person variability) was analyzed using a multigroup multilevel model for questionnaires nested in persons in Mplus (Muthén & Muthén, 1998-2017). We applied the latent variable modeling procedure proposed by Dowling, Raykov, & Marcoulides (2018) to evaluate between-group differences in within-person variability. The model decomposes the total variance into between-person variance

and within-person variance for each group. By using the Mplus MODEL CONSTRAINT option, the statistical significance of the between-group difference in within-person variance can be tested. To compare within-person variability between the experimental groups, within-person variance was estimated on the basis of the three surveys that were scheduled at the same times across groups (i.e., in the low sampling frequency group, the three daily surveys were used; and in the high sampling frequency group, the corresponding first, fifth, and ninth surveys of the day were used; see Table 2). Given that fluctuations in mood follow a diurnal rhythm (see, e.g., Thayer, 1978; Watson, 2000), using only the surveys that were scheduled at the same times across groups allowed us to rule out time of day as an alternative explanation for potential between-group differences in within-person variability. As in Hypothesis 1, we corrected for two multiple tests (momentary mood and state extraversion) by applying the procedure presented by Benjamini and Hochberg (1995). Hypothesis 4 (on between-group differences in within-person relationships) was analyzed with two-level regression models with questionnaires at Level 1 and persons at Level 2. Person-mean-centered state extraversion was used as a Level 1 predictor of momentary mood. Level 2 random intercepts and random slopes were included in the model (LeBeau, Song, & Liu, 2018). To test whether the two experimental conditions differed in their association between state extraversion and momentary mood, we additionally included the cross-level interaction between sampling frequency at the person level and state extraversion at the questionnaire level.

Additionally, as suggested by the reviewers during the peer review process, we conducted several exploratory analyses. First, we explored whether age, gender, and feedback at Level 2 would be found to predict compliance. Previous research has provided a mixed picture of the effects of age and gender on compliance (Ono et al., 2019; Rintala et al., 2019; Soyster et al., 2019; Vachon et al., 2019).

Second, we explored within-person differences in the effects of different sampling frequencies (which were switched after 7 days) on daily perceived burden. To do so, we tested a multilevel model including both weeks of the AA phase with a Week (at the questionnaire level) x Sampling Frequency (at the person level) cross-level interaction.

For Hypothesis 1, the *t* test was computed in the R environment (R Core Team, 2020). All multilevel regression models were computed with the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015), and the *p*-values were created with the R package lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017). The procedure by Dowling et al. (2018) and the multilevel reliabilities were computed in Mplus 8.3 (Muthén & Muthén, 1998-2017).

Results

To test our hypotheses, which involved a directional prediction, we employed one-sided significance tests (Cho & Abe, 2013). Accordingly, we interpreted one-tailed *p*-values and corresponding two-sided 90% confidence intervals for these tests (i.e., for the difference between the group means). Some of the reported tests (e.g., intercorrelations among the study variables, the main effect of the experimental group on momentary mood in the multilevel model) did not refer to a directional prediction, and hence, we reported two-sided *p*-values for these estimates. In the text, we explicitly labeled the *p*-values as one-sided when this applied.

Table 3 presents the means, standard deviations, and within- and between-person correlations between the variables for each group separately. For all momentary and daily measures, there was a substantial amount of within-person variance, ranging from 56% for momentary mood to 79% for state extraversion for the low sampling frequency group and from 54% for momentary mood to 79% for state extraversion for the high sampling frequency group.

Perceived burden. In line with Hypothesis 1, the low sampling frequency group (M = 2.01, SD = 0.64) perceived a lower daily burden than the high sampling frequency group (M =

2.56, *SD* = 0.70), *t*(291) = 7.47, one-tailed *p* < .001, 90% CI [0.44, 0.69], *d* = 0.83. This finding remained significant after we corrected the false discovery rate.

Similarly, for the retrospective measure, the low sampling frequency group (*M* = 2.20, *SD* = 0.75) perceived a significantly lower retrospective burden than the high sampling frequency group (*M* = 2.82, *SD* = 0.75), *t*(192) = 5.78, one-tailed *p* < .001, 90% CI [−0.80, −0.44], *d* = 0.83. This finding also remained significant after we corrected the false discovery rate.[5]

Compliance. Contrary to Hypothesis 2, the low sampling frequency group (*M* = 0.71, *SD* = 0.25) did not demonstrate higher compliance than the high sampling frequency group (*M* = 0.68, *SD* = 0.25), *t*(311) = -1.28, one-tailed *p* =.101, 90% CI [−0.08, 0.01], *d* = -0.15.

Within-person variability. Contrary to Hypothesis 3, the low sampling frequency group (Estimate = 0.83, *SE* = 0.06) did not show a significantly higher within-person variance in momentary mood than the high sampling frequency group (Estimate = 0.75, *SE* = 0.05), *z* = -1.20, one-tailed *p* = .116, 90% CI = [−0.03, 0.21], *d* = -0.10.[6]

Similarly, for state extraversion, the low sampling frequency group (Estimate = 2.01, *SE* = 0.12) did not show a significantly higher within-person variance than the high sampling frequency group (Estimate = 1.94, *SE* = 0.11), *z* = -0.43, one-tailed *p* = .336, 90% CI = [−0.20, 0.33], *d* = -0.08.[3]

Within-person relationships between time-varying variables. As can be seen in Table 4 (Model 2), unexpectedly, the cross-level interaction term (for the interaction between sampling frequency at Level 2 and state extraversion at Level 1) had a positive sign, which means that the low sampling frequency group (b = 0.16) had a descriptively smaller regression coefficient for state extraversion than the high sampling frequency group (b = 0.21). The cross-level interaction term was not significantly different from zero, *t*(198) = 1.35, one-tailed *p* = .911, 90% CI [−0.01, 0.09].

Exploratory data analysis. ***Predictors of compliance.*** None of the variables from the exploratory analyses were significantly related to compliance: gender, $t(203) = 0.20$, $p = .840$, 95% CI [−0.07, 0.09]; age, $t(202) = 0.72$, $p = .475$, 95% CI [-0.004, 0.010]; requesting personal feedback, $t(204) = 0.12$, $p = .905$, 95% CI [-0.09, 0.10].

Effects of sampling frequency as a within-person factor. In a multilevel model with sampling frequency (low vs. high) as a within-person factor, order (low sampling frequency first vs. high sampling frequency first) as a between-person factor, and their cross-level interaction, a main effect of sampling frequency (b = 0.94, SE = 0.06), t(269) = 16.36, p < .001, 95% CI [0.82, 1.05], and a main effect of order (b = -0.38, SE = 0.07), t(282) = -5.32, p < .001, 95% CI [-0.52, -0.24], emerged. The cross-level interaction term was not significantly different from zero (b = 0.02, SE = 0.08), t(270) = 0.25, p = .804, 95% CI [-0.14, 0.18]. That is, during the high sampling frequency phase, participants reported a higher burden than during the low sampling frequency phase, and in the group that had started with the high sampling frequency phase, subjective burden values were lower, on average, than in the group that had started with the low sampling frequency phase.

Discussion

Using an experimental manipulation of sampling frequency in an AA study, we found that, as expected, a higher sampling frequency led to higher perceived burden (H1). However, contrary to our expectations, the high versus low sampling frequency groups did not differ in compliance (H2), within-person variability in momentary mood and state extraversion (H3), or the within-person relationship between momentary mood and state extraversion (H4).

Our finding that the sampling frequency had an effect on perceived burden is in line with previous assumptions (Moskowitz et al., 2009; Santangelo et al., 2013) and with the empirical research by Stone et al. (2003). Contrary to our results, Eisele et al. (2020) found no effect of sampling frequency on perceived burden. A possible explanation could be that in

Eisele et al.'s study, the effect of sampling frequency "was canceled out by the increased motivation due to the higher incentive" (Eisele et al., p. 12) in the high sampling frequency group (40 vs. 80 Euros in the group with three vs. nine AA questionnaires per day, respectively), whereas in our study, financial compensation for the complete study did not differ between the experimental groups.

Our finding that the sampling frequency had no effect on compliance is in line with previous research (Conner & Reid, 2012; Eisele et al., 2020; McCarthy et al., 2015; Ono et al., 2019; Soyster et al., 2019; Stone et al., 2003; Walsh & Brinker, 2016). The results indicate that the higher burden that was reported in the high sampling frequency group did not translate into less effort in responding to the AA prompts. An ad hoc explanation is that participants in the high sampling frequency group might have kept up a similarly high compliance rate (despite perceiving a higher burden) because they wanted the personal feedback after study participation to show a representative picture of their experience and behavior during the study.

To our knowledge, this study is the first to analyze the effect of experimentally manipulated sampling frequency on within-person variability and the within-person relationship between time-varying variables. Although it might sound like good news to researchers applying AA designs to study within-person dynamics that we did not find differences between the low and the high sampling frequency groups, it seems premature to conclude that assessment intensity has no effect on within-person (co)variability. Hence, in Study 2, we aimed to conceptually replicate this finding by using a different manipulation of assessment intensity (questionnaire length).

Study 2

In Study 2, we wanted to conceptually replicate and generalize our findings from Study 1 with a different manipulation of an aspect of assessment intensity. Therefore, we chose to manipulate the questionnaire length per questionnaire as another central design characteristic instead of the sampling frequency. Moreover, we extended the duration of the AA phase from 1 to 2 weeks.

Previous meta-analyses and pooled data analyses have included studies with different ranges of numbers of items (see Table 1). Our aim was to maximize the difference in questionnaire length between the groups (in a range that was realistic for AA studies) while holding constant the constructs that were measured across groups (by using short vs. long forms for each construct).

Method

Study design. The study consisted of an initial online survey, an AA phase across 14 days with three short or long questionnaires per day (depending on the experimental condition that participants had been randomly assigned to), and a retrospective online survey.

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 = 1.65 min, SD = 0.63) was lower, on average, than in the long questionnaire group (M = 3.89 min, SD = 3.42). The two groups answered questions about the same constructs. This allowed us to investigate the effect of questionnaire length without the confounding effect of measuring different constructs between the groups. The difference in the number of items between these groups was achieved by using a short versus a long version of the measures of the constructs (see the Measures section).

During the initial online survey, participants completed a demographic questionnaire and trait self-report measures. In each AA questionnaire, participants rated their momentary motivation, time pressure, state personality, situation characteristics, and momentary mood. In the last AA questionnaire per day, participants additionally rated their perceived burden. In the retrospective online survey, participants rated their retrospective mood, perceived burden, and careless responding regarding the past 14 days. Additionally, participants rated their trait personality again.

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, and the last questionnaire was sent to participants on February 10, 2020. The a priori power analysis was conducted in the same way as in Study 1.

A total of 303 individuals filled out the initial online survey, 284 individuals took part in the AA phase that followed (142 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 data analyses. Data from two participants were excluded from the analyses due to careless responding (see the Data Cleaning section). The final sample consisted of 282 students (short questionnaire group: 83% women; age Range: 18 to 39 years, $M = 23.20$, $SD = 3.45$; long questionnaire group: 87% women; age Range: 18 to 55 years, $M = 22.90$, $SD = 3.81$).

Procedure. All study procedures were approved by the psychological Ethics Committee at the University Koblenz-Landau, Germany. 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 regarding the measured constructs after they participated. In the short questionnaire group (long questionnaire group) 133 (131) participants requested feedback, and 9 (9) participants did not want feedback.

Data cleaning. To screen for careless responding, we analyzed inconsistent responding across reverse-poled items (see below) and a "Use Me" item (Meade & Craig, 2012). First, data from two participants who indicated in the retrospective online survey that their data should not be used in the analyses were excluded from the analyses. The remaining 282 participants had completed 8,611 AA questionnaires. Finally, we removed 26 AA questionnaires (14 AA questionnaires in the short questionnaire group) due to inconsistent responding (Meade & Craig, 2012) across reverse-poled (mood) items.[7] Because these questionnaires had been completed by the participants, compliance was unaffected by the AA questionnaires that were removed (see the Measures section). Hence, our analyses were based on 8,585 AA questionnaires nested in 282 participants (with the exception of the compliance analysis, which was based on 8,611 AA questionnaires nested in 282 participants).

Measures.[8] 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). As a result, we achieved variation in questionnaire length while at the same time measuring the same constructs in the two groups. Only the items that were included in the short questionnaire group (which were the ones analyzed in both groups) will be described in the following (for the additional items assessed in the long questionnaire group, see the supplemental online material).

Male. A factor was used to indicate gender, with a value of 0 for female participants and 1 for male participants.

Feedback. A factor was used to indicate whether participants wanted feedback, with a value of 1 for participants who wanted to receive feedback and 0 for participants who did not want to receive feedback.

Questionnaire length. A factor was used to indicate questionnaire length, with a value of 0 for the short questionnaire and 1 for the long questionnaire.

Momentary mood. To measure momentary (pleasant-unpleasant) mood, we used two items from Study 1 (bad-good [reverse-scored], unwell-well). We calculated a mean score across two mood items so that a higher score indicated more pleasant mood. The within-person α (Geldhof et al., 2014) was .86, and the between-person α was .97.

State extraversion. We measured state extraversion with an adapted version of the adjectives from Saucier's (1994) unipolar Big Five Mini-Markers (Comensoli & MacCann, 2015). Participants indicated how they *behaved in the last half hour* on eight items (bashful [reverse-scored], bold, energetic, extraverted, quiet [reverse-scored], shy [reverse-scored], talkative, and withdrawn [reverse-scored]). The response format was a 5-point Likert scale with each pole labeled (1 = *extremely inaccurate* to 5 = *extremely accurate*). We calculated a mean score across the eight items so that a higher score indicated more extraverted behavior. The within-person ω (Geldhof et al., 2014) was .72, and the between-person ω was .59.

Daily Perceived burden. To measure daily perceived burden, we used the same items as in Study 1. We calculated a mean score across items so that a higher score indicated more perceived burden. The within-person ω (Geldhof et al., 2014) was .71, and the between-person ω was .91.

Retrospective perceived burden. Retrospectively perceived burden was measured with the same three items as in Study 1. Revelle's omega total (McNeish, 2018) was .82.

Compliance. Compliance at the questionnaire level was defined as having responded to the last item on the AA questionnaire (coded 1 = *yes* and 0 = *no*). We calculated the relative compliance across all questionnaires for each person so that a higher score indicated more compliance. When there were technical problems and participants had not received the AA questionnaire in time, they could not respond to the questionnaire. In these cases, we subtracted the number of AA questionnaires (which were missed due to technical problems) from the theoretical maximum number of completed AA questionnaires allowed by our protocol before we calculated the relative compliance.

Data analytic methods. To compare the experimental questionnaire length groups with respect to compliance and perceived burden, the analyses were the same as in Study 1. To compare the experimental groups with respect to within-person variability in mood and state extraversion and the relation between state extraversion and momentary mood, the within-person mood/extraversion scores were based on the items that were displayed in both groups (i.e., items that were displayed exclusively in the long questionnaire were excluded from all analyses). As in Study 1, H1 and H3 were corrected for multiple tests (Benjamini & Hochberg, 1995).

Results

Table 5 presents the means, standard deviations, and within- and between-person correlations between the variables. For all momentary and daily measures, there was a substantial amount of within-person variance, ranging from 59% for daily perceived burden to

81% for state extraversion for the short questionnaire group and from 57% for daily perceived burden to 79% for state extraversion for the long questionnaire group.

Perceived burden. Contrary to Hypothesis 1, daily perceived burden in the short questionnaire group (M = 2.40, SD = 0.67) was not lower than in the long questionnaire group (M = 2.51, SD = 0.77), $t(268) = 1.29$, one-tailed $p = .099$, 90% CI [−0.03, 0.26], $d = 0.14$.

Retrospective burden was also not significantly lower in the short questionnaire group (M = 2.71, SD = 0.85) than in the long questionnaire group (M = 2.77, SD = 0.92), $t(233) = 0.60$, one-tailed $p = .276$, 90% CI [−0.26, 0.12], $d = 0.08$.[9]

Compliance. Contrary to Hypothesis 2, the compliance rate in the short questionnaire group (M = .75, SD = 0.27) was not significantly higher than in the long questionnaire group (M = .72, SD = 0.28), $t(219) = -1.08$, one-tailed $p = .142$, 90% CI [−0.040, 0.008], $d = -0.10$.

Within-person variability. In line with Hypothesis 3, the short questionnaire group (Estimate = 1.19, SE = 0.07) showed a higher degree of within-person variability in momentary pleasant-unpleasant mood than the long questionnaire group (Estimate = 1.00, SE = 0.06), $z = -2.03$, one-tailed $p = .021$, 90% CI [0.04, 0.35], $d = -0.24$.[10] The finding remained significant after we corrected the false discovery rate. Descriptively, the short questionnaire group (Estimate = 0.35, SE = 0.02) also showed a higher degree of within-person variability in state extraversion than the long questionnaire group (Estimate = 0.31, SE = 0.02), but this difference was not significantly different from zero, $z = -1.50$, one-tailed $p = .067$, 90% CI [-0.003, 0.076], $d = -0.13$.[10]

Within-person relationships between time-varying variables. As can be seen in Table 6, the cross-level interaction between questionnaire length and state extraversion was significant, $t(234) = -2.98$, one-tailed $p = .003$, 90% CI [−0.29, −0.08]. As expected, the slope coefficient for state extraversion was larger in the short questionnaire group (b = 0.65) than in the long questionnaire group (b = 0.47). As a quasi R^2 measure of the cross-level interaction,

we calculated the proportional reduction in the Level 2 variance of state extraversion slopes when questionnaire length was added as a predictor of the slopes (Raudenbush & Bryk, 2002). It was .04.

Exploratory data analysis. ***Predictors of compliance.*** As in Study 1, none of the variables from the exploratory analyses were significantly related to compliance: gender, $t(214) = 0.77$, $p = .440$, 95% CI [−0.02, 0.06]; age, $t(217) = 0.05$, $p = .959$, 95% CI [-0.004, 0.004]; requesting personal feedback, $t(220) = 0.31$, $p = .759$, 95% CI [-0.05, 0.07].

Discussion

In Study 2, we experimentally manipulated another aspect of assessment intensity: questionnaire length. Unexpectedly, the questionnaire length groups did not differ in perceived burden (H1) or compliance (H2). In line with our expectations, the within-person variability in momentary mood (but not in extraversion) (H3) and the within-person relationship between state extraversion and momentary mood (H4) were smaller in the long (vs. short) questionnaire group.

Our results are in line with most previous nonexperimental research that found that questionnaire length was unrelated to perceived burden or compliance in an AA study (Jones et al., 2019; Ono et al., 2019; Soyster et al., 2019; Vachon et al., 2019). The only other experimental AA study that we know of that analyzed the effects of manipulated questionnaire length on burden and compliance (Eisele et al., 2020), however, found that longer questionnaires led to higher perceived burden and lower compliance. Whereas the number of items in the experimental groups were similar across studies (30 vs. 60 items per questionnaire in the study by Eisele et al., 2020, and 33 vs. 82 items per questionnaire in our study), the ways in which the greater number of items was achieved differed to some extent across studies: In the study by Eisele et al. (2020), most of the measured constructs were the same across groups, but the long questionnaire group had to answer two additional questions

that referred to the pleasantness of the most important event and the stressfulness of situations since the last questionnaire. These additional questions might have caused participants in the long questionnaire group to think more about their daily negative experiences and therefore could have contributed to the effect that participants in the long (vs. short) questionnaire group perceived the study as more burdensome and showed a lower compliance rate.

Additionally, in the study by Eisele et al. (2020), participants needed to respond within a time frame of 90 s to an AA questionnaire (at random assessment times), whereas in our Study 2, participants had 45 min to respond to an AA questionnaire (at fixed assessment times). Therefore, participants in the study by Eisele et al. might have failed to respond to the questionnaire when they were in a situation that required their full attention (e.g., a conversation, cooking), whereas participants in our Study 2 had the option to simply delay their response by a few minutes in such a situation, thereby maintaining a higher compliance rate. As already discussed in Study 1, another possible explanation for not finding an effect of assessment intensity on compliance in our study is that the personal feedback incentive could have counteracted the decrease in compliance.

The finding that the degree of within-person variability in momentary mood and the within-person relation between state extraversion and mood was smaller in the long (vs. short) questionnaire group is in line with the notion that participants in the long questionnaire group responded in a more heuristic, less nuanced way to the repeated questionnaires (Fuller-Tyszkiewicz et al., 2013; Podsakoff et al., 2019). However, the effect on within-person variability was smaller for state extraversion, and the difference between groups was not significantly different from zero. One possible reason for this difference between momentary mood and state extraversion is that mood was assessed near (or at) the end of the AA questionnaire, whereas state extraversion was assessed at the beginning of the AA questionnaire. In line with this reasoning, research on positioning effects in cross-sectional

surveys (Galesic & Bosnjak, 2009) found lower within-person variance in items that were assessed further away from the beginning of the questionnaire. Taken together, the effects of assessment intensity on burden, within-person variability, and the relation between two time-varying constructs were different between Study 1 (where sampling frequency was manipulated) and Study 2 (where questionnaire length was manipulated). We will come back to differences between these different types of manipulations of assessment intensity in the General Discussion.

General Discussion

The aim of the current paper was to investigate whether differences in assessment intensity have an impact on the aspects of the data from an AA study. To address how assessment intensity was related to perceived burden, compliance, within-person variability, and the within-person relationship between time-varying variables, we used two different experimental manipulations of assessment intensity: sampling frequency (Study 1) and questionnaire length (Study 2). To our knowledge, the present research is the first to study within-person variability and the within-person relationship between time-varying variables as a function of experimentally manipulated assessment intensity in an AA study. Our main findings were that a higher sampling frequency affected only perceived burden but did not affect the other aspects of the AA data we investigated. A longer questionnaire, on the other hand, led to decreased intraindividual variability in momentary mood (but not in state extraversion) and a decreased within-person relationship between momentary mood and state extraversion, but it did not affect perceived burden or compliance.

With respect to compliance as the dependent variable, our experimental results are in line with a large body of previous research that found no impact of sampling frequency and questionnaire length on compliance (Conner & Reid, 2012; Jones et al., 2019; McCarthy et al., 2015; Ono et al., 2019; Soyster et al., 2019; Stone et al., 2003; Walsh & Brinker, 2016).

One exception is Vachon et al.'s (2019) meta-analysis, which found that a higher sampling frequency but not questionnaire length led to a lower compliance. Similar to other studies, we had financially incentivized participants to reach certain levels of compliance in both the low and the high assessment intensity groups. This was done to ensure that the studies followed standard procedures, as financial incentives that are tied to compliance represent a very typical characteristic in AA studies (cf. Trull & Ebner-Priemer, 2020). Future experimental research might investigate the effects that remuneration schedules (e.g., the setting of different thresholds) have on compliance.

With respect to perceived burden (which had not been researched as extensively as compliance in previous research), we had expected that a higher assessment intensity would cause a higher burden, irrespective of the way the difference in objective time needed for study participation was realized (more questionnaires per day, as in Study 1, or more items per questionnaire, as in Study 2). However, as summarized above, these different manipulations of assessment intensity had different effects. One possible explanation for why we found an effect of sampling frequency but not questionnaire length on perceived burden is that participants may be more annoyed by more (vs. less) frequent interruptions in their daily lives than by a longer (vs. shorter) response duration (longer questionnaires). Note that for both questionnaire length groups in Study 2, the sampling frequency was the same as in the low sampling frequency group in Study 1 (i.e., three questionnaires per day). It is conceivable that an effect of questionnaire length on perceived burden shows up only when the study protocol requires a certain number of questionnaires per day (more than three). Given that the study by Eisele et al. (2020) is the only experimental AA study that had manipulated both factors simultaneously (but did not find an interaction between sampling frequency and questionnaire length), future research on the additive and potentially interactive effects of sampling frequency and questionnaire length on perceived burden is needed.

With respect to the degree of within-person variability and relation between time-varying constructs, our results indicated that a longer questionnaire length led to smaller values in these estimates but a higher sampling frequency did not. Note that the length of the questionnaire in Study 1 was similar to the length of the short questionnaire in Study 2 (i.e., around 30 items per questionnaire). That is, we cannot rule out the possibility that an effect of sampling frequency on within-person variability and the relation between time-varying constructs only shows up for longer questionnaires. Clearly, more research on the potential interaction between sampling frequency and questionnaire length on the degree of within-person variability and relation between time-varying constructs is needed. The psychological process behind the differential effect of sampling frequency versus questionnaire length on within-person variability and within-person relations might be that participants in an AA study with a high sampling frequency (but short questionnaires) might implicitly cope differently with the higher assessment intensity than participants in an AA study with many items per questionnaire (but low sampling frequency): If only relatively few items are assessed per measurement occasion, even with a high sampling frequency, participants might be able to produce high data quality (i.e., unbiased responses) by ignoring their perceived burden for the short duration of the questionnaire. On the other hand, long questionnaires might generally reduce participants' effort in responding to the AA questionnaires (particularly at the end of the questionnaires; Galesic & Bosnjak, 2009), thereby reducing the quality of their data (i.e., producing biased responses). However, we can only speculate about how participants cope with the different types of demands an AA study poses. Note that Vachon et al. (2016) and Fuller-Tyszkiewicz et al. (2013) proposed an alternative explanation for a decrease in within-person variability over time. They suggested that over the course of an AA study, individuals' accuracy in reporting momentary experiences increases due to the administration of repeated questionnaires, which might then lead to a reduction in random

variance due to guessing. Further research is needed to replicate our findings and scrutinize the underlying processes.

Limitations

Several limitations have to be considered when interpreting the results of our paper. First, both of our samples were from a student population. We do not know whether certain aspects of our findings depended on the sample characteristics. Therefore, our young, educated samples with large proportions of women restrict the extent to which our findings will generalize to another population. It is possible that the effects of sampling frequency or questionnaire length depend on age or sex, for example. However, we think that our findings should generalize to other student populations that are used in many other AA studies.

Second, it is likely that the motivation of the participants depended on the rewards that were given for participating in the study. For instance, in both of our studies, we offered participants the option to get personal feedback on their answers after the study had been completed. This might have increased participants' motivation to provide more accurate responses or experience the study protocol as less burdensome for individuals who had more interest in the feedback. Additionally, these differences might have been influenced by personal characteristics. For example, participants with high neuroticism might be more interested in tracking momentary mood and state extraversion. Furthermore, we did not compensate the different groups with different rewards regardless of their assessment intensity. Especially in Study 2, the participants in the long questionnaire group were asked to invest more time and energy in participating than those in the short questionnaire group. If this difference between the groups resulted in less motivation to participate (e.g., because the participants thought the reward was not appropriate or they heard that another group in the same study had a shorter questionnaire), this could have resulted in reduced effort while responding to the AA questionnaires, which could have resulted in a reduction in the within-

person variability as well as in the relations between the time-varying constructs. However, we do not know if participants' motivation depended on the rewards that were given. Future research should investigate the effects of rewards on data quantity and quality.

Third, we had limited opportunities to control for careless responders in our sample. We administered one self-reported single item that could indicate invalid responders and a consistency index that could indicate whether some questionnaires were answered inconsistently. If careless responding was not sufficiently controlled for, it could bias the results of our investigation. Furthermore, there might be different types of careless responders in AA studies. Some careless responders might increase the within-person variability, whereas others might decrease the within-person variability. To our knowledge, there are no guidelines for how to deal with careless responders in AA. Future studies should identify the careless responding indices (Meade & Craig, 2012) that are suitable for use in AA studies, identify possible types of careless responders, and establish guidelines for how to deal with careless responders.

