Investigating the effect of experience sampling study design on careless and insufficient effort responding identified with a screen-time-based mixture model

Esther Ulitzsch[1,2,3,*], Wolfgang Viechtbauer[4], Oliver Lüdtke[3,5], Inez Myin-Germeys[6], Gabriel Nagy[3], Steffen Nestler[7], and Gudrun Vera Eisele[6,*]

[*]E.U. and G.V.E. contributed equally to this work

[1]Centre for Educational Measurement, University of Oslo

[2]Centre for Research on Equality in Education, University of Oslo

[3]IPN—Leibniz Institute for Science and Mathematics Education

[4]Maastricht University

[5]Centre for International Student Assessment (ZIB)

[6]Department of Neurosciences, Center for Contextual Psychiatry, KU Leuven

[7]University of Münster

Author Note

Correspondence concerning this article should be sent to Esther Ulitzsch, University of Oslo, Centre for Educational Measurement, Blindernveien 31, Kristine Bonnevies Hus, 0371 Oslo, Norway, email: esther.ulitzsch@cemo.uio.no, or Gudrun Vera Eisele, KU Leuven, Center for Contextual Psychiatry, ON5/B Herestraat 49 - box 1029, 3000 Leuven, Belgium, email: gudrunvera.eisele@kuleuven.be. This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 33160. GVE was supported by an Odysseus grant (G0F8416N) to IMG and a senior project grant (G049023N) to IMG by Research Foundation Flanders. During this work, GVE received a junior postdoctoral fellowship from the Research Foundation Flanders (1223725N). Online materials for this article can be found in the OSF and are available via the following link: https://osf.io/emhgs/?view_only=867a911c11684c85871b5335131283be.

Abstract

When using the experience sampling method (ESM), researchers must navigate a delicate balance between obtaining fine-grained snapshots of phenomena of interest and avoiding undue respondent burden, which can lead to disengagement and compromise data quality. To guide that process, we investigated how questionnaire length and sampling frequency impact careless and insufficient effort responding (C/IER) as an important yet understudied aspect of ESM data quality. To this end, we made use of existing experimental ESM data (Eisele et al., 2022) from 163 students randomly assigned to one of two questionnaire lengths (30/60 items) and one of three sampling frequencies (3/6/9 assessments per day). In our post-registered analyses, we employed a novel mixture modeling approach (Ulitzsch, Nestler, et al., 2024) that leverages screen time data to disentangle attentive responding from C/IER and allows investigating how the occurrence of C/IER evolved within and across ESM study days. We further investigated the relationship between model-implied C/IER and other engagement measures, such as self-reported attentiveness, attention checks, and compliance. We found sampling frequency, but not questionnaire length to impact C/IER, with higher frequencies resulting in higher overall C/IER proportions and sharper increases of C/IER across, but not within days. These effects proved robust across various model specifications. Our findings contrast previous studies on non-compliance, suggesting that respondents may employ different strategies to lower the different types of burden imposed by questionnaire length and sampling frequency. Implications for designing ESM studies are discussed.

Keywords: ecological momentary assessment, sampling frequency, questionnaire length; careless responding; screen times