Finally, we manipulated only two central aspects of the design in an AA study. However, this leaves many other potential aspects (e.g., study duration, distribution of assessments across the day, type of sampling [time-, interval-, or event-contingent sampling], financial compensation, content of the questions, item difficulty, order of the measured items, and the instructions or the software used to signal the participants) that might have effects on the quantity or quality of AA data. Furthermore, we do not know whether our results can be generalized to other (e.g., higher) sampling densities and other (e.g., longer) questionnaire lengths.

Conclusion

The present research is the first to experimentally manipulate assessment intensity to investigate changes in within-person variability and the within-person relationship between time-varying variables. Furthermore, we found that a higher assessment intensity can affect within-person variability and relations between time-varying constructs without increasing participants' perceived burden. Although further validation of the findings is essential, we hope that future researchers will integrate our findings when planning an AA study.

Footnotes

[1] Study 1 was preregistered on February 1, 2019 with all hypotheses and methods of data analysis. The hypotheses for Study 2 were preregistered in January 2020. Please note that the preregistration documents include hypotheses that were not tested/reported in the present paper. The reason is that testing/reporting all hypotheses would have gone beyond the scope of a single paper. The preregistered hypotheses that were not investigated in the current research will be tested and reported in separate papers. Also note that the test of Hypothesis 3 (on within-person variability) was preregistered for momentary mood as the variable of interest but not for state extraversion. When analyzing the data, we realized that an important piece of information would be missing if we did not test and report the effects on the degree of within-person variability in state extraversion, too. The preregistrations can be found on the OSF pages of the respective studies (Study 1: https://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2; Study 2: https://osf.io/xt3rf/?view_only=6e6bf509c6374759b56faac680c55825).

[2] To define an inconsistency index (Meade & Craig, 2012) for each measurement occasion in an AA study, items that are extremely similar in content and demonstrate a very large (negative or positive) within-person correlation are needed. In our study, bipolar momentary mood items (e.g., for the subscale pleasant-unpleasant mood: good-bad vs. happy-unhappy vs. unpleased-pleased vs. unwell-well; within-person intercorrelations across all subscales ranged from $r = |.55|$ to $|.73|$) met these criteria. We defined inconsistent responding at a particular measurement occasion as illogical responses across mood item pairs with responses near (or at) the extremes of the scale (Categories 1 or 2 vs. 6 or 7). For example, response patterns, such as feeling "*very* happy" and "*very* unwell" at the same time or feeling "*very* happy" and "*very* bad" at the same time were categorized as inconsistent responses.

More information about the momentary mood items can be found on the OSF page of Study 1 (https://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2).

[3] Only the relevant scales for the analyses used in this investigation are described. An overview of all measured constructs can be found on the OSF page of this project (https://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2).

[4] Note that independent *t* tests (and not multilevel regression analyses) had been preregistered for testing Hypothesis 1 (on daily perceived burden) and Hypothesis 2 (on compliance). However, for reasons of consistency and to avoid the need to aggregate values by hand, we switched to multilevel analyses during the peer review process. To remain close to the preregistered data analytic method tests, we additionally report means and Cohen's *d* values along with the results of the multilevel regressions.

[5] We performed additional exploratory analyses on the linear effect of the day of the study on daily perceived burden, with the day of the study centered on the 4th day and investigated the interaction between time (the day of the study) and sampling frequency. We conducted the data analytic steps correspondingly for Hypothesis 4. In the final model (with the interaction term), the main effect of the day of the study was significant, $t(239) = 3.69$, $p < .001$, 95% CI [0.03, 0.09]. The cross-level interaction between sampling frequency and the day of the study was significant, $t(240) = -3.52$, $p < .001$, 95% CI [−0.12, −0.04], which means that the low sampling frequency group ($b = 0.06$) had a larger regression coefficient than the high sampling frequency group ($b = -0.02$).

[6] To estimate the effect size of this test, we person-centered momentary mood or state extraversion and aggregated the variances for each person to get a value for the within-person variance. This allowed us to estimate Cohen's *d* for the effect size of this analysis.

[7] We defined inconsistent responding at a particular measurement occasion in the same way as we did in Study 1. However, in Study 2, there were only two momentary mood items

(within the same subscale) that were presented to both experimental groups (short vs. long questionnaire), and hence, the inconsistency index was based on these two items from the momentary pleasant-unpleasant mood subscale (the within-person intercorrelation was r = -.73). More information about the momentary mood items can be found on the OSF page of Study 2 (https://osf.io/xt3rf/?view_only=6e6bf509c6374759b56faac680c55825).

[8] Only the relevant scales for the analyses used in this investigation are described. An overview of all measured constructs can be found on the OSF page of this project (https://osf.io/xt3rf/?view_only=6e6bf509c6374759b56faac680c55825).

[9] We conducted additional exploratory analyses on the linear effect of the day of the study on daily perceived burden, with the day of the study centered on the midpoint of the assessment duration (7.5) and investigated the interaction between time (the day of the study) and questionnaire length. We conducted the data analytic steps correspondingly for Hypothesis 4. In the final model (with the interaction term), the main effect of the day of the study was significant, $t(238) = 6.44$, $p < .001$, 95% CI [0.03, 0.06]. The cross-level interaction between questionnaire length and the day of the study was significant, $t(241) = -2.57$, $p = .011$, 95% CI [−0.04, −0.01], which means that the short questionnaire group ($b = 0.04$) had a larger regression coefficient than the long questionnaire group ($b = 0.02$).

[10] To estimate the effect size of this test, we person-centered momentary mood or state extraversion and aggregated the variances for each person to get a value for the within-person variance. This allowed us to estimate Cohen's d for the effect size of this analysis.

Declarations

Funding

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant GRK2277 (Research Training Group "Statistical Modeling in Psychology").

Conflicts of interest

The author(s) declare that they have no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Ethics approval

The questionnaires and methodologies for both studies were approved by the psychological Ethics Committee at the University Koblenz-Landau, Germany (Ethics approval number for Study 1: 170 and for Study 2: 228).

Consent to participate and consent for publication

Informed consent was obtained from all individual participants included in the study. This consent informed all individual participants regarding publishing their data.

Availability of data, code, and materials

The data sets generated during the current studies and the respective codes are available on the OSF repository (Study 1: https://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2; Study 2: https://osf.io/xt3rf/?view_only=6e6bf509c6374759b56faac680c55825). Studies 1 and 2 were preregistered (see respective URLs).

Authors' contributions

Not applicable

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Beal, D. J. (2015). ESM 2.0: State of the Art and Future Potential of Experience Sampling Methods in Organizational Research. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 383–407. doi: 10.1146/annurev-orgpsych-032414-111335

Benjamini, Y., & Hochberg, Y. (1995). Controlling The False Discovery Rate—A Practical And Powerful Approach To Multiple Testing. *J. Royal Statist. Soc., Series B*, *57*, 289–300. doi: 10.2307/2346101

Carpenter, R. W., Wycoff, A. M., & Trull, T. J. (2016). Ambulatory Assessment: New Adventures in Characterizing Dynamic Processes. *Assessment*, *23*(4), 414–424. doi: 10.1177/1073191116632341

Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, *66*(9), 1261–1266. doi: 10.1016/j.jbusres.2012.02.023

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. doi: 10.1027/1614-0001/a000149

Conner, T. S., & Reid, K. A. (2012). Effects of Intensive Mobile Happiness Reporting in Daily Life. *Social Psychological and Personality Science*, *3*(3), 315–323. doi: 10.1177/1948550611419677

Courvoisier, D. S., Eid, M., & Lischetzke, T. (2012). Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics. *Psychological Assessment*, *24*(3), 713–720. doi: 10.1037/a0026733

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*(5), 478–491. doi: 10.1038/s41562-019-0555-0

Dowling, N. M., Raykov, T., & Marcoulides, G. A. (2018). Examining Population Differences in Within-Person Variability in Longitudinal Designs Using Latent Variable Modeling: An Application to the Study of Cognitive Functioning of Older Adults. *Educational and Psychological Measurement*, 1–12. doi: 10.1177/0013164418758834

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*, 107319112095710. doi: 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*.

Fleeson, W., Malanos, A. B., & Achille, N. M. (2002). An intraindividual process approach to the relationship between extraversion and positive affect: Is acting extraverted as "good" as being extraverted? *Journal of Personality and Social Psychology*, *83*(6), 1409–1422. doi: 10.1037/0022-3514.83.6.1409

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. doi: 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. doi: 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. doi: 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. doi: 10.1177/0963721416666518

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus, 3rd ed.* (pp. xix, 440). New York, NY, US: Routledge/Taylor & Francis Group. doi: 10.4324/9781315746494

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications* (1st ed.). Routledge Academic. doi: 10.4324/9781410604118

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction*, *114*(4), 609–619. doi: 10.1111/add.14503

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

LeBeau, B., Song, Y. A., & Liu, W. C. (2018). Model Misspecification and Assumption Violations With the Linear Mixed Model: A Meta-Analysis. *SAGE Open*, *8*(4), 215824401882038. doi: 10.1177/2158244018820380

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. doi: 10.1016/j.jrp.2012.04.002

Liu, H., Xie, Q. W., & Lou, V. W. Q. (2019). Everyday social interactions and intra-individual variability in affect: A systematic review and meta-analysis of ecological momentary assessment studies. *Motivation and Emotion, 43*(2), 339–353. doi: 10.1007/s11031-018-9735-x

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. doi: 10.1016/j.jpain.2018.01.006

McCabe, K. O., & Fleeson, W. (2016). Are traits useful? Explaining trait manifestations as tools in the pursuit of goals. *Journal of Personality and Social Psychology, 110*(2), 287–301. doi: 10.1037/a0039490

McCarthy, D. E., Minami, H., Yeh, V. M., & Bold, K. W. (2015). An experimental investigation of reactivity to ecological momentary assessment frequency among adults trying to quit smoking: Reactivity to ecological momentary assessment. *Addiction, 110*(10), 1549–1560. doi: 10.1111/add.12996

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

McNiel, J. M., & Fleeson, W. (2006). The causal effects of extraversion on positive affect and neuroticism on negative affect: Manipulating state extraversion and state neuroticism in an experimental approach. *Journal of Research in Personality, 40*(5), 529–550. doi: 10.1016/j.jrp.2005.05.003

McNiel, J. M., Lowman, J. C., & Fleeson, W. (2010). The effect of state extraversion on four types of affect. *European Journal of Personality, 24*(1), 18–35. doi: 10.1002/per.738

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

Mehl, M. R., & Conner, T. S. (Eds.). (2014). *Handbook of research methods for studying daily life* (Paperback ed). New York, NY: Guilford.

Morren, M., Dulmen, S., Ouwerkerk, J., & Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: A systematic review. *European Journal of Pain*, *13*(4), 354–365. doi: 10.1016/j.ejpain.2008.05.010

Moskowitz, D. S., Russell, J. J., Sadikaj, G., & Sutton, R. (2009). Measuring people intensively. *Canadian Psychology/Psychologie Canadienne*, *50*(3), 131–140. doi: 10.1037/a0016625

Muthén, L. K., & Muthén, B. O. (1998). *Mplus User's Guide* (8th edition). Los Angeles, CA: Muthén & Muthén.

Napa Scollon, C., Prieto, C.-K., & Diener, E. (2009). Experience Sampling: Promises and Pitfalls, Strength and Weaknesses. In E. Diener (Ed.), *Assessing Well-Being* (pp. 157–180). Dordrecht: Springer Netherlands. doi: 10.1007/978-90-481-2354-4_8

Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What Affects the Completion of Ecological Momentary Assessments in Chronic Pain Research? An Individual Patient Data Meta-Analysis. *Journal of Medical Internet Research*, *21*(2), e11398. doi: 10.2196/11398

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. doi: 10.1177/1073191119839138

Ottenstein, C., & Werner, L. (2021). *Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors [Manuscript submitted for publication]*.

Phillips, M. M., Phillips, K. T., Lalonde, T. L., & Dykema, K. R. (2014). Feasibility of text messaging for ecological momentary assessment of marijuana use in college students. *Psychological Assessment*, *26*(3), 947–957. doi: 10.1037/a0036612

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

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Thousand Oaks: Sage Publications.

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. doi: 10.1037/pas0000662

Rosen, C. C., Koopman, J., Gabriel, A. S., & Johnson, R. E. (2016). Who strikes back? A daily investigation of when and why incivility begets incivility. *Journal of Applied Psychology*, *101*(11), 1620–1634. doi: 10.1037/apl0000140

Santangelo, P. S., Ebner-Priemer, U. W., & Trull, T. J. (2013). *Experience Sampling Methods in Clinical Psychology*. Oxford University Press. doi: 10.1093/oxfordhb/9780199793549.013.0011

Sitzmann, T., & Yeo, G. (2013). A Meta-Analytic Investigation of the Within-Person Self-Efficacy Domain: Is Self-Efficacy a Product of Past Performance or a Driver of Future Performance?: PERSONNEL PSYCHOLOGY. *Personnel Psychology*, *66*(3), 531–568. doi: 10.1111/peps.12035

Sonnentag, S., Binnewies, C., & Mojza, E. J. (2008). "Did you have a nice evening?" A day-level study on recovery experiences, sleep, and affect. *Journal of Applied Psychology*, *93*(3), 674–684. doi: 10.1037/0021-9010.93.3.674

Soyster, P. D., Bosley, H. G., Reeves, J. W., Altman, A. D., & Fisher, A. J. (2019). Evidence for the Feasibility of Person-Specific Ecological Momentary Assessment Across Diverse Populations and Study Designs. *Journal for Person-Oriented Research*, *5*(2), 53–64. doi: 10.17505/jpor.2019.06

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Mehrdimensionaler Befindlichkeitsfragebogen*. Göttingen: Hogrefe.

Stone, A. A., Broderick, J. E., Schwartz, J. E., Shiffman, S., Litcher-Kelly, L., & Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: Reactivity, compliance, and patient satisfaction: *Pain*, *104*(1), 343–351. doi: 10.1016/S0304-3959(03)00040-X

Thayer, R. E. (1978). Toward a psychological theory of multidimensional activation (arousal). *Motivation and Emotion*, *2*(1), 1–34. doi: 10.1007/BF00992729

Trougakos, J. P., Beal, D. J., Green, S. G., & Weiss, H. M. (2008). Making the Break Count: An Episodic Examination of Recovery Activities, Emotional Experiences, and Positive Affective Displays. *Academy of Management Journal*, *51*(1), 131–146. doi: 10.5465/amj.2008.30764063

Trull, T. J., & Ebner-Priemer, U. (2014). The Role of Ambulatory Assessment in Psychological Science. *Current Directions in Psychological Science*, *23*(6), 466–470. doi: 10.1177/0963721414550706

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, *129*(1), 56–63. doi: 10.1037/abn0000473

Vachon, H., Bourbousson, M., Deschamps, T., Doron, J., Bulteau, S., Sauvaget, A., & Thomas-Ollivier, V. (2016). Repeated self-evaluations may involve familiarization: An exploratory study related to Ecological Momentary Assessment designs in patients with major depressive disorder. *Psychiatry Research*, *245*, 99–104. doi: 10.1016/j.psychres.2016.08.034

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental Disorders: Meta-Analysis and Recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. doi: 10.2196/14475

Walsh, E., & Brinker, J. K. (2016). Temporal Considerations for Self-Report Research Using Short Message Service. *Journal of Media Psychology*, *28*(4), 200–206. doi: 10.1027/1864-1105/a000161

Watson, D. (2000). *Mood and temperament*. New York, NY, US: Guilford Press.

Table 1

Previous Studies on the Effects of Design Features of AA Studies on Burden, Compliance, Within-Person Variability, and Within-Person Relations Between Variables

Article	Study design	Pop.	Design feature (factor levels or range)	DVs	Results for (higher) SF	Results for (higher) QL
Conner & Reid (2012)	Exp. AA	NC	SF (1 vs. 3 vs. 6 qu./day)	Com	No effect on Com	—
Eisele et al. (2020)	Exp. AA	NC	SF (3 vs. 6 vs. 9 qu./day), QL (30 vs. 60 items)	Bur, Com	No effect on Bur, no effect on Com	Increase of Bur
Jones et al. (2019)	Meta/pooled	C	SF (1-9 qu./day)	Com	No effect on Com	—
McCarthy et al. (2015)	Exp. AA	NC	SF (1 vs. 6 qu./day)	Com	No effect on Com	—
Morren et al. (2009)	Meta/pooled	C, NC	SF (1-10 qu./day) QL (1-63 items)	Com	No effect on Com	Related to lower Com
Ono et al. (2019)	Meta/pooled	C	SF (3-12 qu./day), QL (6-63 items)	Com	No effect on Com	No effect on Com
Ottenstein & Werner (2021)	Meta/pooled	C, NC	SF (0.14-44 qu./day), QL (1-150 items)	Com	No effect on Com	No effect on Com
Podsakoff et al. (2019)	Meta/pooled	NC	—	WPV	Related to larger WPV	—
Rintala et al. (2019)	Meta/pooled	C, NC	SF (10 qu./day), QL (42-52 items)	Com	—	No effect on Com

Soyster et al. (2019)	Meta/pooled	C	SF (4 or 8 qu./day), QL (16-40 items)	Com	—	No effect on Com
Stone et al. (2003)	Exp. AA	C	SF (3 vs. 6 vs. 12 qu./day)	Bur, Com	Increase of Bur, no effect on Com	—
Vachon et al. (2019)	Meta/pooled	C	—	Com	Related to lower Com	No effect on Com
Walsh & Brinker (2016)	Exp. AA	NC	SF (20 items across 1 or 2 days)	Com	No effect on Com	—

Note. Pop = Population under study; DV = Dependent variable(s); Exp. AA = Experimental AA study; Meta/pooled = Meta-analysis or pooled data analysis; C = Clinical sample; NC = Nonclinical sample; SF = Sampling frequency; QL = Questionnaire length; qu = Questionnaire; Bur = Burden; Com = Compliance; WPV = Within-person variability.

Table 2

Sampling Scheme of Study 1

Time of day	Experimental group	
	Low sampling frequency	High sampling frequency
9:00-10:40	Questionnaire 1	Questionnaire 1
11:00-13:50		Questionnaires 2-4
14:10-15:50	Questionnaire 2	Questionnaire 5
16:10-19:00		Questionnaires 6-8
19:20-21:00	Questionnaire 3	Questionnaire 9

Note. The displayed sampling scheme refers to the first of the two time schedules from which participants could choose (9:00-21:00 vs. 10:30-22:30). That is, in the second time schedule, each questionnaire was scheduled 90 min later. For each questionnaire, participants had the option to delay their response for up to 15 min. In the high sampling frequency group, Questionnaires 2, 3, and 4 and Questionnaires 6, 7, and 8 were at least 28 min apart.

Table 3

Descriptive Statistics and Bivariate Correlations for the Main Variables Presented Separately for Each Experimental Group (Study 1)

Group	Variable	1	2	3	4
Low sampling frequency	1. Pleasant-unpleasant mood	—	.24***	-.03	
	2. State extraversion	.33***	—	.12**	
	3. Daily perceived burden	-.21**	-.12	—	
	4. Retrospective perceived burden	-.12	.05	.72***	—
	M	5.12	4.31	2.00	2.20
	SD_{within}	0.92	1.42	0.72	-
	$SD_{between}$	0.80	0.70	0.54	0.75
	$N_{persons}$[a]	153	151	149	93
	$N_{questionnaires}$[b]	2,295	1,794	788	
High sampling frequency	1. Pleasant-unpleasant mood	—	.28***	-.10**	
	2. State extraversion	.25**	—	.04	
	3. Daily perceived burden	-.31***	.13	—	
	4. Retrospective perceived burden	-.14	.16	.82***	—
	M	5.00	4.19	2.56	2.82
	SD_{within}	0.87	1.40	0.71	-
	$SD_{between}$	0.79	0.71	0.59	0.75
	$N_{persons}$[a]	160	160	154	101
	$N_{questionnaires}$[b]	2,281	1,769	791	

Note. Between-person correlations are presented below the diagonal. Within-person correlations between the daily measures are presented above the diagonal. All *p*-values are two-sided *p*-values. For all daily measures, we extracted the mean (intercept) and standard deviation from the multilevel null model of the respective variable.

[a]*N* differed between momentary mood and state extraversion because of the "*not applicable*" response option in state extraversion. [b]*N* differed because the respective variables were assessed on different measurement occasions.

*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4

Multilevel Model (Fixed Effects) Predicting Momentary Mood by State Extraversion and Sampling Frequency (Study 1)

Model Predictor	Estimate	*SE*	*df*	*t*
Model 1:				
Intercept	5.16			
State extraversion	0.18	0.02	199.2	11.84***
Sampling frequency	-0.12	0.10	297.0	-1.24
Model 2:				
Intercept	5.17			
State extraversion	0.16	0.02	193.6	7.43***
Sampling frequency	-0.14	0.10	299.0	-1.44
State Extraversion x Sampling Frequency	0.04	0.03	198.5	1.35

Note. State extraversion was centered at the person mean. Sampling frequency was coded as 0 = low sampling frequency group and 1= high sampling frequency group.

***$p < .001$.

Table 5

Descriptive Statistics and Bivariat e Correlations for the Main Variables Presented Separately for Each Experimental Group (Study 2)

Group	Variable	1	2	3	4
Short questionnaire	1. Pleasant-unpleasant mood	—	.35***	-.10***	
	2. State extraversion	.42***	—	-.02	
	3. Daily perceived burden	-.30***	-.24**	—	
	4. Retrospective perceived burden	-.24**	-.19*	.83***	—
	M	4.90	3.13	2.40	2.71
	SD_{within}	1.09	0.59	0.76	-
	$SD_{between}$	0.79	0.29	0.63	0.85
	$N_{persons}$	142	142	139	118
	$N_{questionnaires}$[a]	4,411	4,411	1,500	
Long questionnaire	1. Pleasant-unpleasant mood	—	.27***	-.09**	
	2. State extraversion	.38***	—	.05	
	3. Daily perceived burden	-.32***	-.06	—	
	4. Retrospective perceived burden	-.17	-.02	.81***	—
	M	5.01	3.12	2.51	2.77
	SD_{within}	1.00	0.56	0.78	-
	$SD_{between}$	0.72	0.31	0.71	0.92
	$N_{persons}$	140	140	133	117
	$N_{questionnaires}$[a]	4,174	4,174	1,407	

Note. Between-person correlations are presented below the diagonal. Within-person correlations between the daily measures are presented above the diagonal. All *p*-values are two-sided *p*-values. For all daily measures, we extracted the mean (intercept) and standard deviation from the multilevel null model of the respective variable.

[a]*N* differed because the respective variables were assessed on different measurement occasions.

*$p < .05$. **$p < .01$. ***$p < .001$.

Table 6

Multilevel Model (Fixed Effects) Predicting Momentary Mood by State Extraversion and Questionnaire Length (Study 2)

Model Predictor	Estimate	*SE*	*df*	*t*
Model 1:				
Intercept	4.92			
State extraversion	0.56	0.03	242.2	17.72***
Questionnaire length	0.16	0.09	268.2	1.67
Model 2:				
Intercept	4.90			
State extraversion	0.65	0.04	233.0	14.99***
Questionnaire length	0.20	0.09	268.6	2.01*
State Extraversion x Questionnaire Length	-0.19	0.06	243.6	-2.98**

Note. State extraversion was centered at the person mean. Questionnaire length was coded as 0 = short questionnaire group and 1= long questionnaire group.
*$p < .05$. **$p < .01$. ***$p < .001$.

3 Paper 2: Careless Responding

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

Modeling Careless Responding in Ambulatory Assessment Studies Using Multilevel Latent Class Analysis: Factors Influencing Careless Responding

Kilian Hasselhorn, Charlotte Ottenstein, and Tanja Lischetzke

University of Koblenz-Landau

Author Note

Kilian Hasselhorn's contribution was supported by grant GRK 2277 (Research Training Group "Statistical Modeling in Psychology") from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

The present analyses were preregistered on the OSF. Moreover, all data, analysis code, and research materials are available on the OSF repository: (https://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2).

Results from the present data set were previously published to test different research questions: Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2021). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01683-6

Ideas and results appearing in the manuscript were presented at the 15th Meeting of the Methods and Evaluation Division of the German Psychological Association in Hildesheim, Germany, 2021, the Society of Ambulatory Assessment Conference 2022, and the Network Meeting „Multilevel Modelling in Method and Applied Research" at the University of Zurich, Switzerland, in 2022.

Correspondence concerning this article should be addressed to Kilian Hasselhorn, Department of Psychology, University of Koblenz-Landau, Fortstr. 7, 76829 Landau, Germany. E-mail: hasselhorn@uni-landau.de, Tel: +49 6341 280-31298

Abstract

As the number of studies using ambulatory assessment (AA) has been increasing across diverse fields of research, so has the necessity to identify potential threats to AA data quality such as careless responding. To date, careless responding has primarily been studied in cross-sectional surveys. The goal of the present research was to identify latent profiles of momentary careless responding on the occasions level and latent classes of individuals (who differ in the distribution of careless responding profiles across occasions) on the person level using multilevel latent class analysis (ML-LCA). We discuss which of the previously proposed indices seem promising for investigating careless responding in AA studies, and we show how ML-LCA can be applied to model careless responding in intensive longitudinal data. We used data from an AA study in which the sampling frequency (3 vs. 9 occasions per day, 7 days, n = 310 participants) was experimentally manipulated. We tested the effect of sampling frequency on careless responding using multigroup ML-LCA and investigated situational and respondent-level covariates. The results showed that four Level 1 profiles ("careful", "slow", two types of "careless" responding) and four Level 2 classes ("careful", "frequently careless", two types of "infrequently careless" respondents) could be identified. Sampling frequency did not have an effect on careless responding. On the person (but not the occasion) level, motivational variables were associated with careless responding. We hope that researchers might find the application of an ML-LCA approach useful to shed more light on factors influencing careless responding in AA studies.

Translational Abstract

As the number of studies using ambulatory assessment (AA; also termed ecological momentary assessment or experience sampling method) has been increasing across diverse fields of research, so has the necessity to identify potential threats to AA data quality such as careless responding. To date, careless responding has primarily been studied in cross-sectional surveys. We discuss which of the previously proposed indices seem promising for investigating careless responding in AA studies. The goal of the present research was to identify types of occasions, which differ in the extent and type of momentary careless responding (latent profiles), and types of individuals (latent classes) who differ in the distribution of those types of occasions across measurement occasions. We used data from an AA study in which the sampling frequency (3 vs. 9 occasions per day, 7 days, n = 310 participants) was experimentally manipulated. We tested the effect of sampling frequency on careless responding and investigated situational and respondent-level covariates. The results showed that four types of measurement occasions ("careful", "slow", two types of "careless" responding) and four types of individuals ("careful", "frequently careless", two types of "infrequently careless" respondents) could be identified. Sampling frequency did not have an effect on careless responding. On the individual (but not the occasion) level, motivational variables were associated with careless responding. We hope that researchers might find our approach useful to shed more light on factors influencing careless responding in AA studies.