Countless psychological processes, including symptoms of psychological disorders, are defined by how people think, feel, or behave during their everyday life. Recently, measuring these day-to-day experiences and behaviors has been facilitated by a proliferation of ambulatory assessment techniques, such as the Experience Sampling Method (ESM; Larson & Csikszentmihalyi, 1983; Myin-Germeys & Kuppens, 2021; also referred to as Ecological Momentary Assessment; Stone & Shiffman, 1994). By presenting multiple questionnaires per day, ESM research offers detailed data at a high temporal resolution and reduces retrospective biases associated with cross-sectional assessments (Neubauer et al., 2020; Schwarz, 2012). Increasingly, scholars have recognized the potential of this method for investigating within-person processes (Hamaker & Wichers, 2017) and contextual influences on psychological phenomena (Mestdagh & Dejonckheere, 2021; Myin-Germeys et al., 2018). Yet, the high intensity of ESM assessments can be burdensome for participants, which can, in turn, undermine the quality and quantity of the collected data. Therefore, optimizing study design is crucial to ensure the collection of high-quality ESM data that can inform psychological research. In order to optimize ESM study design, careful investigation of the effects of different design choices on the data are warranted. While the impact of ESM study design on non-compliance and, hence, data quantity is well studied (Conner & Reid, 2012; Eisele et al., 2022; Hasselhorn et al., 2022; McCarthy et al., 2015; Ono et al., 2019; Ottenstein & Werner, 2022; Rintala et al., 2019; Stone et al., 2003; Vachon et al., 2019; Walsh & Brinker, 2016; Wrzus & Neubauer, 2023), its effect on data quality is less well understood. In the current study, we filled this gap and investigated how two major ESM study design choices, namely the questionnaire length (i.e., how many items per assessment?) and the sampling frequency (i.e., how many assessments per day?), relate to careless and insufficient effort responding (C/IER) as an important aspect of data quality. To this end, we drew on a novel mixture modeling approach (Ulitzsch, Nestler, et al., 2024) that leverages screen time data to disentangle attentive responding from C/IER and allows investigating how the occurrence of C/IER

evolved within and across ESM study days. We further investigated the relationship between model-implied C/IER and other engagement measures, such as self-reported attentiveness, attention checks, and compliance.

ESM Design and Respondent Engagement

ESM studies are characterized by a large number of design choices, which allow researchers to adjust the method to their particular research questions and target populations. In the current study, we focus on two decisions that are central in the set-up of every ESM study: The sampling frequency and the questionnaire length. In the literature, the choice of both sampling frequency and questionnaire length varies widely from study to study. For example, in the field of personality disorders, a recent review reported sampling frequencies ranging from 1 to 12 assessments per day and questionnaire lengths ranging from 1 to 100 items (Kaurin et al., 2023) and a review on the use of ESM for chronic pain research observed sampling frequencies from 3 to 12 assessments per day with each assessment including 6 to 63 items (Ono et al., 2019). In theory, increasing the amount of collected information with more detailed and more frequent assessments per day could increase our understanding of the studied phenomena. However, as ESM protocols become more intensive, respondent engagement may be negatively affected. Data quality and quantity may suffer, thereby undermining the added value of extra questions or assessment moments. Understanding how common variations in the study design influence the obtained data is of high importance in guiding the use of ESM for psychological assessment.

Study Design and Compliance

Most past studies on the link between ESM study design factors and the collected data have investigated design-related effects on non-compliance (i.e., instances when a participant does not respond to a scheduled assessment). With few exceptions (for questionnaire length: Morren et al., 2009; for sampling frequency: Vachon et al., 2019; Wrzus & Neubauer, 2023), increases in sampling frequencies and questionnaire length have

not been associated with increases in non-compliance in meta-analyses (no effect of length: Jones et al., 2019; Ono et al., 2019; Ottenstein & Werner, 2022; Rintala et al., 2019; Soyster et al., 2019; Vachon et al., 2019; no effect of frequency: Jones et al., 2019; Morren et al., 2009; Ono et al., 2019; Ottenstein & Werner, 2022). However, such meta-analytical evidence is limited by the variations of design choices in included studies. Design factors tend to be adjusted to each other in applied ESM research. For instance, in an effort to reduce the overall burden for respondents, higher sampling frequencies are typically combined with shorter overall study durations (Kaurin et al., 2023; Wrzus & Neubauer, 2023). This adjustment may equally apply to unreported design aspects of studies, such as the instructions given to participants or incentive strategies to increase engagement, making it challenging to attribute differences in compliance between studies to individual design choices. In addition, studies with outlying low compliance rates may not be published, which can affect meta-analytic findings. Experimental investigations that directly manipulate sampling frequency and questionnaire length, therefore, play an important role in extending our understanding of design-related changes in ESM data because they allow isolating the effect of a single design choice on the data. While two experimental investigations have detected higher reported respondent burden with higher sampling frequencies (Hasselhorn et al., 2022; Stone et al., 2003), no changes in compliance were found in any of the experimental studies (Conner & Reid, 2012; Eisele et al., 2022; Hasselhorn et al., 2022; McCarthy et al., 2015; Stone et al., 2003; Walsh & Brinker, 2016). Less is known about the influence of the questionnaire length on the collected ESM data. Of two recent investigations, only one detected that increases in questionnaire length from 30 to 60 items led to lower compliance (Eisele et al., 2022), while the only other investigations failed to find effects of 33 vs 82 items on compliance (Hasselhorn et al., 2022).

Study Design and Careless and Insufficient Effort Responding

Notably, previous research into ESM study design has focused almost exclusively on compliance. Yet, not responding to a scheduled assessment is only one way for respondents to manage a high assessment burden. Faced with more intensive ESM study designs, respondents may disengage from the study in more subtle ways, for example by engaging in C/IER. C/IER, which is the focus of the current study, has been defined as responding without paying sufficient attention to the content of questions and/or survey instructions (Huang et al., 2015; Meade & Craig, 2012). Only two previous studies have investigated the influence of study design on C/IER in ESM data. In Eisele et al. (2022), increases in questionnaire length from 30 to 60 items were associated with less self-reported attentive responding, while increases in sampling frequency did not have an effect. A second experimental study detected lower within-person variance and weaker within-person relationships with longer questionnaires but not higher sampling frequencies (Hasselhorn et al., 2022). These changes in the data were interpreted as signs of lower data quality.

The apparent lack of research on the effects of study design on C/IER may be associated with a lack of methods to operationalize C/IER. Investigating C/IER occurrence in ESM studies relies on valid and reliable detection approaches. So far, research investigating C/IER in ESM data has predominantly relied on self-reports and behavioral indicators derived from attention check items (as in Eisele et al., 2022) or response patterns (e.g., using the long-string index introduced in Johnson, 2005, to scan for suspiciously low response variability; as in Jaso et al., 2022) and collateral screen time information, flagging observations associated with screen times that are too short to properly evaluate the administered items (as in Hasselhorn et al., 2022, 2023; Jaso et al., 2022). While the application of these C/IER detection techniques allowed for initial investigations on how ESM design factors may affect C/IER occurrence, each of these approaches allows, in our view, only for a rather coarse description of C/IER occurrence and comes with shortcomings that limit the depth and possibly the validity of conclusions

drawn from their application. Using self-reports on C/IER behavior, respectively attentiveness, as in Eisele et al. (2022) rests on the somewhat peculiar assumption that respondents disengaging from the study and exhibiting C/IER on content items nevertheless provided attentive, valid responses when reporting on their behavior. Such items may therefore only catch certain forms of C/IER, where respondents do not fully disengage from the questionnaire (see Meade & Craig, 2012, for a discussion of self-report measures of careless responding). Attention check items are typically administered parsimoniously as repeated and extensive administration may confuse attentive respondents (Meade & Craig, 2012). This infrequent administration, however, may yield a highly underpowered C/IER measure (Eisele et al., 2022).

Finally, behavioral indicators exhibit a certain extent of ambiguity. For instance, when—as it is oftentimes the case in ESM studies—scales are short and repeatedly administered, distributions of behavioral indicators may become almost inseparable. On short scales, choosing the same response option on all items can stem from attentive and careless behavior with equal plausibility. Likewise, familiarization with the items due to repeated exposure may ease and speed up the attentive response process, resulting in strongly overlapping attentive and C/IER screen time distributions. Therefore, finding a criterion that is used to determine whether a response is too short (see Jaso et al., 2022) may be difficult.