Keywords: ambulatory assessment, careless responding, multigroup multilevel latent class analysis, latent profile analysis

Modeling Careless Responding in Ambulatory Assessment Studies Using Multilevel Latent Class Analysis: Factors Influencing Careless Responding

Parallel to the technical development of mobile devices and smartphones, researchers' interest in conducting ambulatory assessment (AA) studies has greatly increased over the last decade in diverse fields of research (e.g., Hamaker & Wichers, 2017; Jaso et al., 2021). AA (also called ecological momentary assessment or experience sampling) is a data collection method used to study individuals' daily life experiences and environmental features in naturalistic and unconstrained settings (Bolger & Laurenceau, 2013; Fahrenberg, 2006; Larson & Csikszentmihalyi, 1983). AA yields intensive longitudinal data and allows researchers to investigate within-person dynamics and individual differences in these within-person dynamics (Hamaker & Wichers, 2017) while ensuring high ecological validity and reduced recall bias (Mehl & Conner, 2014; Trull & Ebner-Priemer, 2014).

As the number of AA studies increases, so does the necessity to identify potential threats to the quality of the intensive longitudinal data that are obtained. One potential threat is careless responding (also called insufficient effort responding), which might compromise the psychometric properties of measurement instruments and potentially bias the correlations between substantive measures (Goldammer et al., 2020; Huang et al., 2015; McGrath et al., 2010). Careless responding refers to response behavior that is characterized by responding to items without sufficient regard to the item content (Huang et al., 2012; Meade & Craig, 2012). In cross-sectional research, careless responding is usually identified by applying either a multiple hurdle approach (e.g., Curran, 2016) or a latent class analysis (LCA) approach (e.g., Meade & Craig, 2012). Both approaches use multiple indices to identify careless responses (Goldammer et al., 2020). In the multiple hurdle approach, a cut-off score for each index is defined, and a participant is classified as a careful responder if they are able to "jump" all the "hurdles." In the LCA approach, the careless responding indices serve as

observed indicators of latent classes of individuals with different patterns of careless responding. By using an LCA approach, researchers can identify careful versus careless responders without having to define cut-off scores for careless responding indices and can potentially differentiate between different types of careless responders (e.g., inconsistent responding vs. invariable responding). To date, however, research on careless responding has focused almost exclusively on cross-sectional data, and there are no guidelines or best practices for identifying careless responses or careless responders in AA (Jaso et al., 2021; van Berkel et al., 2018). Recently, Jaso et al. (2021) investigated careless responding in AA with the multiple hurdle approach and recommended specific data-cleaning protocols that are based on cutoff scores.

To our knowledge, no study to date has applied the LCA approach to investigate careless responding in AA data. Thus, the goal of the present paper was to analyze careless responding in AA data, which is nested in structure (measurement occasions nested in persons), and identify careless versus careful responding (at the level of measurement occasions) and different types of careless versus careful responders (at the person level) using a multilevel LCA (ML-LCA) approach. In addition, we aimed to investigate whether (a) a design factor (sampling frequency), (b) occasion-level (situational) factors, and (c) respondent-level factors (e.g., personality traits) influence the degree of careless responding in an AA study.

In the recent past, an increasing number of researchers have conducted methodological research on the effects of design-related characteristics on aspects of AA data (e.g., data quantity and data quality) by applying an experimental design. These experimental AA studies have focused primarily on participant burden, compliance, sampling contingency, sampling schedule, within-person variability, and within-person relationships as outcome variables (e.g., Eisele et al., 2020; Hasselhorn et al., 2021; Himmelstein et al., 2019; van

Berkel et al., 2019). To our knowledge, the effect of AA design features on careless responding has been analyzed in only one study to date: Eisele et al.'s (2020) results suggested that a higher (vs. lower) sampling frequency has no effect on careless responding, whereas a longer (vs. shorter) questionnaire (at each measurement occasion) increases careless responding. One limitation, however, is that the authors relied exclusively on self-report measures of careless responding, which depends on participants' ability and willingness to report on their response effort in the study. Hence, it still remains an open question whether careless responding that is identified through the use of unobtrusive indices is influenced by an AA study's design features (in particular, sampling frequency).

Previous research has shown that the Big Five personality traits are associated with careless responding (Bowling et al., 2016; Grau et al., 2019). Whereas conscientiousness, extraversion, and agreeableness were negatively related to careless responding, neuroticism was positively related to careless responding, and openness was unrelated to careless responding. Moreover, Bowling et al. (2016) argued that situational factors (e.g., momentary motivation to answer the questions in the current measurement occasion and momentary time pressure) may influence careless responding. To our knowledge, the present research is the first to investigate the relationships between respondent-level and occasion-level (situational) covariates and careless responding in an AA study.

The remaining Introduction is structured as follows: First, we summarize the careless responding indices that have been proposed to identify careless respondents in cross-sectional surveys in previous research. Second, we discuss which of the careless responding indices that have been proposed for cross-sectional surveys seem promising for investigating careless responding in AA studies. Third, we briefly present findings from studies using the LCA approach to model careless responding in cross-sectional data. Fourth, we show how ML-LCA (Vermunt, 2003, 2008) can be applied to model careful versus careless responding on

Level 1 (occasions) and careful versus careless respondents on Level 2 (persons) and how this model can be extended into a multigroup ML-LCA for designs in which multiple experimental groups are studied (in our case, high vs. low sampling frequencies). Finally, we summarize our hypotheses and research questions on the effects of a design feature (sampling frequency), situational factors (time-varying variables), and respondent-level factors (time-invariant variables) on careless responding in an AA study. Note that we refer to the term research questions if we did not have any theoretical basis for the expected effects of the variables.

Careless Responding Indices in Cross-Sectional Research

Cross-sectional research has identified a variety of indices for detecting and removing careless responders and consequently improving data quality (Curran, 2016; Meade & Craig, 2012). These careless responding indices can be divided into two groups: direct and indirect measures. Direct measures of careless responding (e.g., self-reported effort or instructed response items) have to be included as additional items during data collection (Goldammer et al., 2020). Direct measures have been criticized because they can annoy participants or obstruct the flow of the questionnaire and also because participants' responses to the items may be invalid (DeSimone et al., 2018; Edwards, 2019). Indirect measures of careless responding are based on information that is obtained during participants' response process and require additional analyses to be computed after data collection (Goldammer et al., 2020). In the group of indirect measures, multiple careless responding indices have been proposed. These can be classified into three different types: consistency indices, response pattern indices, and response time indices.[1] Each type is aimed at capturing a certain pattern of

[1] We used the categorization of careless responding indices by Goldammer et al. (2020) with three indirect indices. However, indirect careless responding indices have been heterogeneously categorized in the existing literature. Note that, for example, one might consider *outlier indices* as a fourth type of indirect indices (e.g., Curran, 2016). Outlier indices, such as Mahalanobis distance, are used to identify response patterns that strongly deviate from the response pattern of the overall sample (Goldammer et al., 2020; Meade & Craig, 2012). However, the properties of Mahalanobis distance,

careless responding that might not be captured by other types (Curran, 2016). *Consistency indices* are based on the assumption that respondents who give careful responses will provide similar responses across very similar items, consequently providing a response pattern that is internally consistent (Curran, 2016). A lack of internally consistent responses within individuals is denoted as inconsistent responding (or random responding). *Response pattern indices* (also called invariability indices) are based on the assumption that participants who respond carefully will choose different response categories across dissimilar (and often theoretically distinct) items (Curran, 2016; Goldammer et al., 2020; Meade & Craig, 2012). Thus, participants who choose the same response category across all the items on a questionnaire, for example, are identified as careless (long string) responders. *Response time indices* are based on the assumption that very fast responses or responders may be careless (Jaso et al., 2021; Meade & Craig, 2012). For more detailed information on specific indices, see, for example, Curran (2016) or Niessen et al. (2016).

In cross-sectional research, a selection of these indices is used to identify careless responding. However, there are no guidelines regarding which are the "necessary" or "best" indices. Therefore, researchers typically try to use a variety of different indices that capture different types of careless responding. For example, Goldammer et al. (2020) used a total of seven indirect indices, which included several consistency indices, a response pattern index, and a response time index, to identify careless responding. Similarly, Meade and Craig (2012) used a total of 12 careless responding indices, which were a combination of direct and indirect measures with at least one index of each of the types of indirect measures. Taken together, the use of careless responding indices is essential for analyzing careless responding.

Careless Responding Indices in AA Studies

---

which is often used as the only outlier index, have yet to be fully investigated. Thus, the extent to which this index should be used to identify careless responses has yet to be determined (Curran, 2016; Niessen et al., 2016), and it might also capture inconsistent responding (Goldammer et al., 2020).

To our knowledge, the only study to systematically investigate indirect careless responding indices in AA is the study by Jaso et al. (2021). The authors defined three indices on the occasion level: a response pattern index (proportion of items at the mode), a response time index (time to complete a measurement occasion), and another index (standard deviation across items within a measurement occasion) that could be considered a response pattern index (because it identifies extremely similar responses across items as careless responses). However, there might be more useful indices that were developed for (longer) cross-sectional online surveys but that could be adapted to AA data—that is, intensive longitudinal data with measurement occasions nested in persons.

Some consistency indices (e.g., the even-odd consistency index or the polytomous Guttman errors) require relatively long questionnaires (or subscales), or they require additional criteria (e.g., that an item response theory model fits the data) to be fulfilled (Curran, 2016). Therefore, it is not clear whether these consistency indices can maintain their viability and applicability in AA data. However, the *person-total correlation* and the *semantic* (or *psychometric*) *antonyms* indices seem promising for capturing inconsistent responses in AA data. In cross-sectional surveys, the person-total correlation is "a measure of how consistently each person is acting like all other persons" (Curran, 2016, p. 9). As an equivalent measure that can be used in AA data, the occasion-person correlation can measure each participant's consistency in responding to a specific measurement occasion compared with how this person responded on average across all measurement occasions. The semantic (or psychometric) antonym index measures how internally/logically inconsistent a person's responses are to items that have opposite meanings (or demonstrate a very large negative correlation; Curran, 2016). This concept can be directly translated into AA data if suitable items are assessed at each occasion.

Response pattern indices can be directly translated into AA data. However, there are differences to consider when using these indices in AA questionnaires (compared with cross-sectional surveys), which typically use short questionnaires with only a few items per subscale. Response pattern indices are usually computed across all items for each survey page. In AA, the items are usually presented on a mobile phone screen in such a way that participants can complete all the items without having to scroll down (too much). Consequently, response pattern indices have to be computed across fewer items for each questionnaire page. Therefore, the total difference between careful and careless responses is expected to be smaller in AA data (compared with cross-sectional data).

Response time indices can also be directly translated into AA data. However, participants' response times might vary across measurement occasions because the number of items presented at any particular measurement occasion might vary (e.g., depending on the time of day). Additionally, response times might systematically decrease across the duration of the study because participants become more and more familiar with the items over time.

Latent Class Analysis Approach to Careless Responding in Cross-Sectional Research

LCA identifies subtypes of observation units that show similar patterns of scores on (categorical or continuous) observed indicators. When the observed indicators are categorical, these subtypes are typically called *latent classes*, and when the observed indicators are continuous, these subtypes are typically called *latent profiles* (e.g., Masyn, 2013). To study careful versus careless responding using LCA, researcher-defined careless responding indices (e.g., the ones summarized above) serve as observed variables in the LCA model. The LCA approach has been used successfully to identify careless responding in cross-sectional surveys (Goldammer et al., 2020; Kam & Meyer, 2015; Li et al., 2020; Maniaci & Rogge, 2014; McKibben & Silvia, 2017; Meade & Craig, 2012; Paas et al., 2018; Schneider et al., 2018). Previous research has repeatedly shown that participants can be assigned to one of three

classes, with one class representing careful responders and two classes representing careless responders (Goldammer et al., 2020; Kam & Meyer, 2015; Li et al., 2020; Maniaci & Rogge, 2014; Meade & Craig, 2012). Usually, one careless responding class represents long string (i.e., invariable) responses, whereas the other careless responding class represents inconsistent responses. Note that in some studies, a three-class model was not chosen as the best fitting model because the third class was too small (about 2%; Schneider et al., 2018) or the sample size was too small to identify and discriminate between small classes (McKibben & Silvia, 2017). To our knowledge, no study has analyzed whether different types of careless responding classes similar to the ones that have been identified in cross-sectional surveys (e.g., long string vs. inconsistent responding) can be identified in AA data.

Extending the LCA Approach to Model Careless Responding in AA Data (Multilevel LCA, Multigroup Multilevel LCA)

Standard LCA assumes that observations are independent of one another. In nested data resulting from two-stage cluster sampling (i.e., individuals nested in clusters) or longitudinal studies (i.e., occasions nested in persons), this assumption is violated. To account for nested data, Vermunt (2003, 2008) proposed an ML-LCA framework in which the probability that a Level 1 unit belongs to a certain latent class can vary across Level 2 units. In the parametric approach to ML-LCA, the Level 2 unit-specific coefficients are assumed to come from a certain (e.g., normal) distribution (see also Henry & Muthén, 2010). In the nonparametric approach to ML-LCA, a second latent class model is specified at Level 2. The result is a finite number of Level 2 latent classes, which represent different subtypes of Level 2 units that differ in the distribution of Level 1 classes. The nonparametric approach has the advantages of less strong distributional assumptions, lower computational burden, and a potentially easier interpretability (Vermunt, 2003). To date, ML-LCA has primarily been used in cross-sectional nested designs (Mäkikangas et al., 2018; Van Eck et al., 2017; Vuolo et al.,

2012). Recently, ML-LCA has been applied to AA data to model patterns of daily affect regulation (Grommisch et al., 2020; Lischetzke et al., 2022).

In the context of careless responding in AA, we propose that careless responding indices can be operationalized on the occasions level (by translating previously defined indices into AA data) and that ML-LCA can be used to identify occasion-level (Level 1) latent profiles of momentary careless responding and to differentiate between person-level (Level 2) latent classes of individuals who differ in their use of careless responding over time. The selected careless responding indices are entered into the model as the observed Level 1 variables. Note that we use the term latent profiles (and not latent classes) on Level 1 because many of the careless responding indices that have been proposed are continuous variables and because this allowed us to better distinguish between the typologies at the different levels.

For designs in which multiple experimental groups are studied (in our case, high vs. low sampling frequencies), we suggest that the ML-LCA model can be further extended into a multigroup ML-LCA model. When analyzing cross-sectional (single-level) multigroup data, the extension of LCA models into multigroup LCA models (Clogg & Goodman, 1985) in general allows researchers to investigate measurement invariance across groups (i.e., the structural similarity of latent classes/profiles across groups) and equivalence across groups in terms of class/profile sizes (Eid et al., 2003). In a similar vein, the extension of ML-LCA into multigroup ML-LCA allows researchers to investigate the structural similarity of latent classes at each level across groups and equivalence across groups in terms of class sizes at Level 2 (we present a more detailed description of the possible model comparisons to test for measurement equivalence across groups in ML-LCA in the Method section). Whereas multigroup LCA models have previously been applied in cross-cultural studies (Eid et al., 2003; Kankaraš et al., 2010; McMullen et al., 2018), to our knowledge, a multigroup ML-LCA model has not been introduced or applied in previous research.

Research Questions and Hypotheses

The goal of the present research was to identify latent profiles of momentary careless responding at the occasion level (Level 1) and to differentiate between latent classes of individuals at the person level (Level 2) who differed in their use of careless responding over time using (multigroup) ML-LCA and investigate the potential effects of (a) an AA design feature (sampling frequency), (b) situational factors (momentary time pressure and momentary motivation), and (c) respondent-level factors (personality) on careless responding. To address our hypotheses and research questions, we used data from an experimental AA study in which participants were randomly assigned to one of two conditions (low sampling frequency or high sampling frequency). Moreover, the study combined the AA phase with an online survey before the AA phase (pretest) and an online survey after the AA phase (posttest). Specifically, we investigated the following hypotheses and research questions (which were preregistered in April 2021):[2]

Hypotheses on the Latent Typology of Careless Responding and Careless Respondents in the AA Phase

1. We expected that we would be able to identify latent profiles of momentary careless responding in both experimental groups (high and low sampling frequencies) at Level 1 (occasion level) in the AA phase. We expected to find at least two latent profiles at Level 1 (careless vs. careful responding). Additional latent profiles might represent different types of momentary careless responding (e.g., long string careless responding vs. inconsistent careless responding).

[2] The preregistration in April 2021, which was a preregistration of secondary data analyses, is a concretization of the analytic methods that were used to test Hypothesis 2 from the preregistration in February 2019. Moreover, the preregistration of secondary data analyses specified additional analyses to be conducted using the same data to assess the criterion-related validity of the latent class measurement models on the occasion level and the person level.

2. We expected that we would be able to identify individual differences in the frequency of momentary careless responding across the AA phase at Level 2 (person level) in both experimental groups (high and low sampling frequencies). We expected to find at least two latent classes of individuals (careless vs. careful responders). Additional latent classes might represent individual differences in the variability versus stability of careless responding over time (e.g., infrequent vs. frequent careless responding) or in the predominant type of momentary careless responding used during the AA phase (e.g., a predominantly long string type vs. a predominantly inconsistent type).

3. We expected to find the same number of (a) Level 1 latent profiles and (b) Level 2 latent classes across the two experimental groups in the AA phase.

4. We expected the experimental groups to differ in the proportion of participants who were assigned to a latent class of careless responders (at Level 2). In the high (vs. low) sampling frequency group, we expected a higher proportion of participants to be assigned to a careless responding class.

Hypotheses on Careless Responding on the Pretest and the Posttest

Additionally, we aimed to explore associations between careless responding in the AA phase and careless responding on the pretest and the posttest. However, these hypotheses were not the focus of the present research. Hence, we decided to provide detailed information about these hypotheses (Hypotheses 5 to 9 in the preregistration) only in the Supplemental Online Material.

Hypotheses (and Exploratory Research Questions) Regarding Associations of Latent Profiles/Classes With Covariates

Covariates of Latent Profile Membership on the Occasion Level (Level 1) in the AA Phase. 10. We expected momentary motivation to answer the questions in the current measurement occasion to be lower for occasions that were assigned to a latent profile of

momentary careless responding than for occasions that were assigned to a latent profile of momentary careful responding.

11. We expected momentary time pressure to be higher for occasions that were assigned to a latent profile of momentary careless responding than for occasions that were assigned to a latent profile of momentary careful responding.

Covariates of Latent Class Membership on the Person Level (Level 2) in the AA Phase. 12. We expected latent classes of careless responders to have lower values for aggregated (i.e., mean) momentary motivation (to answer the questions in the current measurement occasion) compared with latent classes of careful responders.

13. We expected latent classes of careless responders to have lower values for aggregated (i.e., mean) momentary time pressure compared with latent classes of careful responders.

14. We expected latent classes of careless responders to score lower on conscientiousness than latent classes of careful responders.

15. We expected latent classes of careless responders to score lower on agreeableness than latent classes of careful responders.

16. We expected latent classes of careless responders to score lower on extraversion than latent classes of careful responders.

17. We expected latent classes of careless responders to score higher on neuroticism than latent classes of careful responders.

Exploratory Research Questions. 18. We aimed to explore whether latent classes of careless responders differed from latent classes of careful responders in intellect.

19. We aimed to explore whether participants' reasons for participating in the study (interest in the topic, financial compensation, interest in feedback, or to do the researcher a favor) predicted latent class membership.

Method

Study Design

The study consisted of an initial online survey (pretest), an AA phase across 7 days with either three or nine measurement occasions per day (depending on the experimental condition that participants had been randomly assigned to), and a retrospective online survey (posttest). For the AA phase, participants chose a specific time schedule that best fit their waking hours (9:00-21:00 or 10:30-22:30). In the low sampling frequency group, the three measurement occasions per day were distributed evenly across the day. In the high sampling frequency group, the first, fifth, and ninth measurement occasions were scheduled at the same time of day as the three measurement occasions from the low sampling frequency group, and the six additional measurement occasions were distributed between these questionnaires (see Table 1 for more detailed information).

After the 7-day AA phase, a second 7-day AA phase followed immediately (together with a second posttest after the AA phase). This time, the sampling frequency was switched between the groups. This was done to ensure that each participant invested a comparable amount of time participating in the study so that the financial compensation, which was the same for both groups, was fair. Given that our focus was on the between-group comparison (high vs. low sampling frequency) and not on the effect of switching sampling frequency within persons, the analyses in the present paper were based on the data from the first 7-day AA phase.

At each measurement occasion, participants rated their momentary motivation, momentary time pressure, mood, clarity of mood, state extraversion, and state conscientiousness. At the last measurement occasion each day, participants additionally rated their daily stress and perceived burden. On the pretest, participants rated their dispositional mood, mood regulation, personality (Big Five), emotional clarity, attention to feelings,

reasons for participating in the study (interest in the topic vs. financial compensation vs. interest in feedback vs. to do the researcher a favor), and sociodemographic variables. On each posttest, participants rated their dispositional mood, emotional clarity, attention to feelings, as well as perceived burden and careless responding ("SRSI Use Me" and "SRSI Effort"; Meade & Craig, 2012) with respect to the past 7 days.

Participants

Participants were required to be currently enrolled as a student and to be in possession of an Android smartphone. Participants were recruited via flyers, posters, e-mails, and posts on Facebook during students' semester breaks.

A total of 474 individuals filled out the initial online survey. Due to technical problems with the software for the AA phase, various participants could not synchronize their smartphone with our study and withdrew their participation. One of the reasons for dropout was that participants with an iOS smartphone realized only at this stage that participation required an Android smartphone as had been indicated in the information we gave them about the study. A total of 318 individuals took part in the AA phase that followed (155 individuals in the low sampling frequency condition), and 200 individuals responded in time to the retrospective online survey after the first 7 days (within the prespecified time frame of 12 hr). Participants who did not respond to the retrospective online survey were not excluded from the data analyses. Participants who completed only one measurement occasion during the AA phase were excluded from all analyses because the distribution of their Level 1 profile membership did not differ over time (in the multigroup ML-LCA). Therefore, eight participants (and measurement occasions) were excluded. Note that this exclusion criterion did not change the results of the ML-LCAs or the degree of measurement invariance across the two sampling frequency groups. The final sample consisted of 310 students (150 individuals in the low sampling frequency group: 87% women; age Range: 18 to 34 years, *M*

= 23.19, *SD* = 3.25; high sampling frequency group: 82% women; age Range: 18 to 50 years, M = 24.12, SD = 4.61).

Procedure

All study procedures were approved by the psychological ethics committee at the University Koblenz-Landau, Germany. After informed consent was obtained, the study began with the initial online survey. Subsequently, participants were randomly assigned to one of two experimental conditions (low sampling frequency or high sampling frequency) and randomly assigned to a starting day of the week. Prior to their AA phase, participants received a manual that explained how to install and run the smartphone application movisensXS, Versions 1.4.5, 1.4.6, 1.4.8, 1.5.0, and 1.5.1 (movisens GmbH, Karlsruhe, Germany) and connect to our study as this was required for participation. Participants were told that the number of questionnaires administered per day would range from three to nine times over the 2 weeks. At each measurement occasion, participants could either respond to the questionnaire or delay their response for up to 15 min. Participants who missed the first alarm were signaled 5 min later. If they did not start answering the questionnaire by 15 min after the first signal, the questionnaire became unavailable. At the end of the 7th day of each AA phase (21:00 for participants with the early time schedule or 22:30 for participants with the later time schedule), participants were sent a link to the retrospective online survey via e-mail. This online survey had to be completed within a 12-hr time frame. Students were given 15€ in exchange for their participation if they answered at least 50% of the AA questionnaires, and they had the chance to win an extra 25€ if they answered at least 80% of the AA questionnaires. Furthermore, at the end of the second retrospective online survey, they could indicate whether they wished to receive personal feedback regarding the constructs measured in the study after their participation was complete. In the low sampling frequency group (high

sampling frequency group), 98 (92) participants requested feedback, 10 (10) participants did not want feedback, and 45 (58) participants did not answer the item.

Measures

Careless Responding Indices in the AA Phase

We identified careless responding indices that have previously been used in cross-sectional survey research to study between-person differences in careless responding (Curran, 2016; Maniaci & Rogge, 2014; Meade & Craig, 2012; Niessen et al., 2016) and evaluated whether the proposed indices could be applied as indicators of momentary careless responding in our AA study (i.e., on the occasion level).

Average Long String. At each measurement occasion, we computed a long string for each page of the questionnaire that displayed a set of items with the same response format (i.e., pages displaying setting items with varying response formats were not used to compute a long string). Subsequently, for each measurement occasion, we calculated the average long string across all page-specific long strings. Note that the maximum long string was strongly correlated with the average long string at the within- and between-person levels, *r*s > .85. Hence, we decided to include only the average long string as an index of careless responding.

Occasion-Person Correlation. Similar to the idea of the person-total correlation index in cross-sectional surveys, the occasion-person correlation indicates the degree to which a measurement occasion is a typical measurement occasion within the participant. In particular, a lower occasion-person correlation score indicates an atypical measurement occasion within the participant, indicating a high probability of careless responding. For each participant, we calculated the occasion-person correlation across all substantive measures (i.e., excluding setting questions) at each measurement occasion. We used the Fisher-*z*-transformed version of the occasion-person correlation so that the index followed a normal distribution.

Response Time Index. The software used to administer the questionnaires at each measurement occasion recorded the response time (RT) for each page of the questionnaire. For each measurement occasion, the RTs for each page were aggregated into a sum so that each value represented the total time it took to complete the questionnaire. Note that some constructs (daily stress and daily perceived burden) were assessed only once per day (in the final AA measurement occasion each day). The RTs for these constructs were not included in the calculation of the RT index to ensure that the RT index was based on the same number of pages across all occasions. To accommodate the fact that response times systematically decreased across the duration of the study (because the items became more familiar to the participants over time), we detrended the RT sums. This was done by estimating (for the total sample of individuals across both experimental groups and for all measurement occasions that were scheduled at the same time of day across experimental groups) a multilevel growth curve model for occasions nested in persons, with RT as the dependent variable and the linear and quadratic terms for the number of days since beginning the study as the Level 1 predictors. The Level 1 residuals—which represent the occasion-specific deviations from the expected RT scores based on the linear and quadratic time trend—from this model served as the RT index for our careless responding analyses.

Inconsistency Index. To define an inconsistency index (Meade & Craig, 2012) for each measurement occasion in an AA study, items that are very similar in content and demonstrate a very large (negative or positive) within-person correlation are needed. In our study, momentary pleasant-unpleasant mood items (good-bad vs. happy-unhappy vs. unpleased-pleased vs. unwell-well; within-person intercorrelations ranged from $r = |.63|$ to $|.73|$), momentary calm-tense mood items (tense-relaxed vs. calm-rested; the within-person correlation was $r = -.55$), and momentary wakefulness-tiredness items (tired-awake vs. rested-sleepy; the within-person correlation was $r = -.71$) met these criteria. The response format was

a bipolar 7-point Likert scale with the endpoints verbally labeled (e.g., 1 = *very good* to 7 = *very bad*). Inconsistent responding at a particular measurement occasion is a response pattern that is internally/logically inconsistent. More specifically, we defined inconsistent responding as illogical responses across a mood item pair with responses near (or at) the extremes of the scale (Categories 1 or 2 vs. 6 or 7). For example, response patterns such as feeling "*very* happy" and "*very* unwell" at the same time or feeling "*very* happy" and "*very* bad" at the same time would be categorized as inconsistent responses. We calculated a sum score for all inconsistent responses across item pairs (possible range of values: 0 to 4).

Covariates (Level 1)

Momentary Motivation. Participants rated their momentary motivation to complete the questionnaire with a single item ("I am currently motivated to answer the questions"). The response format was a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*very much*).

Momentary Time Pressure. We measured momentary time pressure with a single item ("I am currently pressed for time"; Ottenstein & Lischetzke, 2019). The response format was a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*very much*).

Covariates (Level 2)

Big Five Personality Traits. We measured extraversion, agreeableness, conscientiousness, neuroticism, and intellect (openness) with unipolar adjective scales (Trierweiler et al., 2002) that had four adjectives per dimension. The response format was a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*very much so*). We calculated a mean score across the four items in each dimension such that a higher value indicated a higher standing on the respective personality trait. Revelle's omega total (McNeish, 2018) ranged from .64 (extraversion) to .83 (intellect).

Reasons for Participating in the Study. Participants were asked to indicate their reasons for participating in the study on a single item with five response options (1 = interest

in the topic of the study, 2 = financial compensation, 3 = interest in the personal feedback, 4 = to do the researcher a favor, and 5 = other). They could select multiple options if they wished to. Each reason for participating in the study was dummy-coded with a value of 1 for selecting the response option.

Data Analytic Models

We conducted ML-LCAs (with occasions nested in persons) on the data from the AA phase of the study to identify latent profiles of momentary careless responding at the occasion level (Level 1) and to differentiate between latent classes of individuals at the person level (Level 2) who differed in their use of careless responding over time. The four indices of careless responding were entered as observed variables (treating the average long string index, the occasion-person correlation, and the response time index as continuous variables, and the inconsistency index as an ordered categorical variable). The ML-LCAs were estimated based on the three measurement occasions per day that were scheduled at the same time of day across experimental groups (i.e., in the low sampling frequency group, the three measurement occasions per day were used; and in the high sampling frequency group, the corresponding first, fifth, and ninth measurement occasions of the day were used).[3]

To identify the correct number of latent profiles/latent classes at Levels 1 and 2, we followed the three-step model-fitting procedure recommended by Lukočiené et al. (2010). For

[3] There are two reasons why we chose to use only the corresponding measurement occasions in the high sampling frequency group (and not all the measurement occasions): First, fluctuations in mood, one of our substantive measures used to compute the careless responding indices, follow a diurnal rhythm. Therefore, the additional measurement occasions (occasions two to four and occasions six to eight, see Table 1) would be more similar to each other compared to the three more distant measurement occasions (one, five, and nine) that matched with the measurement occasions in the low sampling frequency group, which would lead to a larger occasion-person correlation among the additional measurement occasions. Second, we measured additional measures in the last measurement occasion of each day. That is, the average long string index could be higher for this measurement occasion compared to all other measurement occasions. A difference in the total number of analyzed measurement occasions per day between the two experimental groups would therefore introduce a systematic difference in the distribution of the long string index between groups. Taken together, using only the matching measurement occasions allowed us to rule out alternative explanations for potential between-group differences in our careless responding indices.

each experimental group, we first conducted a series of LCAs (ignoring the multilevel structure) to determine the optimal number of Level 1 profiles (at the occasion level). Second, we determined the optimal number of Level 2 classes (at the person level) in a series of ML-LCAs in which the number of Level 1 profiles was fixed to the value of the first step. Third, we redetermined the number of Level 1 profiles using ML-LCAs in which the number of Level 2 classes was fixed to the value of the second step. This three-step model-fitting procedure takes into account the possibility that the number of Level 1 profiles might change after the multilevel structure is accounted for.

As a model selection criterion, we used the BIC that was based on the Level 2 sample size because it has been shown to perform best for models with continuous observed variables (Lukočienė et al., 2010). Given that information criteria values may continue decreasing as the number of latent classes increases even though these classes are not meaningfully different and might not represent distinct groups (e.g., Masyn, 2013), we additionally examined the size of the decrease in the BIC and explored whether the decrease flattened out at some point (Nylund et al., 2007). As additional criteria, we used the quality of the latent class separation (entropy R^2, classification error, and the average posterior class probability) and the substantive interpretability of the latent class solution (Masyn, 2013). With respect to interpretability, we selected the more parsimonious solution if an additional class in a *k* class model represented only a slight variation of a class found in a *k*–1 class model.

To test whether the two experimental groups differed in the proportion of participants who were assigned to a latent class composed of careless responders (Hypothesis 4), we extended the ML-LCA model into a multigroup ML-LCA model and adapted the general model-fitting procedure that has been recommended for "standard" (single-level) multigroup LCA models (Eid et al., 2003; Kankaraš et al., 2010) to our nested data structure. That is, after determining the optimal numbers of latent profiles of momentary careless responding at Level

1 and the latent classes of individuals at Level 2 in each experimental group separately, we analyzed measurement invariance across the two experimental groups. We estimated the following four multigroup ML-LCA models, which differed in the assumed degree of measurement invariance across groups: In the *heterogeneous multigroup ML-LCA model*, the definition of latent profiles at Level 1 (i.e., conditional means for the continuous indicators and conditional response probabilities for the ordinal indicator), the definition of latent classes at Level 2 (i.e., the conditional distribution of Level 1 profiles over time in each latent class at Level 2), and the class sizes at Level 2 were allowed to differ across experimental groups. In the *partially homogeneous multigroup ML-LCA Model A*, the definition of latent profiles at Level 1 was constrained to be equal across groups, whereas the definition of latent classes at Level 2 and the class sizes at Level 2 were still allowed to differ across experimental groups. In the *partially homogeneous multigroup ML-LCA Model B*, the definition of both latent profiles at Level 1 and latent classes at Level 2 were constrained to be equal across experimental groups, but the class sizes at Level 2 were allowed to differ across groups. In the *fully homogeneous multigroup ML-LCA model*, all parameters were constrained to be equal across experimental groups. As the model selection criterion, we again used the BIC that was based on the Level 2 sample size (Lukočienė et al., 2010).

To test our hypotheses on associations of the latent class variables at Levels 1 and 2 (in the AA phase) with time-varying and time-invariant covariates (Hypotheses 10 to 19), we used the adjusted (maximum likelihood) three-step approach recommended by Bakk et al. (2013). Specifically, we examined (a) whether the latent profiles of momentary careless responding (at Level 1) had different mean levels of the time-varying covariates (momentary motivation, momentary time pressure), (b) whether the latent classes of individuals (at Level 2) had different mean levels of the continuous time-invariant covariates (Big Five personality traits, aggregated momentary motivation, and aggregated time pressure), and (c) whether

individuals' reasons for participating (dummy-coded) predicted their latent class membership (at Level 2). The adjusted three-step approach accounts for classification errors in class assignment when analyzing relationships between latent class membership and covariates (Bakk et al., 2013). We evaluated the statistical significance of our results using Wald tests. We corrected for multiple testing using Benjamini and Hochberg's (1995) procedure for controlling the false discovery rate (FDR) in these analyses.

All analyses were computed in Latent Gold 6.0 (Vermunt & Magidson, 2021). To avoid local maxima and nonconvergence, we used 1,000 sets of random starting values. Data management (including calculation of careless responding indices) was done in the R environment (R Core Team, 2022) and with the R package careless (Yentes & Wilhelm, 2018).

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The study procedure and analyses were preregistered on the OSF. Moreover, all data, analysis code, and research materials are available on the OSF repository (https://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2).

Results

Descriptive Statistics

Descriptive statistics for the careless responding indices in the AA phase for each group can be found in Table 2.

Multilevel Latent Class Analysis of Careless Responding Indices

First, we determined the optimal ML-LCA solution for each sampling frequency group separately. Note that in the following, the BIC always refers to the BIC that was based on the Level 2 sample size.

Multilevel Latent Class Analysis for the Low Sampling Frequency Group

To identify the optimal ML-LCA solution for the low sampling frequency group, we determined the number of careless responding profiles on the occasion level (Level 1) in the first step. The upper panel of Table 3 shows the model fit statistics for a series of single-level LCA models with up to eight profiles on Level 1. The BIC continued to decrease as the number of latent profiles increased. As the additional criteria (i.e., the Entropy R^2, the classification error, and the average posterior class probability) continued to get worse as the number of latent profiles increased (with the exception of the Entropy R^2 between the three- and four-profile solutions), we determined the number of latent profiles on the basis of substantive interpretability. The two-, three-, and four-profile solutions each added a qualitatively distinct profile to the previous (less complex) model. Models with more than four profiles did not add a qualitatively distinct profile but rather showed just minor variations from the profiles found in the four-profile solution. Therefore, we selected the four-profile solution at Level 1 in the first step.

In the second step, we determined the optimal number of Level 2 classes (on the person level). The upper panel of Table 4 shows the model fit statistics for a series of ML-LCA models with up to six Level 2 classes (and the number of Level 1 profiles fixed to four). The BIC suggested that a four-class model fit the data best, whereas the additional model fit criteria on Level 2 (i.e., the Entropy R^2, the classification error, and the average posterior class probability) supported the three-class model. With respect to substantive interpretability, the

four-class model added a qualitatively distinct class compared with the three-class model. Hence, we selected four Level 2 classes in the second step.

In the third step, we redetermined the number of Level 1 profiles by comparing models with four Level 2 classes and one to eight Level 1 profiles. The upper panel of Table 5 shows the model fit statistics for this series of ML-LCA models. Again, the BIC continued to decrease as the number of latent profiles increased, and the additional criteria did not show clear support for a specific profile solution (i.e., the classification error and the average posterior class probability continued to get worse as the number of latent profiles increased, and the Level 1 Entropy R^2 pointed to the two-profile model). Hence, we determined the number of profiles on the basis of their substantive interpretability. Again, the four-profile solution was the most parsimonious model to add a qualitatively distinct profile, whereas models with more Level 1 profiles did not add qualitatively distinct profiles. Therefore, we selected the model with four Level 2 classes and four Level 1 profiles as the final model for the low sampling frequency group.

Interpretation of the Latent Class Structure of Careless Responding in the AA Phase for the Low Sampling Frequency Group

Figure 1 shows the configurations of momentary careless responding profiles (Level 1 profiles) for the low sampling frequency group as standard deviations from the overall sample mean. In the case of the inconsistency index, which was the only categorical indicator, the bars represent percent deviations of the probability of having at least one inconsistent response from the mean probability of having at least one inconsistent response (this was done so this indicator could be interpreted intuitively). For the four profiles, we labeled the largest profile the "careful" responding profile because all the careless responding indices were close to the average values of the sample. The second largest profile (presented as the third profile to be consistent across all the ML-LCA models) was characterized by high values

(compared with the sample mean) on the response time and inconsistency indices and average values on the average long string and occasion-person correlation indices. Hence, we labeled this profile the "inconsistent" profile. The third largest profile (presented as the second profile to be consistent across all the ML-LCA models) was characterized by high values (compared with the sample mean) on the average long string and occasion-person correlation indices and low values on the inconsistency index. Therefore, we labeled this profile the "long string" responding profile. We labeled the smallest profile the "slow" profile because it was characterized by extremely high values (compared with the sample mean) on the response time index (and average values on the other indices). In line with Hypothesis 1 (for the low sampling frequency group), we identified multiple careless responding profiles that differentiated between the careful and careless occasions and represented different types of momentary careless responding.

Figure 2 shows the distribution of classes of individuals who differed in Level 1 profile membership over time for the four classes for the low sampling frequency group. We labeled the class of individuals that almost always (> 90%) used the careful profile and rarely used any other profile (< 5%) the "careful class." Two classes of individuals predominantly used the careful profile most of the time (62% and 73%) and either the long string or the inconsistent profiles relatively often. Accordingly, we labeled these classes of individuals the "infrequently careless class of the long string type" and the "infrequently careless class of the inconsistent type." We labeled the class of individuals that predominantly used the inconsistent profile the "frequently careless class." In line with Hypothesis 2 (for the low sampling frequency group), we identified different classes of individuals that represented individual differences in the variability versus stability of careless responding over time and in the predominant type of momentary careless responding used during the AA phase.

Multilevel Latent Class Analyses for the High Sampling Frequency Group

To identify the optimal ML-LCA solution for the high sampling frequency group, we followed the same steps as described for the low sampling frequency group. The lower panel of Table 3 shows the model fit statistics for the series of (single-level) LCA models with up to eight profiles at Level 1. The BIC continued to decrease as the number of latent profiles increased. However, the drop in the BIC flattened out after the five-profile solution, suggesting that more than five profiles were not needed to describe the data adequately. The additional criteria (entropy R^2, classification error, and average posterior class probability) supported the model with two profiles. In terms of substantive interpretability, the four-profile solution was the most parsimonious model to add a qualitatively distinct profile. The five-profile solution included only minor variations in the profiles found in the four-profile solution. Therefore, we selected the four-profile solution at Level 1 in the first step.

In the second step, we determined the optimal number of Level 2 classes (at the person level), whereas we fixed the number of Level 1 profiles to four profiles. The lower panel of Table 4 shows the model fit statistics for up to six Level 2 classes. The BIC suggested that a three-class model fit the data best. Moreover, the additional criteria on Level 2 supported the three-class solution. With respect to substantive interpretability, the four-class model added a very small, qualitatively different class compared with the three-class model. This fourth class corresponded to the "frequently careless class" in the low sampling frequency group. The difference between the two classes across the two sampling frequency groups is the extent to which the careless responding profiles were used (95% of the time in the low sampling frequency group and 54% of the time in the high sampling frequency group). However, these classes had a very similar substantive interpretation (frequently careless). This suggests that the four-class model in the high sampling frequency group might be necessary to capture the differences between the two sampling frequency groups appropriately (to be directly

compared in a multigroup ML-LCA model after the separate group-specific models). Hence, we selected the model with four Level 2 classes in the second step.

In the third step, we redetermined the number of Level 1 profiles, whereas we fixed the number of Level 2 classes to four. The lower panel of Table 5 shows the model fit statistics for up to eight profiles at Level 1. Again, the BIC continued to decrease as the number of latent profiles increased. As the additional criteria did not show clear support for a specific profile solution (i.e., the Level 1 entropy R^2 value and the average posterior class probability supported the two-profile model, whereas the classification error continued to increase as the number of latent profiles increased), we determined the number of profiles on the basis of their substantive interpretability. Again, the four-profile solution was the most parsimonious model to add a qualitatively distinct profile. Therefore, we selected the model with four Level 2 classes and four Level 1 profiles as the final model.

Interpretation of the Latent Class Structure of Careless Responding in the AA Phase for the High Sampling Frequency Group

Figure 3 shows the configurations of momentary careless responding profiles (Level 1 profiles) for the high sampling frequency group as standard deviations from the overall sample mean ordered by size. Again, for the inconsistency index, the bars represent percent deviations of the probability of having at least one inconsistent response from the mean probability of having at least one inconsistent response. For the four profiles, we labeled the largest profile the "careful" responding profile because all careless responding indices were near the average values of the sample (in this profile). The second largest profile was characterized by high values (compared with the sample mean) on the average long string and the occasion-person correlation indices and low values on the inconsistency index. Therefore, we labeled this profile the "long string" responding profile. The third largest profile was characterized by high values (compared with the sample mean) on the response time and

inconsistency indices and average values on the average long string and the occasion-person correlation indices. Hence, we labeled this profile the "inconsistent" profile. We labeled the smallest profile the "slow" profile because it was characterized by extremely high values (compared with the sample mean) on the response time index (and average values on the other indices). In line with Hypothesis 1 (for the high sampling frequency group), we identified multiple careless responding profiles that differentiated between careful and careless occasions and represented different types of momentary careless responding.

Figure 4 shows the distribution of classes of individuals who differed in their Level 1 profile membership over time for the four classes in the high sampling frequency group. We labeled the class of individuals that almost always (> 90%) used the careful profile and rarely used any other profile (< 5%) the "careful class." Two classes of individuals predominantly used the careful profile most of the time (75% and 80%) and either the long string or the inconsistent profile relatively often. Accordingly, we labeled these classes of individuals the "infrequently careless class of the long string type" and the "infrequently careless class of the inconsistent type." We labeled the class of individuals that used careless responding profiles 54% of the time profile the "frequently careless class." In line with Hypothesis 2 (for the high sampling frequency group), we identified different classes of individuals that represented individual differences in the variability versus stability of careless responding over time and in the predominant type of momentary careless responding used during the AA phase. In line with Hypothesis 3, we found very similar typologies of momentary careful versus careless responding (on Level 1) and careful versus careless respondents (on level 2) were identified across the two experimental groups.

Measurement Invariance Across Experimental Groups

To analyze measurement invariance across the two experimental groups, we estimated four multigroup ML-LCA models with different degrees of equality restrictions across groups. The fully homogeneous multigroup ML-LCA model, in which all parameters were set equal across experimental groups, showed the lowest BIC value (BIC = 13633.735), followed by the partially homogeneous Model B (BIC = 13633.750), the partially homogeneous Model A (BIC = 13682.746), and the heterogeneous model (BIC = 13709.274). This means that the definition of Level 1 profiles, the definition of Level 2 classes, and the class sizes at Level 2 could be assumed to be equal across the experimental groups. Therefore, we rejected Hypothesis 4, which predicted that the experimental groups would differ in the proportion of participants who were assigned to a latent class of careless responders (at Level 2). Figure 5 shows the configurations of momentary careless responding profiles (Level 1 profiles) as standard deviations from the overall sample mean ordered by size, and Figure 6 shows the distribution of classes of individuals who differed in Level 1 profile membership over time for the four classes resulting from the fully homogeneous multigroup ML-LCA model. We do not describe the interpretation of the latent class structure of careless responding in the AA phase for the final model because it is very similar to the interpretation for the low sampling frequency group.

Covariates of Latent Class Membership at Levels 1 and 2 in the AA Phase

Differences Between Occasion Level Latent Profile Membership (Level 1) in Time-Varying Covariates

Results of the Wald tests of the mean differences in (time-varying) covariates across the careless responding profiles are presented in the left panel of Table 6. Contrary to Hypotheses 10 and 11, there was no evidence of mean differences in time-varying momentary motivation and momentary time pressure across the momentary careless responding profiles.

Differences Between Individual Level Latent Class Membership (Level 1) in Time-Invariant Covariates

Results of the Wald tests of the mean differences in (time-invariant) covariates across classes of individuals are presented in the right panel of Table 6, and the findings that remained significant after the false discovery rate were corrected are indicated in bold. Latent classes differed in values of aggregated momentary motivation, $\chi^2(3) = 36.39$, $p < .001$, $R^2 = .072$. In line with Hypothesis 12, the frequently careless class ($M = 2.39$) had lower values of aggregated momentary motivation compared with the careful class ($M = 3.13$), $\chi^2(1) = 19.66$, $p < .001$, the infrequently careless class of the long string type ($M = 3.49$), $\chi^2(1) = 33.37$, $p < .001$, and the infrequently careless class of the inconsistent type ($M = 3.16$), $\chi^2(1) = 18.18$, $p < .001$. Contrary to Hypothesis 12, differences between the other classes remained nonsignificant after the false discovery rate was corrected. Contrary to Hypotheses 13 to 17, and Research Question 18, the classes of individuals (Level 2) did not show mean differences in aggregated momentary time pressure or in Big Five personality traits.

Exploratory Covariate Analyses of Participants' Reasons for Participating in the Study

Exploratory analysis of Research Question 19 revealed that interest in the topic, financial compensation, and interest in feedback were not associated with individuals' latent class membership. However, the latent classes differed in the extent to which they indicated that their reason for participating was *to do the researcher a favor*, $\chi^2(3) = 20.87$, $p < .001$, $R^2 = .027$. The frequently careless class less often indicated (1%) that their reason for participating was *to do the researcher a favor* compared with the careful class (36%), $\chi^2(1) = 18.66$, $p < .001$, the infrequently careless class of the long string type (35%), $\chi^2(1) = 17.77$, $p < .001$, and the infrequently careless class of the inconsistent type (20%), $\chi^2(1) = 9.74$, $p = .002$. Differences between the other classes were nonsignificant.

Discussion

To our knowledge, the current study is the first to model careless responding in AA using an LCA approach. This was done by identifying indirect careless responding indices that could be applied to AA data and by entering these indices of momentary careless responding as observed variables into an ML-LCA model—that is, a multilevel extension of LCA that accommodates the nested data structure of AA data. In the current study, we also investigated how (a) AA design features (sampling frequency), (b) situational factors (momentary motivation and momentary time pressure), and (c) respondent-level factors (personality and reasons for participating in the study) were associated with careless responding in AA. To test whether experimentally manipulated sampling frequency had an effect on careless responding, we extended the ML-LCA model into a multigroup ML-LCA model and showed how different degrees of measurement invariance across experimental groups could be analyzed in the multigroup ML-LCA framework. In our AA data, we identified four momentary careless responding profiles on Level 1 (the careful, inconsistent, long string, and slow responding profiles) and four classes of individuals who differed in their use of the momentary careless responding profiles over time on Level 2 (a careful class, an infrequently careless class of the long string type, an infrequently careless class of the inconsistent type, and a frequently careless class). Our results demonstrated that sampling frequency (3 vs. 9 occasions per day for 7 days) did not influence careless responding in AA. On the occasion level, momentary careless responding was unrelated to momentary time pressure and momentary motivation to complete the questionnaire. On the person level, careless responding in AA was related to motivational variables (aggregated motivation to answer questionnaires and participating in the study because the participant wanted *to do the researcher a favor*) but was unrelated to the Big Five personality traits.

Using indirect careless responding indices that fall into different, previously proposed types of indices (i.e., categorized by consistency, response pattern, and response time), the present research demonstrated that an LCA approach that is adapted to the nested data structure (ML-LCA) can be used to model careless responding in AA. Our results on the latent typology of momentary careless responding were largely in line with previous research (Goldammer et al., 2020; Kam & Meyer, 2015; Li et al., 2020; Maniaci & Rogge, 2014; Meade & Craig, 2012), thereby demonstrating that different types of careless responding (long string vs. inconsistent careless responding) can also be identified in AA data. Similar to recommendations for studying careless responding in cross-sectional surveys (e.g., Curran, 2016), we recommend that researchers who want to investigate patterns of careless responding in their AA data use indices from these different types to capture different types of careless responding. The only difference between our findings and the findings from cross-sectional survey studies on careless responding was that we additionally identified a slow response profile, which had not been reported in previous cross-sectional research. An explanation for finding an additional slow profile could be that participants in AA are more likely to be disturbed while completing repeated daily measurement occasions (by their daily life activities) compared with participants who devote a specific time period to completing a one-time (cross-sectional) survey (e.g., in a laboratory setting or on a home computer). To our knowledge, this study is the first to investigate classes of individuals who differ in their use of the momentary careless responding over time in AA. In our sample, more than half of the participants (54%) were assigned to either the infrequently careless class of the long string type or the infrequently careless class of the inconsistent type. This result suggests that the majority of participants did not provide high quality data at all measurement occasions across the duration of the AA study. This finding is in line with the study by Jaso et al. (2021), in which 59% of participants exhibited carelessness at a minimum of one measurement occasion.

Additionally, they found that 1% of their participants exhibited carelessness at 75% of the measurement occasions, which corresponds to our frequently careless class, which comprised 2% of our sample.

By selecting ML-LCA as the data-analytic approach, the latent classes of individuals that were estimated on the person level represented individual differences in the distribution of momentary profiles of careless responding across the study period. If researchers are interested in additionally modeling the time structure of careless responding across the study period, mixture latent Markov models (van de Pol & Langeheine, 1990; Vermunt, 2010; Vermunt et al., 2008) could be applied as an alternative latent class analytic approach. In latent Markov models, initial probabilities (i.e., the probability to show a particular pattern of careless or careful responding at the first measurement occasion) and transition probabilities (i.e., the probability of remaining in the same careless responding profile over time or moving to another profile) are estimated. As an extension, a mixture latent Markov model could be used to identify latent classes of individuals who differ in the transition pattern of careless responding across the study period. This type of modeling approach could be fruitful if individual differences in the "timing" of careless responding in an AA study (and potential predictors there-of) are of particular interest (e.g., to potentially identify types of individuals who show careless responding mainly at the beginning of the study vs. mainly at the end of the study or types of individuals who mainly show careless responding of one particular type in the first half of the study and later switch to a different type of careless responding).

Situational factors that influence careless responding might give insight into possible reasons for why participants sometimes provide low quality data. Whereas our results suggest that momentary motivation to answer the questions in the current measurement occasion and momentary time pressure did not influence careless responding, other situational factors might lead to situational careless responding. One possible explanation for not finding an

effect of momentary time pressure on momentary careless responding is that participants who experienced high time pressure might have selected to skip the questionnaire at the measurement occasion entirely, in which case an effect of time pressure cannot be detected. One possible explanation for not finding an effect of momentary motivation on momentary careless responding is that the motivation to comply with the study procedure might be more important on the person level than on the occasion level. Once participants have strongly committed themselves to participate in an intensive data collection phase, within-person fluctuations in momentary motivation might be "overruled". Future research should aim to investigate other situational factors in AA to provide greater insight into the processes that lead to momentary careless responding.

Using the multigroup ML-LCA model, we demonstrated that this model can be used to investigate the influence of design features (in our case, sampling frequency) on careless responding in AA. We expected that a higher sampling frequency would lead to more careless responding because we expected that greater effort by the participants to comply with the study procedure (which was operationalized as 3 vs. 9 measurement occasions per day) would lead directly to lower data quality or would increase participants' perceived burden and therefore lead to lower quality data. Contrary to our expectations, we did not find an effect of sampling frequency on careless responding. This finding is in line with the study by Eisele et al. (2020), who found that the sampling frequency (3 vs. 6 vs. 9 occasions per day for 14 days) did not influence careless responding. (Note that the study by Eisele et al., 2020, had not yet been published at the time when we preregistered our study). Furthermore, Hasselhorn et al.'s (2021) study, which used the same data as we did in the current paper, suggested that sampling frequency did not impair other aspects of data quality (within-person variance and relationships between time-varying constructs) but increased perceived burden through study participation. It is interesting to note that the results of the studies by Eisele et al. (2020) and

Hasselhorn et al. (2021) suggested that a different design feature (questionnaire length) might influence careless responding and other aspects of data quality. Future research should investigate in more detail the effects of questionnaire length and other design features on careless responding in AA.

Our findings that the Big Five personality dimensions were unrelated to careless responding are not in line with previous cross-sectional research (Bowling et al., 2016; Grau et al., 2019), which found that conscientiousness, agreeableness, extraversion, and emotional stability were negatively related to careless responding. One notable difference between previous research and our study is that previous research on the link between personality and careless responding investigated indirect careless responding indices in one survey and analyzed how these indices were correlated with the Big Five dimensions, whereas we analyzed whether latent class membership (representing individual differences in the distribution of momentary careless responding patterns across the AA phase) was related to personality scores. Additionally, we did not use Mahalanobis distance as an indirect careless responding indicator (as discussed in Footnote 1). Furthermore, personality traits might have a different influence on careless responding in AA (compared with cross-sectional research) because of the differences in the study procedures between these designs. Compared with participation in a cross-sectional survey, the commitment to participate in an intensive longitudinal study and the conditions under which this intensive assessment takes place (e.g., receiving repeated beeps or app notifications in daily life) might represent a stronger situation and hence lead to a smaller influence of personality traits on careless responding in AA compared with cross-sectional survey research. In line with this reasoning, we found that higher motivation (aggregated momentary motivation and participating in the study because a participant wanted *to do the researcher a favor*) was associated with less careless responding in AA. Future research might investigate whether the amount of (financial) compensation that

participants receive for their participation in an AA study has an effect on the degree of careless responding.