Screen-Time-Based Mixture Model as a Novel Tool for Investigating C/IER Occurrence in ESM Studies

Recently developed confirmatory mixture modeling approaches for identifying and investigating C/IER in ESM studies (Ulitzsch, Nestler, et al., 2024; Vogelsmeier et al., 2024) overcome limitations of previously employed C/IER detection techniques (see Ulitzsch, Nestler, et al., 2024, for a detailed discussion), thereby opening new paths for gaining a fine-grained understanding of C/IER occurrence in ESM data. The overarching

principle of confirmatory mixture modeling approaches to C/IER is the translation of theoretical considerations on respondent behavior into two mixture component models—one representing an assumed attentive and the other one an assumed inattentive data-generating process. For instance, in approaches for item responses, attentive item responses are assumed to reflect the to-be-measured traits, i.e., to follow standard measurement models (Arias et al., 2020; Roman et al., 2023; Ulitzsch, Pohl, et al., 2022, 2023; Ulitzsch, Yildirim-Erbasli, et al., 2022; van Laar & Braeken, 2022). Inattentive item responses, in contrast, are assumed to be driven by respondents' category preferences (Arias et al., 2020) or to be random (van Laar & Braeken, 2022). So far, such approaches have predominantly been developed for cross-sectional data, and approaches targeted to the ESM context only emerged recently (Ulitzsch, Nestler, et al., 2024; Vogelsmeier et al., 2024). These allow for attentiveness to vary on the respondent-by-occasion level, facilitating to investigate both respondent-level and contextual determinants of C/IER behavior. Model formulations for both item responses (Vogelsmeier et al., 2024) and screen times from electronically administered ESM studies (Ulitzsch, Nestler, et al., 2024) exist. In the present study, we draw on the screen-time-based model by Ulitzsch, Nestler, et al. (2024). The model is rooted in the common presumption that differences in timing are indicative of differences in response strategies and behavior (De Boeck & Jeon, 2019), and that, hence, atypical screen times may signal aberrant respondent behavior. It targets identification of aberrantly short screen times, presumably too short to evaluate the displayed content and select suitable responses (Bowling et al., 2021; Jaso et al., 2022). By drawing on screen times rather than item responses, the model avoids measurement invariance assumptions for attentive item responses and can also be applied when constructs are measured with single indicators.

The screen-time-based mixture model by Ulitzsch, Nestler, et al. (2024) models the median screen time t_{ibd} respondent $i \in \{1, \ldots, N\}$ requires to respond to beep $b \in \{1, \ldots, B\}$ on day $d \in \{1, \ldots, D\}$. Note that beeps are nested within days. When the

same number of beeps is administered on consecutive days, the measurement occasion can be denoted as $o = (d-1) \cdot B + b$. Whether or not respondent i is attentive when responding to beep b administered on day d is encoded in the unobserved latent class variable s_{ibd}, taking the value 1 for attentive interactions and 0 otherwise.

Attentive log median screen times are assumed to be governed by a law-of-practice effect, capturing potential speed-up due to repeated exposure to the same questionnaire and the ESM delivery system, which may be particularly pronounced in the early stages of the study, when respondents familiarize themselves with the ESM measures and delivery system. More specifically, attentive log median screen times are modeled as a function of (a) a respondent-specific initial time expenditure parameter τ_{1i}, determining the time an attentive respondent requires at the very first measurement occasion, (b) a respondent-specific habitual decay parameter capturing potential speed-up across measurement occasions with an exponential decay function τ_{2i} and (c) a lower asymptote μ_β, incorporating the assumption that such speed-up effects will not result in median screen times of zero:

$$\ln(t_{ibd}|s_{ibd} = 1) \sim \mathcal{N}(\mu_\beta + \tau_{1i} \cdot \tau_{2i}^{o-1}, \sigma_1^2). \tag{1}$$

Inattentive log median screen times are assumed to be independent of respondent and occasion characteristics and are assumed to follow a common normal distribution, that is

$$\ln(t_{ibd}|s_{ibd} = 0) \sim \mathcal{N}(\mu_0, \sigma_0^2). \tag{2}$$

The mean of this normal distribution is constrained to be below the lower asymptote of the attentive component model, i.e., $\mu_0 \leq \mu_\beta$, incorporating the assumption that, on average, inattentive responding requires less time than attentive responding.