Limitations

Several limitations need to be considered when interpreting the results of the current investigation. First, our student sample (with a large proportion of women) was not representative of the general population. Therefore, our findings might not be generalizable beyond a young adult population.

Second, as substantive measures, we assessed momentary mood and personality states in the AA phase of our study by using scales that included reverse-scored items. This allowed us to compute different types of careless responding indices, including an inconsistency index. Other AA studies might use different types of measures (e.g., with all items in the same direction), and hence, it might not be possible to compute the same momentary careless responding indices. Moreover, different types of time-varying constructs (e.g., affective vs. behavioral vs. cognitive constructs) and different response formats (e.g., response categories vs. slider scales) might be associated with a different degree of burden for participants and might thereby yield different types and different degrees of careless responding in an AA study. Future research could manipulate other aspects of AA studies (e.g., constructs or response formats) and investigate their effects on data quality.

Finally, in the current investigation, we were unable to control for the potential effects of response styles on our results. It is possible that controlling for the effects of response styles could change some of the momentary careless responding profiles or careless classes of individuals. One example might be the long string (invariability) profile (on Level 1), which, to some extent, might capture the midpoint response style (participants' preference for choosing the middle category). The results of a cross-sectional study by Grau et al. (2019) were in line with this assumption as they found that careless responding and response styles

were correlated. However, their results suggest that careless responding and response styles are distinct constructs. Additionally, response styles have not yet been modeled with AA data. Therefore, future research should investigate the association between careless responding and response styles in AA.

Conclusions

Although the generalizability of the current results should be scrutinized by future research, the present research contributes to AA research by demonstrating how careless responding can be modeled in AA data with a multilevel extension of the LCA approach (ML-LCA). We also showed how this model could be extended further into a multigroup ML-LCA model to investigate measurement invariance with respect to the latent typologies at the different levels across groups. We hope that future research on careless responding in AA might find the application of (multigroup) ML-LCA useful for shedding more light on the effects of design features of AA studies on careless responding, and ultimately, to be able to enhance data quality in AA.

References

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the Association between Latent Class Membership and External Variables Using Bias-adjusted Three-step Approaches. *Sociological Methodology*, *43*(1), 272–311. https://doi.org/10.1177/0081175012470644

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

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218–229. https://doi.org/10.1037/pspp0000085

Clogg, C. C., & Goodman, L. A. (1985). Simultaneous latent structure analysis in several groups. In *Sociological methodology 1985*. (pp. 81–110). Jossey-Bass.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The Differential Impacts of Two Forms of Insufficient Effort Responding: IMPACT OF DIFFERENT TYPES OF IER. *Applied Psychology*, *67*(2), 309–338. https://doi.org/10.1111/apps.12117

Edwards, J. R. (2019). Response invalidity in empirical research: Causes, detection, and remedies. *Journal of Operations Management*, *65*(1), 62–76. https://doi.org/10.1016/j.jom.2018.12.002

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing Typological Structures Across Cultures By Multigroup Latent Class Analysis: A Primer. *Journal of Cross-Cultural Psychology*, *34*(2), 195–210. https://doi.org/10.1177/0022022102250427

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*. 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*.

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), 101384. https://doi.org/10.1016/j.leaqua.2020.101384

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural Differences in Careless Responding. *Journal of Cross-Cultural Psychology*, *50*(3), 336–357. https://doi.org/10.1177/0022022119827379

Grommisch, G., Koval, P., Hinton, J. D. X., Gleeson, J., Hollenstein, T., Kuppens, P., & Lischetzke, T. (2020). Modeling individual differences in emotion regulation repertoire in daily life with multilevel latent profile analysis. *Emotion*, *20*(8), 1462–1474. https://doi.org/10.1037/emo0000669

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. (2021). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person

relationships in ambulatory assessment. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01683-6

Henry, K. L., & Muthén, B. (2010). Multilevel Latent Class Analysis: An Application of Adolescent Smoking Typologies With Individual and Contextual Predictors. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 193–215. https://doi.org/10.1080/10705511003659342

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

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. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2021). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychological Methods*. https://doi.org/10.1037/met0000312

Kam, C. C. S., & Meyer, J. P. (2015). How Careless Responding and Acquiescence Response Bias Can Influence Construct Dimensionality: The Case of Job Satisfaction. *Organizational Research Methods*, *18*(3), 512–541. https://doi.org/10.1177/1094428115571894

Kankaraš, M., Moors, G., & Vermunt, J. K. (2010). Testing for Measurement Invariance With Latent Class Analysis. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.),

Cross-Cultural Analysis (2nd ed., pp. 393–419). Routledge. https://doi.org/10.4324/9781315537078-14

Larson, R., & Csikszentmihalyi, M. (1983). The Experience Sampling Method. *New Directions for Methodology of Social & Behavioral Science*, *15*, 41–56.

Li, C. R., Follingstad, D. R., Campe, M. I., & Chahal, J. K. (2020). Identifying Invalid Responders in a Campus Climate Survey: Types, Impact on Data, and Best Indicators. *Journal of Interpersonal Violence*, 088626052091858. https://doi.org/10.1177/0886260520918588

Lischetzke, T., Schemer, L., In-Albon, T., Karbach, J., Könen, T., & Glombiewski, J. A. (2022). Coping under a COVID-19 lockdown: Patterns of daily coping and individual differences in coping repertoires. *Anxiety, Stress, & Coping*, *35*(1), 25–43. https://doi.org/10.1080/10615806.2021.1957848

Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). 6. The Simultaneous Decision(s) about the Number of Lower- and Higher-Level Classes in Multilevel Latent Class Analysis. *Sociological Methodology*, *40*(1), 247–283. https://doi.org/10.1111/j.1467-9531.2010.01231.x

Mäkikangas, A., Tolvanen, A., Aunola, K., Feldt, T., Mauno, S., & Kinnunen, U. (2018). Multilevel Latent Profile Analysis With Covariates: Identifying Job Characteristics Profiles in Hierarchical Data as an Example. *Organizational Research Methods*, *21*(4), 931–954. https://doi.org/10.1177/1094428118760690

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Masyn, K. E. (2013). *Latent Class Analysis and Finite Mixture Modeling*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0025

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. https://doi.org/10.1037/a0019216

McKibben, W. B., & Silvia, P. J. (2017). Evaluating the Distorting Effects of Inattentive Responding and Social Desirability on Self-Report Scales in Creativity and the Arts. *The Journal of Creative Behavior*, *51*(1), 57–69. https://doi.org/10.1002/jocb.86

McMullen, J., Van Hoof, J., Degrande, T., Verschaffel, L., & Van Dooren, W. (2018). Profiles of rational number knowledge in Finnish and Flemish students – A multigroup latent class analysis. *Learning and Individual Differences*, *66*, 70–77. https://doi.org/10.1016/j.lindif.2018.02.005

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

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

Mehl, M. R., & Conner, T. S. (Eds.). (2014). *Handbook of research methods for studying daily life* (Paperback ed). Guilford.

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

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Ottenstein, C., & Lischetzke, T. (2019). Development of a Novel Method of Emotion Differentiation That Uses Open-Ended Descriptions of Momentary Affective States. *Assessment*, 107319111983913. https://doi.org/10.1177/1073191119839138

Paas, L. J., Dolnicar, S., & Karlsson, L. (2018). Instructional Manipulation Checks: A longitudinal analysis with implications for MTurk. *International Journal of Research in Marketing*, *35*(2), 258–269. https://doi.org/10.1016/j.ijresmar.2018.01.003

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schneider, S., May, M., & Stone, A. A. (2018). Careless responding in internet-based quality of life assessments. *Quality of Life Research*, *27*(4), 1077–1088. https://doi.org/10.1007/s11136-017-1767-2

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 Berkel, N., Ferreira, D., & Kostakos, V. (2018). The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys*, *50*(6), 1–40. https://doi.org/10.1145/3123988

van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., & Kostakos, V. (2019). Context-Informed Scheduling and Analysis: Improving Accuracy of Mobile Self-Reports. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300281

van de Pol, F., & Langeheine, R. (1990). Mixed Markov Latent Class Models. *Sociological Methodology*, *20*, 213. https://doi.org/10.2307/271087

Van Eck, K., Johnson, S. R., Bettencourt, A., & Johnson, S. L. (2017). How school climate relates to chronic absence: A multi–level latent profile analysis. *Journal of School Psychology*, *61*, 89–102. https://doi.org/10.1016/j.jsp.2016.10.001

Vermunt, J. K. (2003). 7. Multilevel Latent Class Models. *Sociological Methodology*, *33*(1), 213–239. https://doi.org/10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*(1), 33–51. https://doi.org/10.1177/0962280207081238

Vermunt, J. K. (2010). Longitudinal Research Using Mixture Models. In K. van Montfort, J. H. L. Oud, & A. Satorra (Eds.), *Longitudinal Research with Latent Variables* (pp. 119–152). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11760-2_4

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, Measurement, and Analysis* (pp. 373–385). Elsevier.

Vuolo, M., Staff, J., & Mortimer, J. T. (2012). Weathering the great recession: Psychological and behavioral trajectories in the transition from school to work. *Developmental Psychology*, *48*(6), 1759–1773. https://doi.org/10.1037/a0026047

Yentes, R. D., & Wilhelm, F. (2018). *careless: Procedures for computing indices of careless responding*.

Table 1

Sampling Scheme of the AA Study

Time of day	Experimental group	
	Low sampling frequency	High sampling frequency
9:00-10:40	Occasion 1	Occasion 1
11:00-13:50		Occasions 2 to 4
14:10-15:50	Occasion 2	Occasion 5
16:10-19:00		Occasions 6 to 8
19:20-21:00	Occasion 3	Occasion 9

Note. The displayed sampling scheme refers to the first of the two time schedules from which participants could choose (9:00-21:00 vs. 10:30-22:30). That is, in the second time schedule, each occasion was scheduled 90 min later. At each measurement occasion, participants had the option to delay their response for up to 15 min. In the high sampling frequency group, Occasions 2, 3, and 4 and Occasions 6, 7, and 8 were at least 28 min apart.

Table 2

Descriptive Statistics and Bivariate Correlations for the Careless Responding Indices Presented Separately in Each Experimental Group

Group	Variable	1	2	3	4
Low sampling frequency	1. Average long string	—	.17***	-.06*	-.02
	2. Occasion-person correlation	.17*	—	-.07**	-.17***
	3. Response time index	-.21**	-.07	—	.02
	4. Inconsistency index	.08	-.15	.10	—
	M	1.38	1.11	0.07	0.08
	SD_{within}	0.30	0.42	0.61	0.35
	$SD_{between}$	0.12	0.25	0.35	0.23
High sampling frequency	1. Average long string	—	.11***	-.01	-.02
	2. Occasion-person correlation	.10	—	-.12***	-.12***
	3. Response time index	-.13	-.20*	—	.04
	4. Inconsistency index	-.18*	-.10	-.13	—
	M	1.36	1.15	-0.09	0.04
	SD_{within}	0.28	0.44	0.59	0.25
	$SD_{between}$	0.12	0.27	0.31	0.11

Note. Between-person correlations ($N_{Persons}$ = 150 in the low sampling frequency group and $N_{Persons}$ = 160 in the high sampling frequency group) are presented below the diagonal. Within-person correlations ($N_{occasions}$ = 2,296 in the low sampling frequency group and $N_{occasions}$ = 2,293 in the high sampling frequency group) between the careless responding indices are presented above the diagonal. All *p*-values are two-sided.

*p < .05. **p < .01. ***p < .001.

Table 3

Model Fit Statistics for LCA Models With Different Numbers of Level 1 Profiles (and One Level 2 Class) in Both Sampling Frequency Groups

Number of profiles	LL	BIC	Size of drop in BIC	Npar	Entropy R^2 (L1)	Class. Err. (L1)	Smallest AvePP (L1)	Size of smallest P
Low sampling frequency group								
1	-4577.258	9204.621		10	1	0	1.000	1.000
2	-4121.357	8317.873	922.860	15	0.916	0.005	0.946	0.036
3	-3838.164	7776.540	564.551	20	0.813	0.041	0.880	0.035
4	**-3664.971**	**7455.208**	**432.664**	**25**	**0.813**	**0.052**	**0.880**	**0.012**
5	-3583.877	7318.074	285.138	30	0.801	0.064	0.869	0.005
6	-3529.966	7235.304	242.057	35	0.784	0.083	0.828	0.005
7	-3487.301	7175.027	176.228	40	0.786	0.089	0.772	0.005
8	-3449.419	7124.316	80.175	45	0.783	0.095	0.832	0.002
High sampling frequency group								
1	-4132.137	8309.951		9	1	0	1	
2	-3538.288	7147.628	1162.323	14	0.950	0.003	0.935	0.036
3	-3302.478	6701.385	446.243	19	0.915	0.01	0.935	0.008
4	**-3085.061**	**6291.927**	**409.458**	**24**	**0.838**	**0.037**	**0.897**	**0.008**
5	-2933.099	6013.378	278.549	29	0.833	0.047	0.882	0.006
6	-2896.334	5965.224	48.154	34	0.787	0.074	0.828	0.006
7	-2863.981	5925.894	39.330	39	0.789	0.076	0.830	0.005
8	-2826.956	5877.219	48.675	44	0.764	0.102	0.780	0.005

Note. LCA = Latent class analysis; LL = log-likelihood; BIC = Bayesian information criterion (based on the Level 2 sample size); Npar = number of parameters; Class. Err. = classification error; AvePP = average posterior class probability; L1 = Level 1; P = careless responding profile on the occasion level. Bold indicates the selected model.

Table 4

Model Fit Statistics for ML-LCA Models With Different Numbers of Level 2 Classes (and Four Level 1 Profiles) in Both Sampling Frequency Groups

Number of classes	LL	BIC	Size of drop in BIC	Npar	Entropy R^2 (L1)	Class. Err. (L1)	Smallest AvePP (L1)	Entropy R^2 (L2)	Class. Err. (L2)	Smallest AvePP (L2)	Size of smallest C
Low sampling frequency group											
1	-3664.971	7455.208		25	0.813	0.052		1.000	0.000	1.000	
2	-3591.438	7328.185	127.023	29	0.810	0.059	0.884	0.788	0.036	0.892	0.162
3	-3569.457	7304.260	23.925	33	0.799	0.068	0.886	0.715	0.097	0.824	0.054
4	**-3557.092**	**7299.578**	**4.682**	**37**	**0.805**	**0.064**	**0.888**	**0.617**	**0.193**	**0.760**	**0.047**
5	-3554.254	7313.944	-14.366	41	0.804	0.065	0.886	0.579	0.244	0.640	0.045
6	-3552.573	7330.625	-16.681	45	0.803	0.067	0.889	0.592	0.245	0.746	0.015
High sampling frequency group											
1	-3085.061	6291.927		24	0.838	0.037	0.897	1.000	0.000	1.000	
2	-3043.964	6230.034	61.893	28	0.843	0.036	0.878	0.504	0.167	0.823	0.441
3	-3018.914	6200.233	29.801	32	0.851	0.033	0.887	0.589	0.151	0.831	0.119
4	**-3016.141**	**6214.989**	**-14.756**	**36**	**0.852**	**0.033**	**0.886**	**0.538**	**0.215**	**0.716**	**0.030**
5	-3014.119	6231.246	-16.257	40	0.854	0.032	0.889	0.530	0.229	0.660	0.009
6	-3013.253	6249.814	-18.568	44	0.854	0.032	0.889	0.484	0.276	0.611	0.009

Note. ML-LCA = Multilevel latent class analysis; LL = log-likelihood; BIC = Bayesian information criterion (based on the Level 2 sample size); Npar = number of parameters; Class. Err. = classification error; AvePP = average posterior class probability; C = classes of individuals; L1 = Level 1; L2 = Level. Bold indicates the selected model.

Table 5

Model Fit Statistics for ML-LCA Models With Different Numbers of Level 1 Profiles (and Four Level 2 Classes) in Both Sampling Frequency Groups

Number of profiles	LL	BIC	Size of drop in BIC	Npar	Entropy R^2 (L1)	Class. Err. (L1)	Smallest AvePP (L1)	Size of smallest P	Entropy R^2 (L2)	Class.Err. (L2)	Smallest AvePP (L2)	Size of smallest C
Low sampling frequency group												
1	-4577.258	9219.653		13	1.000	0.000	1.000		1.000	0.000	1.000	
2	-4095.785	8296.793	922.860	21	0.908	0.007	0.942	0.044	0.241	0.358	0.526	0.01
3	-3793.467	7732.242	564.551	29	0.818	0.040	0.895	0.042	0.495	0.217	0.764	0.011
4	**-3557.092**	**7299.578**	**432.664**	**37**	**0.805**	**0.064**	**0.888**	**0.019**	**0.617**	**0.193**	**0.760**	**0.047**
5	-3394.481	7014.440	285.138	45	0.805	0.096	0.883	0.012	0.759	0.104	0.754	0.024
6	-3253.410	6772.383	242.057	53	0.827	0.100	0.848	0.012	0.857	0.070	0.897	0.070
7	-3145.253	6596.155	176.228	61	0.803	0.123	0.842	0.005	0.860	0.072	0.833	0.051
8	-3085.123	6515.980	80.175	69	0.804	0.126	0.825	0.005	0.872	0.062	0.855	0.058
High sampling frequency group												
1	-4132.137	8325.177		12	1.000	0.000	1.000		1.000	0.000	1.000	
2	-3514.230	7129.964	1195.213	20	0.949	0.003	0.967	0.037	0.207	0.449	0.000	0.038
3	-3251.289	6644.683	485.281	28	0.846	0.026	0.898	0.036	0.465	0.254	0.675	0.040
4	**-3016.141**	**6214.989**	**429.694**	**36**	**0.852**	**0.033**	**0.886**	**0.008**	**0.538**	**0.215**	**0.716**	**0.029**
5	-2836.749	5896.807	318.182	44	0.779	0.098	0.891	0.007	0.758	0.115	0.868	0.053
6	-2678.182	5620.273	276.534	52	0.787	0.105	0.876	0.006	0.747	0.122	0.825	0.075
7	-2594.120	5492.751	127.522	60	0.806	0.113	0.853	0.006	0.800	0.109	0.816	0.217
8	-2547.785	5440.682	52.069	68	0.791	0.124	0.829	0.006	0.812	0.101	0.822	0.223

Note. ML-LCA = Multilevel latent class analysis; LL = log-likelihood; BIC = Bayesian information criterion (based on the Level 2 sample size); Npar = number of parameters; Class. Err. = classification error; AvePP = average posterior class probability; P = careless responding profile on the occasion level; C = classes of individuals; L1 = Level 1; L2 = Level. Bold indicates the selected model.

Table 6

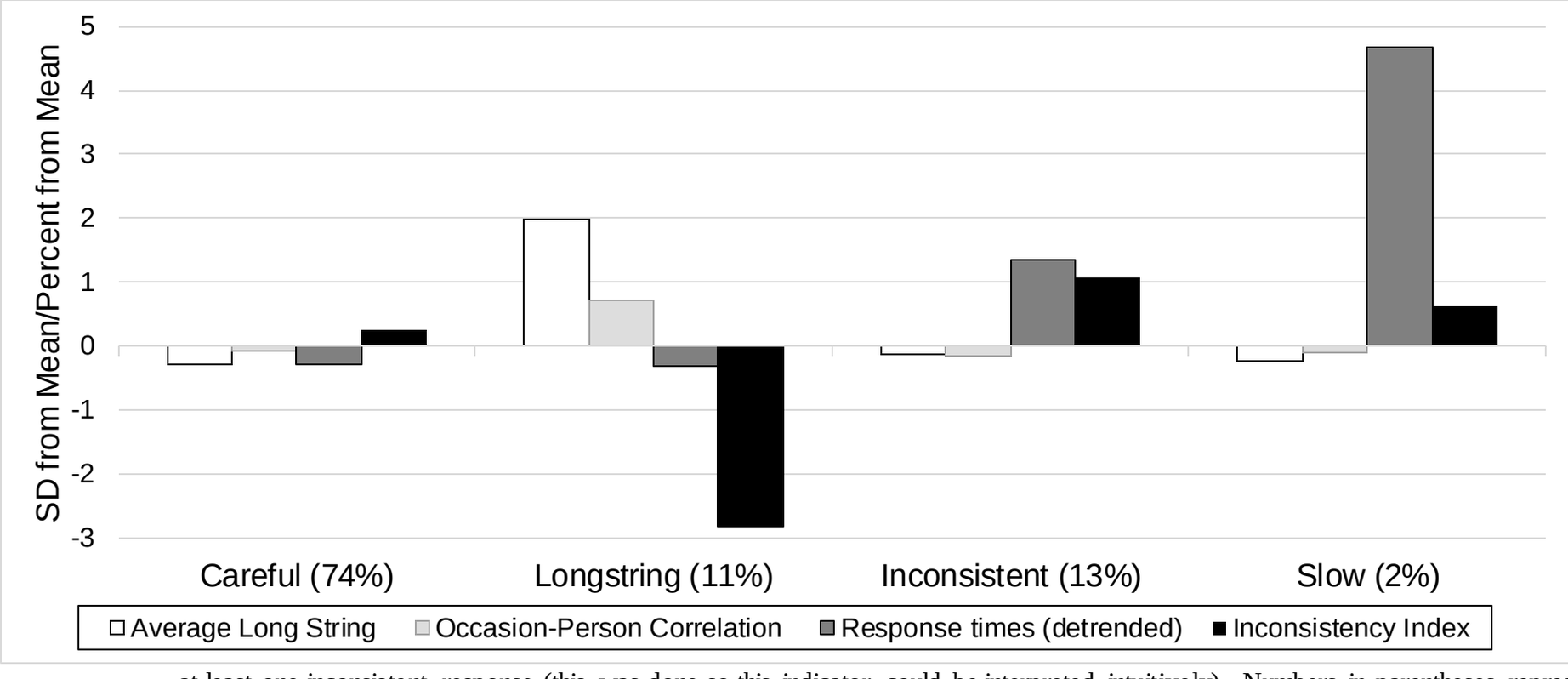
Global Wald Tests and Effect Sizes for Mean Differences in Covariates Across Latent Profiles of Careless Responding at Level 1 and Across Latent Classes of Individuals at Level 2

	Careless responding profiles				Classes of individuals			
Covariate	Wald test χ^2	*df*	*p*	R^2	Wald test χ^2	*df*	*p*	R^2
Time-varying variables								
Momentary motivation	2.807	3	.422	0.002	**36.391**	**3**	**< .001**	**0.072**
Momentary time pressure	7.681	3	.053	0.003	4.720	3	.194	0.029
Time-invariant variables								
Conscientiousness	—	—	—	—	3.005	3	.391	0.016
Agreeableness	—	—	—	—	1.400	3	.706	0.009
Extraversion	—	—	—	—	3.574	3	.311	0.014
Neuroticism	—	—	—	—	2.047	3	.563	0.016
Intellect	—	—	—	—	4.840	3	.184	0.032
Interest in the topic	—	—	—	—	5.082	3	.166	0.037
Financial compensation	—	—	—	—	0.918	3	.821	0.005
Interest in feedback	—	—	—	—	5.346	3	.148	0.032
To do the researcher a favor	—	—	—	—	**20.873**	**3**	**< .001**	**0.027**

Note. Wald test results are based on the recommended maximum likelihood adjusted three-step approach. All *p*-values are two-sided. Bold indicates that the finding remained significant following false discovery rate correction.

Figure 1

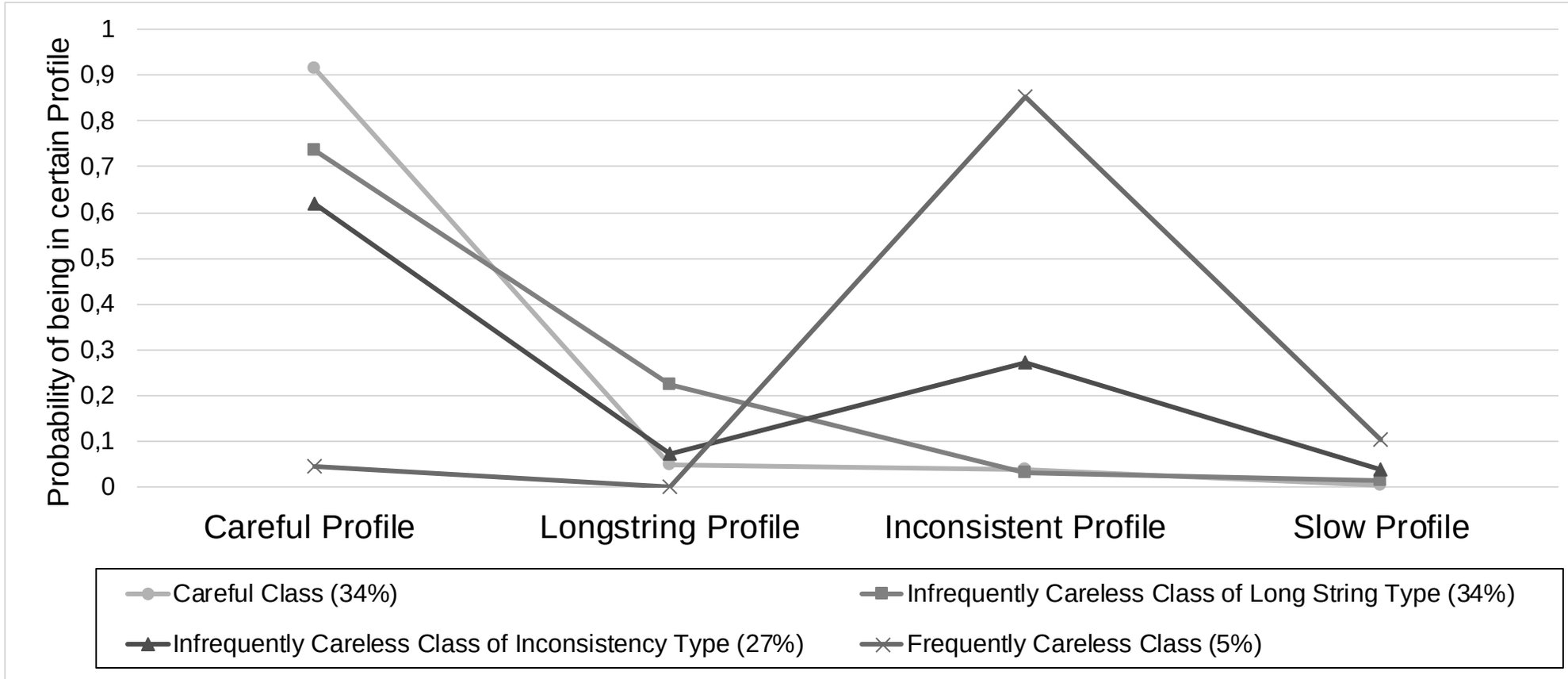
Latent Profiles of Careless Responding Indices at Level 1 (Occasion Level) for the Low Sampling Frequency Group



Note. Bars represent standard deviations from the overall sample mean. In the case of the inconsistency index, which was the only categorical indicator, the bars represent percent deviations of the probability of having at least one inconsistent response from the mean probability of having at least one inconsistent response (this was done so this indicator could be interpreted intuitively). Numbers in parentheses represent profiles sizes (i.e., the percentage of measurement occasions that are assigned to a profile). This is the ML-LCA model with four profiles on Level 1 and four classes of individuals on Level 2 for the low sampling frequency group.

Figure 2

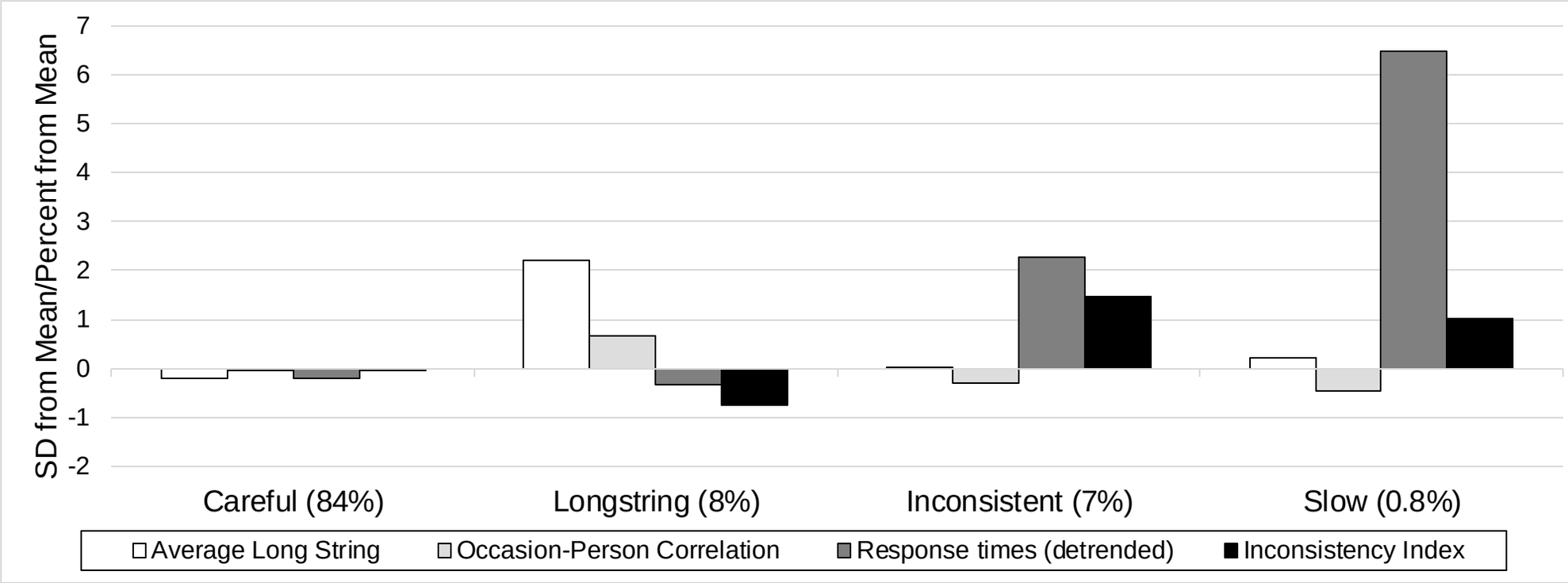
Latent Classes of Individuals Differing in the Distribution of Careless Responding Profiles Over Time for the Low Sampling Frequency Group



Note. Latent classes of individuals differing in the distribution of careless responding profiles (Level 1) over time (ML-LCA model with four profiles on Level 1 and four classes of individuals on Level 2 for the low sampling frequency group). Numbers in parentheses represent class sizes (i.e., the percentage of individuals assigned to a class).

Figure 3

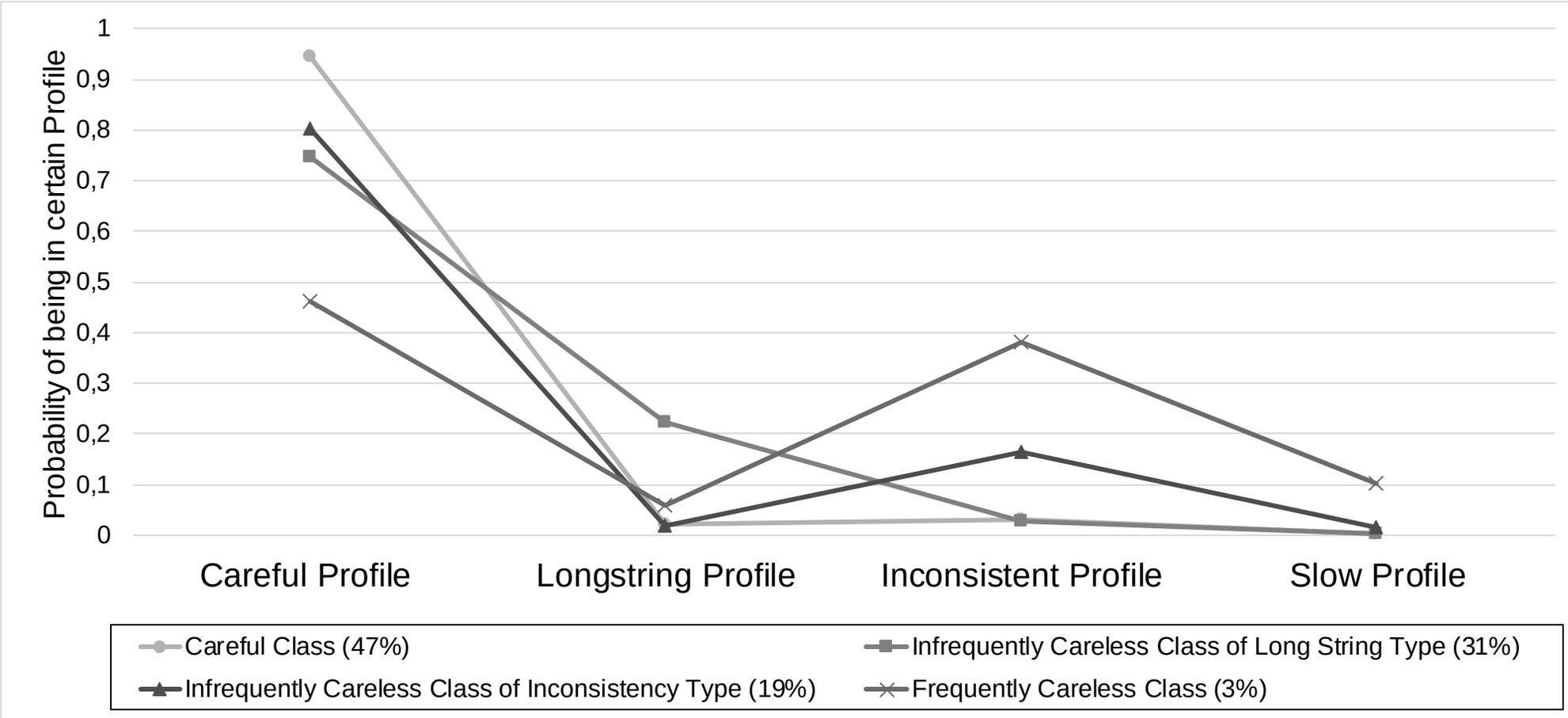
Latent Profiles of Careless Responding Indices at Level 1 (Occasion Level) for the High Sampling Frequency Group



Note. Bars represent standard deviations from the overall sample mean. In the case of the inconsistency index, which was the only categorical indicator, the bars represent percent deviations of the probability of having at least one inconsistent response from the mean probability of having at least one inconsistent response (this was done so this indicator could be interpreted intuitively). Numbers in parentheses represent profiles sizes (i.e., the percentage of measurement occasions that are assigned to a profile). This is the ML-LCA model with four profiles on Level 1 and four classes of individuals on Level 2 for the high sampling frequency group.

Figure 4

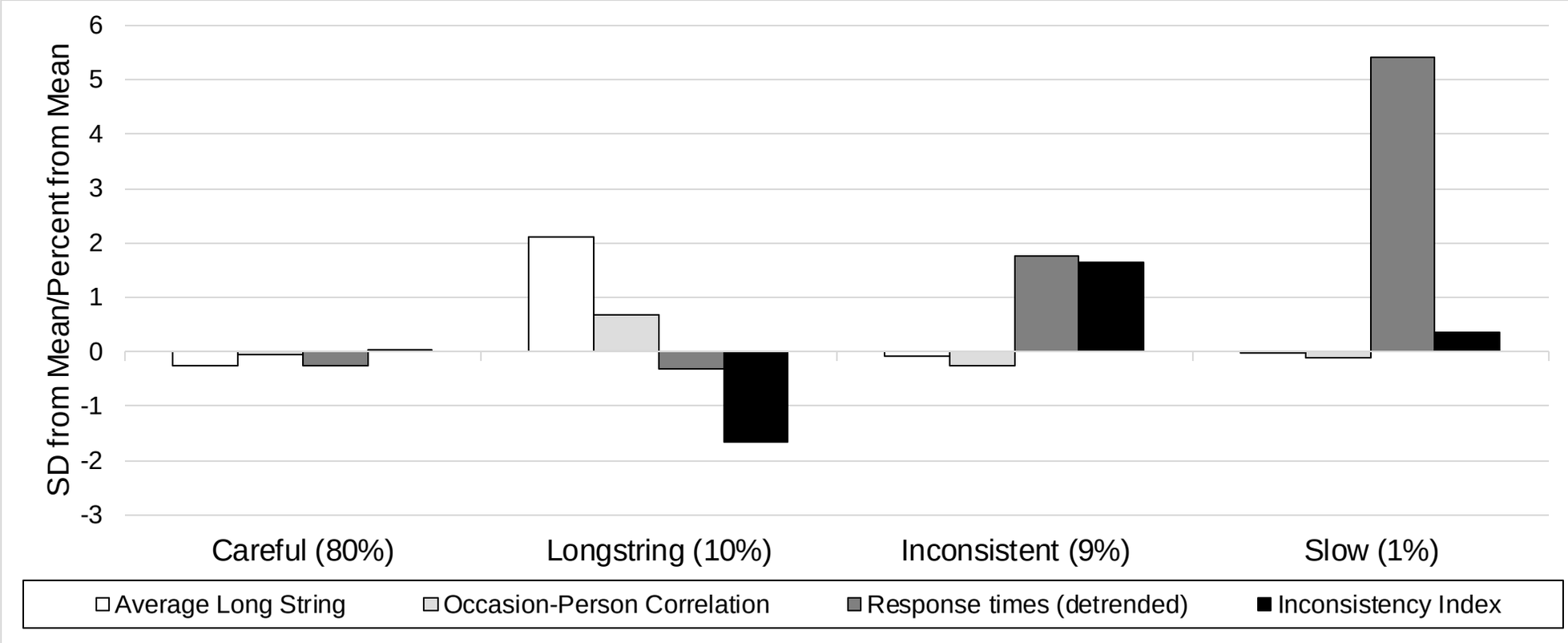
Latent Classes of Individuals Differing in the Distribution of Careless Responding Profiles Over Time for the High Sampling Frequency Group



Note. Latent classes of individuals differing in the distribution of careless responding profiles (Level 1) over time (ML-LCA model with four profiles on Level 1 and four classes of individuals on Level 2 for the high sampling frequency group). Numbers in parentheses represent class sizes (i.e., the percentage of individuals assigned to a class).

Figure 5

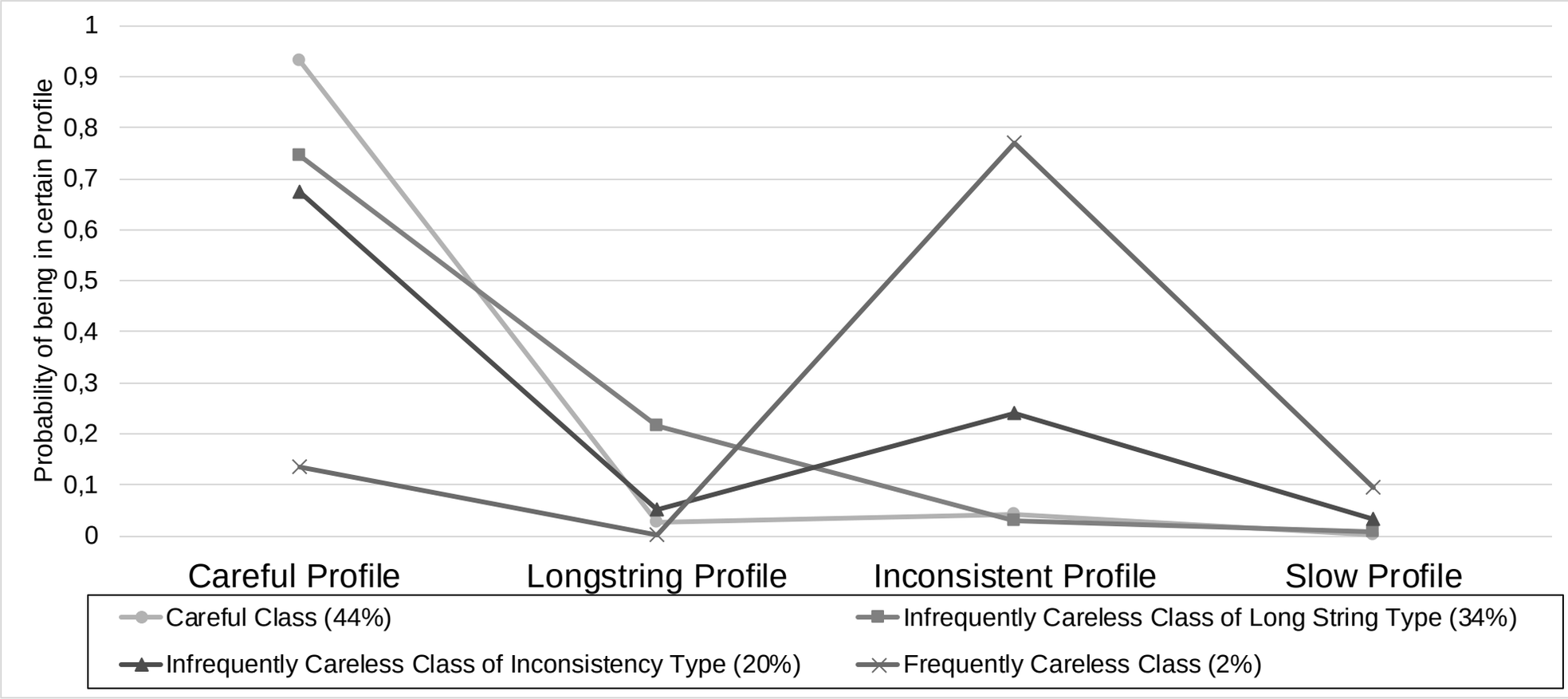
Latent Profiles of Careless Responding Indices at Level 1 (Occasion Level) for the Both Sampling Frequency Groups



Note. Bars represent standard deviations from the overall sample mean. In the case of the inconsistency index, which was the only categorical indicator, the bars represent percent deviations of the probability of having at least one inconsistent response from the mean probability of having at least one inconsistent response (this was done so this indicator could be interpreted intuitively). Numbers in parentheses represent profiles sizes (i.e., the percentage of measurement occasions that are assigned to a profile). This is the final multigroup ML-LCA model with four profiles on Level 1 and four classes of individuals on Level 2.

Figure 6

Latent Classes of Individuals Differing in the Distribution of Careless Responding Profiles Over Time for Both Sampling Frequency Groups



Note. Latent classes of individuals differing in the distribution of careless responding profiles (Level 1) over time (final ML-LCA model with four profiles on Level 1 and four classes of individuals on Level 2). Numbers in parentheses represent class sizes (i.e., the percentage of individuals assigned to a class).

4 Paper 3: Response Styles

Hasselhorn, K., Ottenstein, C., Meiser, T., & Lischetzke, T. (2023). The effects of questionnaire length on response styles in ambulatory assessment. [Manuscript submitted for publication].

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

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

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

Ambulatory assessment (AA) is becoming an increasingly popular research method in the fields of psychology and life science (Hamaker & Wichers, 2017). AA (also known as daily diary, experience sampling, or 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 questionnaire (questionnaire length). Mehl and Conner (2014) provide a detailed explanation about study design considerations and methods of data collection. In the present research, we focused on questionnaire length as a design feature.

The Effects of Questionnaire Length in AA

Considering the very large number of studies that have applied AA in the last decade across diverse fields of research, knowledge about the effects that design choices (e.g.,

questionnaire length) have on aspects of AA data (e.g., data quantity and quality) or participant burden is surprisingly limited (Eisele et al., 2020; Himmelstein et al., 2019). However, over the last few years, an increasing number of researchers have applied experimental designs in AA studies to investigate the effects of design choices on participant burden and aspects of AA data (e.g., data quantity and quality). These experimental AA studies have primarily focused on participant burden, compliance, sampling schedule, sampling contingency, within-person variability, within-person relationships, and careless responding as outcome variables (e.g., Eisele et al., 2020; Hasselhorn et al., 2021; Himmelstein et al., 2019; van Berkel et al., 2019).

Empirical evidence of the effects of questionnaire length was provided by two studies (Eisele et al., 2020; Hasselhorn et al., 2021). With respect to perceived burden and compliance, the study by Hasselhorn et al. (2021) found that participants in the long questionnaire group did not differ from participants in the short questionnaire group in perceived burden or compliance. By contrast, the study by Eisele et al. (2020) found that a longer questionnaire was associated with higher perceived burden and lower compliance. In line with the study by Hasselhorn et al. (2021), most meta-analyses and pooled data analyses have found no association between questionnaire length and compliance (Ottenstein & Werner, 2022; Rintala et al., 2019), but see Morren et al. (2009) for an exception. Given these different results on perceived burden and compliance as outcome variables, more research is needed to elucidate the effects of questionnaire length. With respect to other outcome variables, the study by Hasselhorn et al. (2021) found that a longer questionnaire was associated with lower within-person variability (for momentary mood but not for state extraversion) and a weaker within-person relationship between state extraversion and momentary mood. Furthermore, the study by Eisele et al. (2020) found that a longer questionnaire was associated with higher amounts of careless responding. These results might

indicate that longer questionnaires can lead to reduced (aspects of) data quality in an AA study. This interpretation is in line with the study by Galesic and Bosnjak (2009), who concluded that, in cross-sectional research, a longer questionnaire can produce lower data quality. However, research on effects of questionnaire length on (aspects of) data quality in AA studies has not yet analyzed the potential effects on response styles (RS) as reflections of heuristic processing in the context of AA.

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). Baumgartner and Steenkamp (2001) distinguished between seven common RS: Midpoint RS (MRS), Extreme RS (ERS), acquiescence and disacquiescence RS, noncontingent responding, net acquiescence RS, and response range. In the present research, we focused on MRS and 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., 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

Previous research has proposed a variety of different methods for modeling and accounting for RS, such as count procedures, latent class analytic approaches, or Item Response Theory (IRT) models (Van Vaerenbergh & Thomas, 2013). In recent decades, IRT models have seen an increase in the literature. These models can be divided into two groups: extensions of traditional IRT models for ordinal responses and IRTree models. Traditional IRT models for ordinal responses 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 models represent an alternative framework for assessing and controlling for RS. 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., extensions of traditional IRT models for ordinal responses 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).

IRTree Models

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 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 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 questionnaire) in an AA study on RS. 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. 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 traditional IRT models for ordinal responses see Deng et al., 2018).

Our preregistered[4] hypothesis was that a longer questionnaire would lead to a greater reliance on RS (relative to the substantive trait) in an AA study. In terms of the parameters of the IRTree model in Table 2, our hypothesis can be reformulated as follows: We hypothesized that a longer questionnaire would 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}. The original (preregistered) hypothesis and the reformulated hypothesis can be considered equivalent because 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. Our hypothesis was based on the assumption that a longer questionnaire would be more cognitively demanding for participants than a shorter questionnaire, so that responses are less driven by effortful trait-based processes but mainly by heuristic processes like RS. In a similar vein, Bolt and Johnson (2009) argued that 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.

[4] The preregistration can be found on the OSF repository (https://osf.io/xt3rf/?view_only=c760d31883e649b08c76b66dbfdc7ae3). 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 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. 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.

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. 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).

Procedure

All study procedures were approved by the psychological ethics committee at the university. 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 .72, and the between-person ω was .59. For state conscientiousness, the within-person ω (Geldhof et al., 2014) was .79, and the between-person ω was .80.

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 .84, and the between-person ω was .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 scale ranging from 1 (*not at all*) to 5 (*very much so*). We calculated a mean score across the four items in each dimension such that a higher value indicated a higher standing on the respective personality trait. Revelle's omega total (McNeish, 2018) was .73 for global self-report of extraversion and .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 Table 1 and Table 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, $\theta conc_v$, η_{1v}, η_{2v}, θext_{vj}, $\theta conc_{vj}$, η_{1vj}, η_{2vj}, see Figure 2). We used the Akaike information criterion (AIC) and the Bayesian information criterion (BIC; Burnham & Anderson, 2004) as model selection criteria.

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). [5]

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

[5] 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.

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. 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://osf.io/vw3gf/?view_only=b6f9f08a6b5941eb9c17a4951d1d0cd2).

All models were computed in Mplus (Muthén & Muthén, 1998-2022).

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 (α = 0.277, *SE* = .02) than in the short questionnaire group (α = 0.402, *SE* = .02). Similarly, the α parameter for trait conscientiousness was smaller in the long questionnaire group (α = 0.189, *SE* = .01) than in the short questionnaire group (α = 0.438, *SE* = .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

Impact of RS and Questionnaire Length on Regression Coefficients With Pleasant-Unpleasant Mood as the External Criterion

In the short questionnaire group, state extraversion significantly predicted pleasant-unpleasant mood at the within-person level, β = 0.599, *SE* = 0.012, *z* = 50.607, *p* < .001, and at the between-person level, β = 0.231, *SE* = 0.117, *z* = 1.972, *p* = .049, when RS were accounted for. When RS were not accounted for, the estimated standardized regression coefficients for pleasant-unpleasant mood were β = 0.448, *SE* = 0.017, *z* = 25.621, *p* < .001 at the within-person level, and β = 0.577, *SE* = 0.085, *z* = 6.780, *p* < .001 at the between-person level in the short questionnaire group.

In the long questionnaire group, state extraversion significantly predicted pleasant-unpleasant mood at the within-person level, β = 0.571, *SE* = 0.010, *z* = 55.875, *p* < .001, and at the between-person level, β = 0.560, *SE* = 0.094, *z* = 5.945, *p* < .001, when RS were accounted for. When RS were not accounted for, the estimated standardized regression coefficients for pleasant-unpleasant mood were β = 0.436, *SE* = 0.018, *z* = 24.827, *p* < .001 at the within-person level and β = 0.511, *SE* = 0.088, *z* = 5.816, *p* < .001 at the between-person level. Three out of four regression coefficients were descriptively greater in the models that accounted for RS compared with the models that did not account for RS (at the within-person and between-person levels)

Impact of RS and Questionnaire Length on Regression Coefficients With Global Self-Report of Extraversion and Conscientiousness as External Criteria

In the short questionnaire group, when RS were accounted for, trait extraversion at the between-person level significantly predicted the global self-report of extraversion (measured in the initial online survey), β = 0.276, *SE* = 0.101, *z* = 2.725, *p* = .006, and when RS were not accounted for, the estimated standardized regression coefficient for trait extraversion was β = 0.363, *SE* = 0.100, *z* = 3.624, *p* < .001. A similar picture emerged for conscientiousness in the short questionnaire group: When RS were accounted for, trait conscientiousness at the between-person level significantly predicted the global self-report of conscientiousness (measured in the initial online survey), β = 0.372, *SE* = 0.091, *z* = 4.074, *p* < .001, and when RS were not accounted for, the estimated standardized regression coefficient for trait extraversion was β = 0.498, *SE* = 0.079, *z* = 6.325, *p* < .001.

In the long questionnaire group, when RS were accounted for, trait extraversion at the between-person level significantly predicted the global self-report of extraversion, β = 0.358, *SE* = 0.097, *z* = 3.690, *p* < .001, and when RS were not accounted for, the estimated standardized regression coefficient for trait extraversion was β = 0.523, *SE* = 0.077, *z* = 6.822, *p* < .001. For conscientiousness, in the long questionnaire group, when RS were accounted for, trait conscientiousness at the between-person level significantly predicted the global self-report of conscientiousness β = 0.329, *SE* = 0.120, *z* = 2.753, *p* = .006, and when RS were not accounted for, the estimated standardized regression coefficient for trait conscientiousness was β = 0.299, *SE* = 0.107, *z* = 2.791, *p* = .005. Three out of four regression coefficients were descriptively smaller in the models that accounted for RS compared with the models that did not account for RS (at the within-person and between-person levels).

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 main finding is in line with the study by Hasselhorn et al. (2021), who used the same data as we did in the current paper, and the study by Eisele et al. (2020). The former study found that a longer questionnaire was associated with smaller within-person variability and a weaker within-person relationship between time-varying constructs, and the latter study found that a longer questionnaire was associated with higher careless responding. These findings suggest that a long questionnaire might impair (aspects of) data quality in an AA study. However, these findings were based on only two experimental AA studies that had a limited range of questionnaire lengths (33 vs. 82 items, Hasselhorn et al., 2021; 30 vs. 60 items, Eisele et al., 2020). Clearly, more experimental AA studies on the effects of questionnaire length are needed. Furthermore, on the basis of these 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 & 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. 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 on the impact of RS on the relationship between the modeled traits and criterion variables, we did not observe differences in the interpretation of regression analyses between models that accounted for RS compared with models that did not account for RS. Specifically, the regression coefficients varied (seemingly) unsystematically across the types of models (models that accounted for RS vs. models that did not account for RS), and there were differences in these effects between experimental groups. For example, most (three out of four) regression coefficients were descriptively greater in the models that accounted for RS compared with the models that did not account for RS for pleasant-unpleasant mood as the external criterion (at the within-person and between-person levels). However, for global self-reports of extraversion and conscientiousness (measured in the initial online survey) as external criteria, most regression coefficients (three out of four) were descriptively smaller in the models that accounted for RS compared with the models that did not account for RS, which replicates earlier findings that correlations between different constructs can be inflated if response styles are not controlled for (Böckenholt & Meiser, 2017). Whereas it seems promising that we could not observe any substantial differences in the regression analyses between the types of models in our data, 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 weeks 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. It is possible that the effect of questionnaire length on RS depends, for example, on age (with a more prominent effect in older participants).

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 led to a greater magnitude of RS in our AA study. Although further validation of our findings is necessary, we hope that researchers will consider our findings when planning an AA study in the future.

Declaration of Interest Statement

The author(s) declare that they have no potential conflicts of interest with respect to the research, authorship, or publication of this article.

References

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *British Journal of Mathematical and Statistical Psychology*. 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*. https://doi.org/10.1037/met0000294

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

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 New York. 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). https://doi.org/10.18637/jss.v048.c01

Deng, S., E. McCarthy, D., E. Piper, M., B. Baker, T., & 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* * The original data upon which this paper is based are available at www.hhpub.com/journals/ejpa. *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. (2020). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*. 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*.

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

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. (2021). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01683-6

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

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction*, *114*(4), 609–619. https://doi.org/10.1111/add.14503

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. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12158

Morren, M., Dulmen, S., Ouwerkerk, J., & Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: A systematic review. *European Journal of Pain*, *13*(4), 354–365. https://doi.org/10.1016/j.ejpain.2008.05.010

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 edition). Muthén & Muthén.

Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What Affects the Completion of Ecological Momentary Assessments in Chronic Pain Research? An Individual Patient Data Meta-Analysis. *Journal of Medical Internet Research*, *21*(2), e11398. https://doi.org/10.2196/11398

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

Ottenstein, C., & Werner, L. (2022). Compliance in Ambulatory Assessment Studies: Investigating Study and Sample Characteristics as Predictors. *Assessment*, *29*(8), 1765–1776. https://doi.org/10.1177/10731911211032718

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 *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Elsevier. 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

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. https://doi.org/10.1037/pas0000662

Soyster, P. D., Bosley, H. G., Reeves, J. W., Altman, A. D., & Fisher, A. J. (2019). Evidence for the Feasibility of Person-Specific Ecological Momentary Assessment Across Diverse Populations and Study Designs. *Journal for Person-Oriented Research*, *5*(2), 53–64. https://doi.org/10.17505/jpor.2019.06

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Mehrdimensionaler Befindlichkeitsfragebogen*. Hogrefe.

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

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental Disorders: Meta-Analysis and Recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., & Kostakos, V. (2019). Context-Informed Scheduling and Analysis: Improving Accuracy of Mobile Self-Reports. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300281

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: *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

Table 7

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	$\frac{\exp\left(y_{1vi}\left(\eta_{1v} - \beta_{1i}\right)\right)}{1 + \exp\left(\eta_{1v} - \beta_{1i}\right)}$
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	$\frac{\exp\left(y_{3vi}\left(\eta_{2v} - \beta_{3i}\right)\right)}{1 + \exp\left(\eta_{2v} - \beta_{3i}\right)}$

Table 8

Definition of Pseudoitems and Node Probabilities for the Two-Dimensional Parametrization of Extreme Responding for the IRTree Model in Figure 1

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

Table 9

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

Model	Within-person level			Between-person level			AIC	BIC	Npar
	unidim	two-dim		unidim	two-dim				
	α = 0	equal α	freely est. α	α = 0	equal α	freely est. α			
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 α= α parameters set equal across the two substantive constructs (extraversion, conscientiousness); freely est. α = α 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.

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)

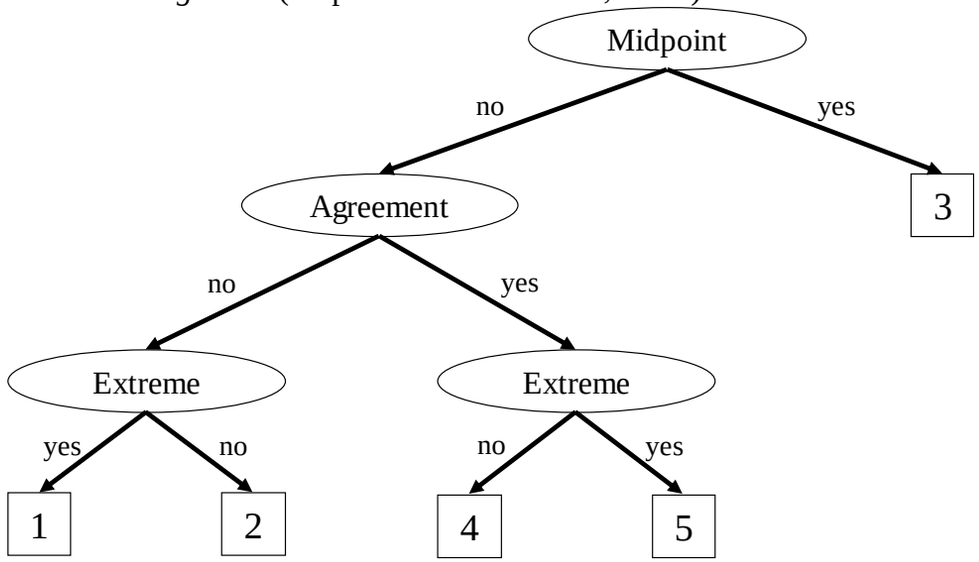
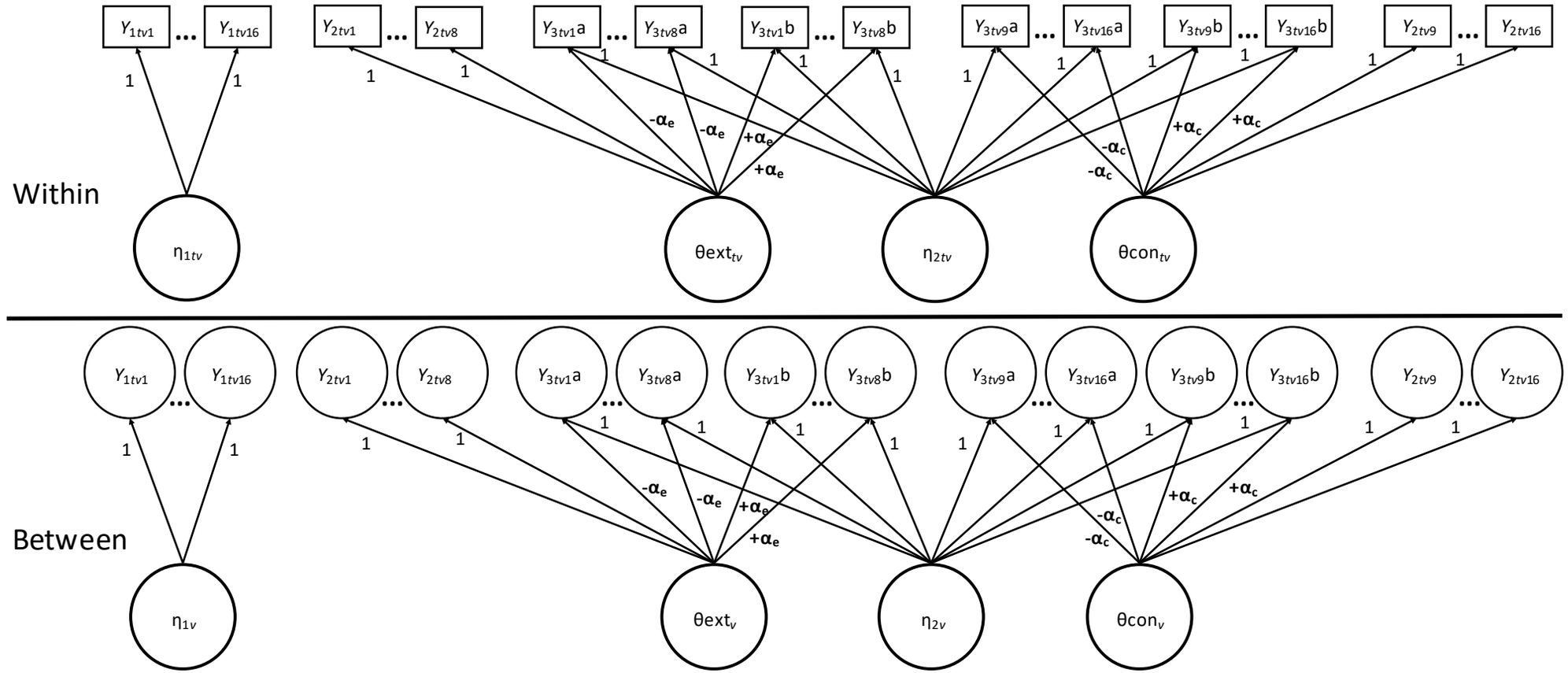


Figure 2

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.

Figure Captions (as a list)

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)

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

5 General Discussion

This present dissertation aimed to experimentally manipulate aspects of an AA study's sampling strategy - sampling frequency (Study 1) and questionnaire length (Study 2) - and to investigate their impact on perceived burden, data quantity, and aspects of data quality. Specifically, I aimed to examine the effects of sampling frequency and questionnaire length on participant burden, compliance, within-person variability, and the within-person relationship between time-varying variables. Furthermore, I wanted to investigate the effects of sampling frequency on careless responding, and the effects of questionnaire length on the relative impact of RS in the context of AA.

In the General Discussion, I will first summarize the key findings for each research question, critically reflect on the findings, and discuss their specific implications for future researchers that are interested in conducting AA studies (5.1). I will then discuss general implications of the findings that go beyond the scope of a single outcome variable (5.2). Finally, I will discuss some limitations and future research directions (5.3), before drawing a conclusion (5.4).

5.1 Summary and Outcome-Specific Implications

5.1.1 Experimental Manipulations of Sampling Frequency and Questionnaire Length

To investigate my research questions, two experimental AA studies were conducted, manipulating either sampling frequency (Study 1) or questionnaire length (Study 2) with two experimental conditions. In Study 1, participants received either 3 or 9 questionnaires per day for the first 7 days of the study.[6] In Study 2, participants received either a 33-item or an 82-

[6] Note that after the first 7 days, a second 7-day AA phase followed immediately in which the sampling frequency was switched between groups. This was done to ensure that each participant spent a comparable amount of time participating in the study, so that the financial compensation, which was the same for both groups, was fair. However, the analyses presented in this dissertation are based on the first 7-day AA phase, because the focus of this dissertation was on the between-group comparisons (high vs. low sampling frequency).

item questionnaire three times a day for 14 days. For an overview of which outcome variables were associated with which empirical studies, see Figure 1.

5.1.2 Participant Burden

In the first paper (Chapter 2), I examined the effects of sampling frequency and questionnaire length on (perceived) participant burden. Thereby, I investigated two different measures of participant burden, which were the same for both AA studies: The first measure was daily perceived burden, which referred to participants' perceived burden of each day (measured in the last questionnaire per day), and the second measure was retrospectively perceived burden, which referred to participants' perceived burden across the AA phase (measured at the end of the AA phase). I hypothesized that a higher sampling frequency (Study 1, RQ1A) and a higher questionnaire length (Study 2, RQ1B) would lead to higher (daily and retrospective) participant burden. My analyses revealed that a higher sampling frequency, but not a higher questionnaire length, led to higher (daily and retrospective) participant burden.

With respect to sampling frequency the results were in line with previous assumptions (Moskowitz et al., 2009; Ono et al., 2019; Piasecki et al., 2007; Roekel et al., 2019; Santangelo et al., 2013; Wen et al., 2017) and with the empirical study by Stone et al. (2003). By contrast, Eisele et al. (2020) found no effect of sampling frequency on participant burden (daily and retrospective). This may have been due to the fact that in the study by Eisele et al., the effect of sampling frequency "was canceled out by the increased motivation due to the higher incentive" (Eisele et al., 2020, p. 12) in the high sampling frequency group (40€ vs. 80€ in the group with three vs. nine AA questionnaires per day, respectively), whereas in our study the financial compensation for the entire study did not differ between the experimental groups.

With respect to questionnaire length, the results were contrary to previous assumptions (Dworak, 2022; Ono et al., 2019; Piasecki et al., 2007; Roekel et al., 2019) and the empirical study by Eisele et al. (2020) who found that a longer questionnaire led to a higher participant burden (daily and retrospective). While the experimental groups (across studies) had a comparable number of items (30 vs. 60 items per questionnaire in Eisele et al.'s study, and 33 vs. 82 items per questionnaire in our study), there were some differences in how the greater number of items was achieved, and in the sampling strategy, which might explain the different results across studies: In Eisele et al.'s study, most of the measured constructs were the same across groups, but the long questionnaire group had to respond to two extra questions about the pleasantness of the most important event and the stressfulness of situations since the last questionnaire. These questions may have caused participants in the long questionnaire group to contemplate their daily negative experiences more and may have contributed to the finding that participants in the long (versus short) questionnaire group perceived the study as more burdensome. Furthermore, Eisele et al. (2020) implemented a 90-second time limit for participants to start an AA assessment (at random assessment times), whereas in my Study 2, participants were given 45 minutes to complete an AA assessment (at fixed assessment times). Therefore, participants in my Study 2 were able to organize their (daily life) activities around the fixed assessment times, compared to the random assessment times in Eisele et al.'s study, which may have further reduced the effect of a longer questionnaire on participant burden in our study, because the assessment prompts interfered less with the daily activities in our Study 2.

I had expected that a higher sampling frequency and a longer questionnaire would have the same effect on participant burden, because I predicted that an objective increase in the time required to participate in the study (more questionnaires per day, as in Study 1, or more items per questionnaire, as in Study 2) would lead to a higher participant burden. This

assumption is consistent with previous research, suggesting that the sampling frequency and the questionnaire length (together with the study duration) are measures of participant burden (e.g., Ono et al., 2019; Wrzus & Neubauer, 2022). However, the different manipulations had different effects in the studies presented in this dissertation and in the study by Eisele et al. (2020).

In fact, the assumptions that sampling frequency and questionnaire length are measures of participant burden, or that the objective time required to participate in a study, is the reason why study participants experience greater burden, may be incorrect. Furthermore, I hypothesize that there are different psychological processes driving the effects of sampling frequency compared to questionnaire length: Participants in AA studies have agreed in advance to participate in the study and to the study's sampling strategy. Participants who are unwilling to exert the effort required to respond to the items in the given study protocol would most likely not participate in the AA study. Note that selection bias in AA studies is discussed in Chapter 5.3. In the AA phase of the study, participants are prompted to complete a set of items within a given time frame in their daily life. This prompt interferes with their current activities and thus induces a burden on participants. Therefore, studies with a higher sampling frequency interfere more often with the daily life of study participants, thus inducing a higher participant burden. Regarding the effect of questionnaire length on participant burden, I hypothesize that the psychological process that is affected by the questionnaire length is the effort (or cognitive load) required to respond to the items of the questionnaire and provide data of sufficient quality. As a result, a longer questionnaire would cause a higher response effort, thus possibly affecting other aspects of response behavior (e.g., data quantity or aspects of data quality), as participants may attempt to reduce their effort to respond to the items of the questionnaire, but not to higher participant burden. In the following, I will refer to these psychological processes, which are assumed to be different psychological processes that drive

the effects of sampling frequency versus questionnaire length in the current dissertation, as *hypothetical psychological processes*.

Importantly, the hypothetical psychological processes for sampling frequency and questionnaire length can explain the results and the differences between the results of the current dissertation and the results of Eisele et al. (2020). I measured (daily and retrospective) participant burden with items, which were adapted from the study by Stone et al. (2003), that directly relate to the number of interferences that participants experience through participating in the studies (e.g., with the item "How much did participating in the study interfere with your usual activities?"). In contrast, in the study by Eisele et al., they measured participant burden with items that related to the effort required to respond to the items of the assessment for momentary perceived burden and (e.g., "Filling in the questionnaire took effort") and for retrospective perceived burden (e.g., "Did the questionnaire become boring during the study?", or "Did your motivation to respond to the beeps decrease during the weeks?"). Consistent with the hypothetical psychological processes, I would conclude for the studies presented in the current dissertation that a higher sampling frequency, but not a longer questionnaire, would increase the participant burden. For the study by Eisele et al. (2020), I would conclude that a longer questionnaire, but not a higher sampling frequency, would increase the participant burden. These expectations are in line with the results of the current dissertation and the results of Eisele et al. (2020). Additionally, there is some indirect evidence to suggest that the effects of sampling frequency and questionnaire length are not based on the same psychological process: Eisele et al. (2020) found no interaction between the effects of sampling frequency and questionnaire length on participant burden. However, the presented empirical studies did not include measures that capture both hypothetical psychological processes. Therefore, these hypothetical psychological processes remain up for

debate. Clearly, more research is needed to demonstrate whether the effects of sampling frequency and questionnaire length are based on the same psychological process(es).

5.1.3 Compliance

In the first paper (Chapter 2), I examined the effects of sampling frequency and questionnaire length on compliance. I hypothesized that a higher sampling frequency (Study 1, RQ2A) and a longer questionnaire (Study 2, RQ2B) would lead to lower compliance. Contrary to my expectations, my analyses revealed that a higher sampling frequency and a longer questionnaire did not lead to lower compliance. These results are consistent with a large body of previous research (Conner & Reid, 2012; Eisele et al., 2020; Jones et al., 2019; McCarthy et al., 2015; Ono et al., 2019; Ottenstein & Werner, 2022; Soyster et al., 2019; Stone et al., 2003; Walsh & Brinker, 2016). Some exceptions to these are the meta-analysis by Vachon et al. (2019), which found that a higher sampling frequency, but not a longer questionnaire, led to lower compliance, and the meta-analysis by Morren et al. (2009), who found that a longer questionnaire led to lower compliance. In addition, the study by Eisele et al. (2020), which to my knowledge is the only study to experimentally manipulate questionnaire length (and the sampling frequency), found that longer questionnaires, but not sampling frequency, led to lower compliance. However, the study by Eisele et al. (2020) differed in two ways that may explain the differences in the results regarding the effects of questionnaire length. Eisele et al. (2020) implemented a 90-second time limit for participants to start an AA assessment (at random assessment times), whereas in my Study 2, participants were given 45 minutes to complete an AA assessment (at fixed assessment times). Thus, participants in Eisele et al.'s study may not have been able to respond to the questionnaire during situations requiring their full attention (e.g., a conversation, cooking), whereas participants in my Study 2 could delay their response for a few minutes, leading to a higher compliance rate. Additionally, as described in Chapter 5.1.2, participants in my Study 2 were

able to organize their (daily life) activities around the fixed assessment times (to ensure that they were able to respond to the questionnaire with their full attention), compared to the random assessment times in Eisele et al.'s study, which may have further reduced the effect of a longer questionnaire on compliance.

Another possible explanation for not finding an effect of sampling frequency or questionnaire length on compliance in our study is that I provided personal feedback as an incentive, which could have counteracted the decrease in compliance. Furthermore, similar to many other studies, I financially incentivized participants to reach certain levels of compliance in both groups within both studies. This was necessary to maintain standard procedures, as the use of financial incentives tied to compliance is a very typical characteristic in AA studies (cf. Trull & Ebner-Priemer, 2020).

With regard to the psychological processes, the results of the current dissertation indicate that the higher participant burden reported in the high sampling frequency group (see Chapter 5.1.2) did not translate into a lower data quantity. This is contrary to previous assumptions that participants might try to reduce the burden by not completing a particular measurement occasion (Stone et al., 2003; Vachon et al., 2019). One possible explanation is that participants in the high sampling frequency group maintained high levels of compliance (despite perceiving a higher burden) because they desired accurate post-study feedback that reflected their experience and behavior during the study.

With regard to the hypothetical psychological processes (see Chapter 5.1.2), I would conclude that there is no effect of burden (induced by a higher sampling frequency) on compliance. However, I would conclude that a longer questionnaire would increase the effort required to respond to the items of the questionnaire and thus reduce compliance. One possible explanation for not finding an effect of questionnaire length on compliance in my Study 2 and in most meta-analytic or pooled data analyses is that the true effect may be very

small, because compliance may depend on other factors, such as being alone (Rintala et al., 2020) or experiencing high positive affect (Sokolovsky et al., 2014), and participants might engage in more subtle response behaviors in order to cope with the lower momentary motivation (to start and complete the questionnaire). Therefore, participants might be more likely to provide data that is not of sufficient quality, such as responding carelessly (Jones et al., 2019), compared to not responding to the questionnaire. In addition, the effect of questionnaire length on compliance could be counteracted by the link between the incentives given to participants and compliance (described above), or by different study characteristics (e.g., the amount of personal contact with participants). Note that in the study by Eisele et al. (2020), the incentives given to participants were not linked directly to compliance, which might explain why they found an effect of questionnaire length on compliance. In their study, participants were told that they needed to complete "enough beeps" to receive full payment at the end of the study. Clearly, more (experimental) research is needed to investigate the effects of questionnaire length on compliance, and whether these effects depend on the link between the incentives given to participants and compliance. In the following, I will discuss the results of the effects of sampling frequency and questionnaire length on (aspects of) data quality, which represent more subtle response behaviors (mentioned above).

5.1.4 Within-Person Variability and Within-Person Relationship Between Time-Varying Variables

In the first paper (Chapter 2), I examined the effects of sampling frequency and questionnaire length on within-person variability and the within-person relationship between time-varying variables. I hypothesized that a higher sampling frequency (Study 1) and a longer questionnaire (Study 2) would lead to lower within-person variability (RQ3A for sampling frequency and RQ3B for questionnaire length) and a lower within-person relationship between time-varying variables (RQ4A for sampling frequency and RQ4B for

questionnaire length). Thereby, I investigated two different constructs (for RQ3A and RQ3B), namely momentary pleasant-unpleasant mood (momentary mood) and state extraversion, which were the same for both AA studies, and the association between these constructs. My analyses revealed that, contrary to my expectations, higher sampling frequency did not lead to a lower within-person variability (RQ3A) in momentary mood or state extraversion, or to a lower within-person relationship between momentary mood and state extraversion (RQ4A). However, my analyses showed that a longer questionnaire led to a lower within-person variability in momentary mood, but not in state extraversion, and to a lower within-person relationship between momentary mood and state extraversion. To my knowledge, the studies presented in the current dissertation were the first to analyze the effects of experimentally manipulated sampling frequency and questionnaire length on within-person variability, and on the within-person relationship between time varying variables.[7] The results, that the degree of within-person variability in momentary mood and the relationship between state extraversion and momentary mood were lower in the long questionnaire group (vs. the short questionnaire group) support the idea that participants in the long questionnaire group provided less nuanced responses to repeated questionnaires (assessments), using a more heuristic approach (Fuller-Tyszkiewicz et al., 2013; Podsakoff et al., 2019). However, the effect on within-person variability was less pronounced for state extraversion, and the difference between groups was not statistically significant. This could be due to the fact that mood was assessed at or near the end of the AA questionnaire, whereas state extraversion was evaluated at the beginning of the AA questionnaire. Studies on positioning effects in cross-sectional surveys (Galesic & Bosnjak, 2009) have shown that items evaluated further away from the beginning of the questionnaire have lower within-person variance, which could

[7] Note that the recent study by Eisele et al. (2023) was published after the manuscript of the empirical Chapter 2 was published.

explain the discrepancy between the effects of questionnaire length on momentary mood and state extraversion.

The result that questionnaire length leads to a lower within-person variability is important for future researchers to consider when planning an AA study. If researchers try to examine within-person relationships (between time-varying variables), and the constructs of interest have a relatively low within-person variability in general, a long questionnaire might cause a further decline in the within-person variability. Therefore, the constructs might no longer have meaningful within-person variance anymore, which is required to examine within-person relationships between time-varying variables (Heck & Thomas, 2015; Hox, 2002; Raudenbush & Bryk, 2002). I would suggest, based on the results of the current dissertation, that researchers interested in studying within-person relationships between time-varying variables that generally have a relatively low within-person variability should avoid long questionnaires, or at least place the most important questions at the beginning of each questionnaire, in order to ensure that meaningful within-person variability can be observed.

With respect to the assumed psychological processes, that the sampling frequency and the questionnaire length (together with the sampling duration) have the same effects on the within-person variance and on the within-person relationship (between momentary mood and state extraversion), the finding that a higher sampling frequency did not lead to a lower within-person variance and a lower the within-person relationship (between momentary mood and state extraversion) is difficult to explain. However, these findings are in line with the hypothetical psychological processes (described in Chapter 5.1.2), that a higher sampling frequency only increases participant burden, but not other aspects of data quantity and quality. Additionally, the hypothetical process that questionnaire length increases the effort (cognitive load) required to respond to the items of the questionnaire might explain the discrepancy between the effects of questionnaire length on the within-person variability of momentary

mood and state extraversion. If the cognitive load of participants is low at the beginning of each questionnaire, and increases as the participants respond to more items, the participants are more likely to provide data that is not of sufficient quality at the end of the questionnaire. In other words, the participants' willingness to exert effort required to respond to the items, and provide high data quality, declines over the course of the questionnaire. This idea is supported by the study on positioning effects by Galesic and Bosnjak (2009) described above. Nevertheless, we can only speculate about how different design features may place different types of demands on participants in an AA study. Further research is needed to scrutinize the underlying processes.

5.1.5 Careless Responding

The aim of the second paper (Chapter 3) was to identify latent profiles of momentary careless responding at the occasion level (Level 1) and to differentiate between latent classes of individuals at the person level (Level 2) who differed in their use of careless responding over time using (multigroup) ML-LCA and investigate the potential effects of sampling frequency on careless responding. I expected to identify the same number of latent profiles of momentary careless responding at the occasion level (Level 1) and latent classes of individuals at the person level (Level 2) who differed in their use of careless responding over time using both sampling frequency groups. Furthermore, I expected that a high sampling frequency (vs. a low sampling frequency) would lead to a higher proportion of participants that are assigned to a careless responding class on Level 2 (RQ5). Within my analyses, I could identify four momentary careless responding profiles on Level 1 (the careful, inconsistent, long string, and slow responding profiles) and four classes of individuals who differed in their use of the momentary careless responding profiles over time on Level 2 (a careful class, an infrequently careless class of the long string type, an infrequently careless class of the inconsistent type, and a frequently careless class) in both sampling frequency groups in Study

1. The latent typology of momentary careless responding profiles on Level 1 corresponded closely with previous cross-sectional research (Goldammer et al., 2020; Kam & Meyer, 2015; Li et al., 2022; Maniaci & Rogge, 2014; Meade & Craig, 2012). This indicates that different types of careless responding, such as long strings and inconsistent careless responding, can also be identified in AA data using unobtrusive indices. Contrary to my expectation, the sampling frequency had no effect on careless responding in AA because all parameters were equal across groups (including the class sizes at Level 2). This finding is consistent with the study by Eisele et al. (2020), who found that the sampling frequency had no effect on careless responding. Moreover, this result indicates that, contrary to the assumption of Jones et al. (2019), the higher participant burden reported in the high sampling frequency group (see Chapter 5.1.2) did not translate to more careless responding. The finding that a higher sampling frequency does not lead to more careless responding is in line with the hypothetical psychological processes (described in Chapter 5.1.2), that a higher sampling frequency only increases participant burden, but not other aspects of data quantity and quality. Interestingly, the hypothetical psychological processes and the results of the studies by Eisele et al. (2020) suggest that the questionnaire length (in AA) may influence careless responding (and aspects of data quality). The idea that the participants' response effort to complete the questionnaire is influenced by the questionnaire length is supported by the result in Chapter 3 that the person-variable (aggregated) motivation to answer questionnaires was related to careless responding, because a higher motivation might increase the effort that participants are willing to exert while responding to the items of the questionnaire. However, future research should elucidate the effects of questionnaire length and other design features on careless responding in AA.

5.1.6 Response Styles

The aim of the third paper (Chapter 4) was to investigate the effects of questionnaire length on RS in an AA study. I used multigroup multidimensional IRTree models in an

MSEM framework to test whether a longer questionnaire would lead to a greater impact of RS relative to the substantive trait in an AA study. In line with my research question (RQ6), the result showed that in the group with a longer questionnaire (vs. a shorter questionnaire) the substantive trait had less influence on the fine-grained decision between an extreme and a nonextreme response. That implies that, as predicted, participants in the longer questionnaire group were more strongly affected by RS in relation to the trait. This result is in line with the study by Eisele et al. (2020), who found that a longer questionnaire was associated with more careless responding, which is another threat to the data quality in AA studies.[8]

With respect to the psychological process(es) behind the differential effects of questionnaire length, I assumed that a longer questionnaire (vs. a shorter) may impose a higher cognitive load, or a higher perceived burden (see RQ1B), on participants as they attempt to complete such a questionnaire. Furthermore, I assumed that, as a consequence, participants may be more driven by heuristic processes like RS in an attempt to reduce cognitive load or perceived burden (Bolt & Johnson, 2009; Jones et al., 2019; Knowles & Condon, 1999). However, the results of Chapter 2 suggest that participants did not perceive a higher burden in the long questionnaire group (vs. the short questionnaire group), thus indicating that a higher participant burden, contrary to Jones et al.'s (2019) assumption, may not be the underlying psychological process that led to a greater relative impact of RS. On the other hand, it appears obvious that participants would experience a higher cognitive load when attempting to complete a questionnaire with more items than one with fewer items (Bolt & Johnson, 2009; Knowles & Condon, 1999). Note that I propose the same hypothetical psychological processes (described in Chapter 5.1.2) as Bolt and Johnson (2009) and Knowles and Condon (1999). However, cognitive load was not directly assessed in Study 2, so

[8] Note that, if we assume that the different manipulations (sampling frequency and questionnaire length) have the same effects on the data quality in AA studies, the results of Chapter 3 are in contrast to the results observed in Chapter 4.

alternative explanations of the underlying psychological processes cannot be ruled out. Furthermore, future research should elucidate the effects of sampling frequency and other design features on careless responding in AA.

5.2 General Implications

The investigation of the effects of sampling frequency and questionnaire length on several outcome variables has implications on methodological, theoretical, and practical levels. In the following, I will describe these general implications of the results of the current dissertation that go beyond the scope of a single outcome variable.

At a theoretical level, the results of the present dissertation were that sampling frequency only affects participants' perceived burden, but not other aspects of AA data (compliance, within-person variability, within-person relationship between time-varying variables, and careless responding), whereas questionnaire length does not affect participants' perceived burden or compliance, but other aspects of AA data (within-person variability, within-person relationship between time-varying variables, and RS). These results indicate that researchers, who plan to conduct an AA study in the future, should not consider these design features (sampling frequency and questionnaire length) as interchangeable when trying to strike a balance between being able to collect rich information, not overburdening participants (Carpenter et al., 2016), and not compromising aspects of AA data (e.g., data quantity and data quality; Arslan et al., 2020; May et al., 2018). Specifically, researchers can expect that an AA study that assesses 30 items per measurement occasion nine times a day (over 14 days) would not result in the same participant burden or the same data quality as an AA study that assesses 90 items per measurement occasion three times a day (over 14 days). Therefore, I suggest that researchers should try to decide on each design feature separately when planning an AA study.

The results that the sampling frequency only affects participants' perceived burden, but not other aspects of AA data, suggests that researchers can increase the sampling frequency as much as they like, but at least to 9 assessments per day, without compromising data quantity and (aspects of) data quality. However, it does not seem realistic that participants would endure any sampling frequency (e.g., 50 measurement occasions per day over 14 days)[9] without trying to reduce the amount of effort they have to invest as they attempt to complete such a sampling strategy. One way to reduce their effort could be for individuals would not to enroll in AA studies that have a very high sampling frequency in the first place. Another possibility is that the sampling frequency used in Study 1 (nine measurement occasions per day) was not high enough to induce changes in data quantity and (aspects of) data quality. There might be a threshold (in terms of number of measurement occasions per day) at which changes in data quantity or (aspects of) data quality occur. Furthermore, this threshold may be influenced by factors other than the sampling frequency, such as the population of interest or participants' momentary positive and negative affect. For example, participants currently suffering from affective disorders could perceive higher momentary burden by comparison to healthy participants (van Genugten et al., 2020). Participants perceiving higher momentary positive affect might be reporting less perceived burden, whereas a higher negative affect might lead to a higher perceived burden (van Genugten et al., 2020). Because AA has been used extensively in the field of clinical psychology (Vachon et al., 2019), these factors are important to consider when researchers plan an AA study.

In AA studies, researchers typically use short questionnaires with only a few items per subscale or even single item subscales and there are no common guidelines or best practices

[9] Note that a sampling frequency of 50 measurement occasions per day has already been used, over four days, in the study by Kuppens et al. (2010).

for how many items researchers should use to measure subscales. Additionally, there are different suggestions and (cross-sectional) empirical studies providing mixed evidence on how many items should be used: For example, Shrout and Lane (2012) suggest that using at least three items for each construct under study, while cross-sectional studies have found that single-item subscales perform as well as multi-item scales (Bergkvist & Rossiter, 2009; Wanous et al., 1997), and other cross-sectional studies have found that multi-item measures are more predictively valid (Warren & Landis, 2007). On the other hand, while the results of the current dissertation suggest that questionnaire length influences (aspects of) data quality, I did not observe differences in the interpretation of the regression analyses in the empirical Chapters 2 and 4 between both experimental groups in Study 2.[10] Therefore, using very few items per subscale, or even single item subscales, may be more detrimental to researchers' goals (e.g., studying within-person relationships between time-varying variables) or the data quality (e.g., the validity and reliability of the subscales) than the costs of including at least three items per subscale (Shrout & Lane, 2012). However, there may be AA studies that have to use single items measures in order to enable the participants to complete the questionnaire in a short timeframe. Therefore, I suggest that researchers explain why they could not include more items per subscale, or at least critically reflect on the number of items they used to measure each subscale.

5.3 Limitations and Future Research Directions

I have discussed several limitations in empirical Chapters 2 to 4. In the following, I will highlight some more general limitations of the current dissertation and future research directions.

[10] The interpretation of the relationship between state extraversion and momentary mood in Chapter 2, and the interpretation between models that accounted for RS compared with models that did not account for RS in Chapter 4 remained the same across experimental groups.

As a starting point, empirical Chapter 4 of this dissertation focused on the effects of questionnaire length on the relative impact of RS and did not analyze the effects of sampling frequency on the relative impact of RS. Therefore, I do not know whether a higher sampling frequency may lead to a higher relative impact of RS. In most empirical studies, and in the current dissertation, the sampling frequency only influenced participant burden, but not the data quantity or any aspect of data quality (Conner & Reid, 2012; Eisele et al., 2020; McCarthy et al., 2015; Ono et al., 2019; Soyster et al., 2019; Stone et al., 2003; Walsh & Brinker, 2016). However, as described in Chapter 5.2, it seems unrealistic that participants would endure any sampling frequency without attempting to reduce the amount of effort they invest to complete the sampling strategy of an AA study. Future research should investigate the effects of sampling frequency on the relative impact of RS, as participants may rely more on their RS if the sampling frequency is experienced as too high.

While the results of the current dissertation suggest that a long questionnaire leads to lower within-person variability, a lower within-person relationship between time-varying variables, and a higher relative impact of RS, I did not investigate whether RS are the cause of the observed lower within-person variability and lower within-person relationship. Participants may rely more on RS (relative to the substantive trait) in order to reduce their cognitive load (or effort required to respond to the items) in long questionnaires, which could cause a lower within-person variability and a lower within-person relationship. For example, an individual's true score on a five-point Likert scale might be a four in one measurement occasion and a five on the next measurement occasion. If this individual relies more on extreme RS (relative to the substantive trait), he or she may choose to report a five on both measurement occasions as his or her respective score. Consequently, the within-person variability and the within-person relationships could be biased across many participants and many measurement occasions. However, with the chosen approach to model RS in empirical

Chapter 4, it is not possible to investigate the impact of RS on the within-person variability and the within-person relationships, so future research should investigate the association between RS and these outcome variables.

The observed lower within-person variability, and lower within-person relationship between time-varying variables in the long questionnaire group may also be caused by more careless responding. Because this dissertation focused on the effects of sampling frequency on careless responding and did not analyze the effects of questionnaire length on careless responding, I do not know whether questionnaire length influences careless responding. However, in line with this idea, Eisele et al. (2020) found that a higher questionnaire length leads to more careless responding. In the example described above, participants that engage in more careless responding, in order to reduce their cognitive load (or effort required to respond to the items) in long questionnaires, would report a five on both measurement occasions, irrespective of item content and their true scores.[11] Additionally, it is possible that participants respond carelessly on some occasions and respond with a higher reliance on their RS (relative to their substantive trait) on other occasions. This assumption is in line with the study by Alarcon and Lee (2022), who found that careless responding and RS are moderately correlated but refer to different underlying reasons why participants display either careless responding or RS. They state that RS occur when participants respond to the item superficially, whereas careless responding occurs when participants respond to items without reading the item (Alarcon & Lee, 2022). More research is needed on the effects of questionnaire length on careless responding, the associations between careless responding and RS, and on their influence on aspects of the data quality (e.g., within-person variability or within-person relationship between time-varying variables) in AA data.

[11] In this example careless responding refers to a participants' response to the items of the questionnaire without regard to item content (see e.g., Meade & Craig, 2012).

Another limitation of the current dissertation is that the choices during data management (preprocessing choices) may have biased the results. We only reported a single set of preprocessing choices and their results, which could undermine the validity of the conclusions (Weermeijer et al., 2022). For example, we screened for careless responding in empirical Chapter 2 and excluded the data of some participants and some measurement occasions from all reported analyses. It is possible that different preprocessing choices would have yielded different results, as shown, for example, in the (cross-sectional) study by Steegen et al. (2016). However, Weermeijer et al. (2022), who analyzed the effects of preprocessing choices related to data exclusion on the conclusions drawn from the analyses in five different AA studies, did not find an effect of the preprocessing choices on the results of the analyses. Regarding the results of the current dissertation, I do not know whether the chosen preprocessing choices made biased the results. Clearly, future researchers should clarify the effects of preprocessing choices in AA studies on the results obtained.

In the current dissertation, I did not investigate whether most of the outcome variables, except for participant burden, change over time in an AA study. Therefore, I do not know whether participants provide the same data quantity or data quality in the first few days of an AA study compared to last few days of an AA study. Participants may be more fatigued or bored with the study protocol near the end of an AA study, thus refraining from completing the measurement occasions or providing more data that is not of sufficient quality. Furthermore, the study duration, as a design feature in AA studies, might influence the effects of sampling frequency and questionnaire length. For example, in an AA study with a high sampling frequency and a long study duration participants have to respond to the same items many times, which may lead to boredom or fatigue, potentially affecting data quantity or (aspects of) data quality. Similarly, a longer questionnaire combined with a long study duration requires participants to complete many different items repeatedly, which can lead to

boredom or fatigue. Future research should investigate whether the study duration affects the strength of the effects of the sampling frequency and questionnaire length and whether the effects of sampling frequency and questionnaire length vary over the duration of an AA study.

Another limitation, which has been discussed in empirical Chapters 2 and 4, is that our student sample (participants who were young and highly educated, with a large proportion of women) may restrict the generalizability of the findings of this dissertation. For example, the effects of questionnaire length or sampling frequency may be different in a population of employees, because employees may not be able, or allowed, to complete the measurement occasions of a high sampling frequency AA study, or they may be unwilling to exert the effort required to respond to all the items in a long questionnaire due to of their work-related stressors. Moreover, selection bias may limit the generalizability of the findings to the population. While there may be a selection bias in AA in general (Stone, Schneider, Smyth, Junghaenel, Wen, et al., 2023), it is unlikely that our chosen sampling frequencies, which participants knew about prior to enrolling in our AA studies, resulted in a different selection bias than other AA studies (Stone, Schneider, Smyth, Junghaenel, Couper, et al., 2023). Therefore, I believe that the results of this dissertation should be generalized to other student populations that are used in many other AA studies.

In this dissertation, I have manipulated only two of many central design features in AA studies. However, many other design features, such as the study duration, the item order, the complexity of the items, the cognitive load involved in answering each item, may affect participant burden, or aspects of AA data (e.g., data quantity and data quality). Additionally, there are more outcomes that can be considered. For example, a longer questionnaire may lead to a lower reliability of the measurement. Future research should investigate the effects of other design features on the outcome variables identified in this dissertation, and on other outcome variables, in AA studies.

5.4 Conclusion

By shedding light on the effects of the sampling frequency and questionnaire length as central design choices in AA studies, this dissertation contributes to a better understanding of the extent to which, and under what conditions, these two design choices might affect the identified outcome variables. The central findings regarding sampling frequency were that a higher sampling frequency led to a higher perceived participant burden, but did not affect other aspects of data quantity and quality investigated in the AA study presented. With regard to questionnaire length, I found that a longer questionnaire did not affect perceived participant burden or data quantity, but did lead to a lower within-person variability, a lower within-person relationship between time-varying variables and a greater relative impact of RS in the AA study presented. Although further validation of the results is essential, I hope that future researchers will integrate the results of this dissertation when designing an AA study.

6 References

The references for Papers 1, 2, and 3 are listed in the respective papers.

References

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

Alarcon, G. M., & Lee, M. A. (2022). The Relationship of Insufficient Effort Responding and Response Styles: An Online Experiment. *Frontiers in Psychology*, *12*, 784375. https://doi.org/10.3389/fpsyg.2021.784375

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*. https://doi.org/10.1037/met0000294

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

Beal, D. J. (2015). ESM 2.0: State of the Art and Future Potential of Experience Sampling Methods in Organizational Research. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 383–407. https://doi.org/10.1146/annurev-orgpsych-032414-111335

Bergkvist, L., & Rossiter, J. R. (2009). Tailor-made single-item measures of doubly concrete constructs. *International Journal of Advertising*, *28*(4), 607–621. https://doi.org/10.2501/S0265048709200783

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

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

Conner, T. S., & Reid, K. A. (2012). Effects of Intensive Mobile Happiness Reporting in Daily Life. *Social Psychological and Personality Science*, *3*(3), 315–323. https://doi.org/10.1177/1948550611419677

Courvoisier, D. S., Eid, M., & Lischetzke, T. (2012). Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality

characteristics. *Psychological Assessment*, *24*(3), 713–720. https://doi.org/10.1037/a0026733

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

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-Based Item Response Models of the GLMM Family. *Journal of Statistical Software*, *48*(Code Snippet 1). https://doi.org/10.18637/jss.v048.c01

Dworak, E. M. (2022). Balancing Missing Data, Items, and Assessment Frequency in Experience Sampling Methods Data. *Multivariate Behavioral Research*, *57*(1), 159–160. https://doi.org/10.1080/00273171.2021.2009328

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*. https://doi.org/10.1177/1073191120957102

Eisele, G., Vachon, H., Lafit, G., Tuyaerts, D., Houben, M., Kuppens, P., Myin-Germeys, I., & Viechtbauer, W. (2023). A mixed-method investigation into measurement reactivity to the experience sampling method: The role of sampling protocol and individual characteristics. *Psychological Assessment*, *35*(1), 68–81. https://doi.org/10.1037/pas0001177

Fahrenberg, J. (2006). *Assessment in daily life. A review of computer-assisted methodologies and applications in psychology and psychophysiology, years 2000—2005*.

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

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), 101384. https://doi.org/10.1016/j.leaqua.2020.101384

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. (2021). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01683-6

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (in press). Modeling careless responding in ambulatory assessment studies using multilevel latent class analysis: Factors influencing careless responding. *Psychological Methods*.

Hasselhorn, K., Ottenstein, C., Meiser, T., & Lischetzke, T. (2023). *The Effects of Questionnaire Length on the Relative Impact of Response Styles in Ambulatory Assessment*. Manuscript submitted for publication.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus, 3rd ed.* (pp. xix, 440). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315746494

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

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications* (1st ed.). Routledge Academic. https://doi.org/10.4324/9781410604118

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2021). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychological Methods, 27*(6), 958–981. https://doi.org/10.1037/met0000312

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction, 114*(4), 609–619. https://doi.org/10.1111/add.14503

Kam, C. C. S., & Meyer, J. P. (2015). How Careless Responding and Acquiescence Response Bias Can Influence Construct Dimensionality: The Case of Job Satisfaction. *Organizational Research Methods*, *18*(3), 512–541. https://doi.org/10.1177/1094428115571894

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

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042–1060. https://doi.org/10.1037/a0020962

Larson, R., & Csikszentmihalyi, M. (1983). The Experience Sampling Method. *New Directions for Methodology of Social & Behavioral Science*, *15*, 41–56.

Li, C. R., Follingstad, D. R., Campe, M. I., & Chahal, J. K. (2022). Identifying Invalid Responders in a Campus Climate Survey: Types, Impact on Data, and Best Indicators. *Journal of Interpersonal Violence*, 088626052091858. https://doi.org/10.1177/0886260520918588

Liu, H., Xie, Q. W., & Lou, V. W. Q. (2019). Everyday social interactions and intra-individual variability in affect: A systematic review and meta-analysis of ecological momentary assessment studies. *Motivation and Emotion*, *43*(2), 339–353. https://doi.org/10.1007/s11031-018-9735-x

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, *59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

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

McCarthy, D. E., Minami, H., Yeh, V. M., & Bold, K. W. (2015). An experimental investigation of reactivity to ecological momentary assessment frequency among adults trying to quit smoking: Reactivity to ecological momentary assessment. *Addiction*, *110*(10), 1549–1560. https://doi.org/10.1111/add.12996

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

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

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*(2), 271–298. https://doi.org/10.1080/1359432X.2010.550680

Morren, M., Dulmen, S., Ouwerkerk, J., & Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: A systematic review. *European Journal of Pain*, *13*(4), 354–365. https://doi.org/10.1016/j.ejpain.2008.05.010

Moskowitz, D. S., Russell, J. J., Sadikaj, G., & Sutton, R. (2009). Measuring people intensively. *Canadian Psychology/Psychologie Canadienne*, *50*(3), 131–140. https://doi.org/10.1037/a0016625

Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What Affects the Completion of Ecological Momentary Assessments in Chronic Pain Research? An Individual Patient Data Meta-Analysis. *Journal of Medical Internet Research*, *21*(2), e11398. https://doi.org/10.2196/11398

Ottenstein, C., & Werner, L. (2021). *Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors [Manuscript submitted for publication]*.

Ottenstein, C., & Werner, L. (2022). Compliance in Ambulatory Assessment Studies: Investigating Study and Sample Characteristics as Predictors. *Assessment*, *29*(8), 1765–1776. https://doi.org/10.1177/10731911211032718

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 *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). Elsevier. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Piasecki, T. M., Hufford, M. R., Solhan, M., & Trull, T. J. (2007). Assessing clients in their natural environments with electronic diaries: Rationale, benefits, limitations, and barriers. *Psychological Assessment*, *19*(1), 25–43. https://doi.org/10.1037/1040-3590.19.1.25

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. *Journal of Applied Psychology*, *104*(6), 727–754. https://doi.org/10.1037/apl0000374

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Sage Publications.

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. https://doi.org/10.1037/pas0000662

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2020). Momentary predictors of compliance in studies using the experience sampling method. *Psychiatry Research*, *286*, 112896. https://doi.org/10.1016/j.psychres.2020.112896

Roekel, E., Keijsers, L., & Chung, J. M. (2019). A Review of Current Ambulatory Assessment Studies in Adolescent Samples and Practical Recommendations. *Journal of Research on Adolescence*, *29*(3), 560–577. https://doi.org/10.1111/jora.12471

Rosen, C. C., Koopman, J., Gabriel, A. S., & Johnson, R. E. (2016). Who strikes back? A daily investigation of when and why incivility begets incivility. *Journal of Applied Psychology*, *101*(11), 1620–1634. https://doi.org/10.1037/apl0000140

Santangelo, P. S., Ebner-Priemer, U. W., & Trull, T. J. (2013). *Experience Sampling Methods in Clinical Psychology*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793549.013.0011

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In *Handbook of research methods for studying daily life*. (pp. 302–320). The Guilford Press.

Sitzmann, T., & Yeo, G. (2013). A Meta-Analytic Investigation of the Within-Person Self-Efficacy Domain: Is Self-Efficacy a Product of Past Performance or a Driver of Future

Performance?: PERSONNEL PSYCHOLOGY. *Personnel Psychology*, *66*(3), 531–568. https://doi.org/10.1111/peps.12035

Sokolovsky, A. W., Mermelstein, R. J., & Hedeker, D. (2014). Factors Predicting Compliance to Ecological Momentary Assessment Among Adolescent Smokers. *Nicotine & Tobacco Research*, *16*(3), 351–358. https://doi.org/10.1093/ntr/ntt154

Sonnentag, S., Binnewies, C., & Mojza, E. J. (2008). "Did you have a nice evening?" A day-level study on recovery experiences, sleep, and affect. *Journal of Applied Psychology*, *93*(3), 674–684. https://doi.org/10.1037/0021-9010.93.3.674

Soyster, P. D., Bosley, H. G., Reeves, J. W., Altman, A. D., & Fisher, A. J. (2019). Evidence for the Feasibility of Person-Specific Ecological Momentary Assessment Across Diverse Populations and Study Designs. *Journal for Person-Oriented Research*, *5*(2), 53–64. https://doi.org/10.17505/jpor.2019.06

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stone, A. A., Broderick, J. E., Schwartz, J. E., Shiffman, S., Litcher-Kelly, L., & Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: Reactivity, compliance, and patient satisfaction: *Pain*, *104*(1), 343–351. https://doi.org/10.1016/S0304-3959(03)00040-X

Stone, A. A., Schneider, S., Smyth, J. M., Junghaenel, D. U., Couper, M. P., Wen, C., Mendez, M., Velasco, S., & Goldstein, S. (2023). A population-based investigation of participation rate and self-selection bias in momentary data capture and survey studies. *Current Psychology*. https://doi.org/10.1007/s12144-023-04426-2

Stone, A. A., Schneider, S., Smyth, J. M., Junghaenel, D. U., Wen, C., Couper, M. P., & Goldstein, S. (2023). Shedding light on participant selection bias in Ecological

Momentary Assessment (EMA) studies: Findings from an internet panel study. *PLOS ONE*, *18*(3), e0282591. https://doi.org/10.1371/journal.pone.0282591

Trougakos, J. P., Beal, D. J., Green, S. G., & Weiss, H. M. (2008). Making the Break Count: An Episodic Examination of Recovery Activities, Emotional Experiences, and Positive Affective Displays. *Academy of Management Journal*, *51*(1), 131–146. https://doi.org/10.5465/amj.2008.30764063

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

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, *129*(1), 56–63. https://doi.org/10.1037/abn0000473

Vachon, H., Bourbousson, M., Deschamps, T., Doron, J., Bulteau, S., Sauvaget, A., & Thomas-Ollivier, V. (2016). Repeated self-evaluations may involve familiarization: An exploratory study related to Ecological Momentary Assessment designs in patients with major depressive disorder. *Psychiatry Research*, *245*, 99–104. https://doi.org/10.1016/j.psychres.2016.08.034

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental Disorders: Meta-Analysis and Recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

van Berkel, N., Ferreira, D., & Kostakos, V. (2018). The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys*, *50*(6), 1–40. https://doi.org/10.1145/3123988

van Genugten, C. R., Schuurmans, J., Lamers, F., Riese, H., Penninx, B. W. J. H., Schoevers, R. A., Riper, H. M., & Smit, J. H. (2020). Experienced Burden of and Adherence to Smartphone-Based Ecological Momentary Assessment in Persons with Affective Disorders. *Journal of Clinical Medicine*, *9*(2), 322. https://doi.org/10.3390/jcm9020322

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

Walsh, E., & Brinker, J. K. (2016). Temporal Considerations for Self-Report Research Using Short Message Service. *Journal of Media Psychology*, *28*(4), 200–206. https://doi.org/10.1027/1864-1105/a000161

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, *82*(2), 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Warren, C. R., & Landis, R. (2007). One is the loneliest number: A meta-analytic investigation on single-item measure fidelity. *Ergometrika*, *4*, 32–53.

Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., Vaessen, T., Kuppens, P., & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods*, *54*(6), 2981–2992. https://doi.org/10.3758/s13428-021-01777-1

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance With Mobile Ecological Momentary Assessment Protocols in Children and Adolescents: A Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *19*(4), e132. https://doi.org/10.2196/jmir.6641

Wrzus, C., & Neubauer, A. B. (2022). Ecological Momentary Assessment: A Meta-Analysis on Designs, Samples, and Compliance Across Research Fields. *Assessment*, 107319112110675. https://doi.org/10.1177/10731911211067538

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

Lebenslauf

Persönliche Angaben

Name:	Hasselhorn
Vorname:	Kilian
E-Mail-Adresse:	Kilian.Hasselhorn@rptu.de

Berufstätigkeit

Seit 10/2018	Wissenschaftlicher Mitarbeiter Arbeitseinheit Diagnostik, Differentielle und Persönlichkeitspsychologie, Methoden und Evaluation an der Universität Koblenz-Landau, Campus Landau (heute RPTU Kaiserslautern-Landau)

Akademischer Werdegang

10/2016 – 09/2018	Master of Science, Psychologiestudium mit Schwerpunkt Wirtschaftspsychologie an der Universität Koblenz-Landau, Campus Landau
10/2012 – 09/2016	Bachelor of Science, Psychologiestudium an der Universität Koblenz-Landau, Campus Landau

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die Dissertation selbst angefertigt habe und keine anderen Hilfsmittel verwendet habe als die in der Arbeit angegebenen.
Bei den gemeinsam verfassten Publikationen habe ich folgende individuelle Beiträge erbracht:

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2021). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01683-6

Formulierung der Fragestellung erfolgte in Zusammenarbeit mit T. Lischetzke, und C. Ottenstein. Die Konzeption der Studie 1 erfolgte in Zusammenarbeit mit T. Lischetzke, und C. Ottenstein. Die Konzeption der Studie 2 erfolgte in Zusammenarbeit mit T. Lischetzke. Die Durchführung der Studie 1 erfolgte (in Zusammenarbeit mit C. Ottenstein) durch mich. Die Auswertung der Studie 1 erfolgte (nach Beratung durch T. Lischetzke) durch mich. Die Durchführung der Studie 2 erfolgte durch mich. Die Auswertung der Studie 2 erfolgte (nach Beratung durch T. Lischetzke) durch mich. Ich habe das Manuskript selbstständig (nach Korrekturen von T. Lischetzke und C. Ottenstein) verfasst.

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

Formulierung der Fragestellung erfolgte in Zusammenarbeit mit T. Lischetzke, und C. Ottenstein. Die Konzeption der Studie erfolgte in Zusammenarbeit mit T. Lischetzke, und C. Ottenstein. Die Durchführung der Studie erfolgte (in Zusammenarbeit mit C. Ottenstein) durch mich. Die Auswertung der Studie erfolgte (nach Beratung durch T. Lischetzke) durch mich. Ich habe das Manuskript selbstständig (nach Korrekturen von T. Lischetzke und C. Ottenstein) verfasst.

Hasselhorn, K., Ottenstein, C., Meiser, T., & Lischetzke, T. (2023). The effects of questionnaire length on response styles in ambulatory assessment. [Manuscript submitted for publication].

Formulierung der Fragestellung erfolgte in Zusammenarbeit mit T. Lischetzke, und C. Ottenstein. Die Konzeption der Studie erfolgte in Zusammenarbeit mit T. Lischetzke und T. Meiser. Die Durchführung der Studie erfolgte durch mich. Die Auswertung der Studie erfolgte (nach Beratung durch T. Lischetzke und T. Meiser) durch mich. Ich habe das Manuskript selbstständig (nach Korrekturen von T. Lischetzke, T. Meiser, und C. Ottenstein) verfasst.

Diese Arbeit oder Teile davon wurden noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Weder diese, noch eine andere Arbeit wurden von mir bei einer anderen Hochschule als Dissertation eingereicht.

Landau, den 25.04.2023

(Kilian Hasselhorn, M.Sc.)