The probability of being in an attentive state is modeled using an item response theory (IRT) model as a function of a person-specific attentiveness parameter ψ_i and attentiveness difficulty. Attentiveness difficulty is allowed to vary linearly within and

between days, being modeled as a function of an initial attentiveness difficulty parameter γ_0 as well as the within- and between-day effect parameters γ_W and γ_B, that is

$$p(s_{ibd} = 1) = \pi_{ibd} = \frac{\exp(\psi_i - [\gamma_0 + \gamma_W \cdot (b-1) + \gamma_B \cdot (d-1)])}{1 + \exp(\psi_i - [\gamma_0 + \gamma_W \cdot (b-1) + \gamma_B \cdot (d-1)])}. \quad (3)$$

Conceptually, attentiveness difficulty captures the study's C/IER-evoking characteristics, while person attentiveness can be understood as a parameter that quantifies how much a respondent can "resist" these characteristics. Relating these parameters to person and contextual characteristics allows investigating how these are related to C/IER occurrence, rendering this approach a promising tool for gaining a fine-grained understanding of the determinants of C/IER behavior.

The Present Study

In the present study, we leveraged the novel approach by Ulitzsch, Nestler, et al. (2024) to investigate how ESM design choices affect C/IER occurrence as an important yet understudies way of C/IER aspect of ESM data quality. Specifically, we re-analyzed existing experimental data (Eisele et al., 2022) to investigate whether increases in sampling frequency (3, 6, or 9 assessments per day) and questionnaire length (30 vs 60 items per assessment) lead to changes in the data that are indicative of C/IER. This allows us to derive guidelines for designing ESM studies that curb the occurrence of C/IER and yield data of higher quality. Our two main post-registered hypotheses were that the model-based approach identifies lower proportions of attentive observations in the long compared to the short questionnaire condition (H1), as well as in the higher compared to lower sampling frequency conditions (H2). Note that the terms higher C/IER and lower attentiveness are used interchangeably throughout the registration and text. We also explored differences in patterns of linear within- and between-day changes in attentiveness difficulties as a function of different questionnaire lengths (E1.1) and sampling frequencies (E2.1).

We scrutinized conclusions on our primary hypotheses and explorations against

choices in model set-up. In particular, we explored whether conclusions on C/IER (respectively, attentiveness) differences and trajectories across questionnaire length (E1.2) and sampling frequency conditions (E2.2) are robust against (a) the choice of screen time measure in a branched ESM design, (b) whether habitual decay is modeled as a function of completed surveys or time passed since the start of the ESM study, and (c) whether within-day changes in attentiveness difficulty are modeled as a function of the number of administered beeps or time of the day. We further evaluated whether there is evidence for more complex (i.e., quadratic) within- and/or between-day changes in attentiveness difficulty in any of the questionnaire length conditions (E1.3) and any of the sampling frequency conditions (E2.3).

Besides our main goal of investigating the link between study design and C/IER and its robustness across model specifications, we aimed to investigate how the model-based C/IER detection approach relates to previously employed measures of C/IER and compliance. We hypothesized that levels of model-implied respondent attentiveness are lower for respondents who failed attention check items than for those who did not (H3.1) and that observations with failed attention check items show lower posterior attentive class probabilities according to the model-based approach (H3.2). In addition, we expected a positive association between self-reported attentive responding and model-implied respondent attentiveness (H4). Finally, we hypothesized that there would be a positive association between compliance (i.e., completed measurement occasions) and model-implied respondent attentiveness (H5). Two additional exploratory research questions were added after the registration, namely whether patterns of C/IER (respectively, attentiveness) trajectories obtained from self-reports mirror their model-implied counterparts (E4) and whether compliance trajectories mirror model-implied attentiveness trajectories (E5).

Methods

Transparency and Openness

The analyses and hypotheses were post-registered (https://osf.io/8nkpa/?view_only=771ce46a0eb142edac7488521dd17c4e) and deviations from the post-registration are described in the text. All study materials are available on the OSF page of the project (https://osf.io/emhgs/?view_only=867a911c11684c85871b5335131283be) and data are available upon request. We report all data manipulations and exclusions.

Sample

To be eligible for participation in the study, participants had to be fluent in Dutch, between 18 and 30 years old, and currently enrolled as a student at a University. In addition, participants were excluded if they had previously taken part in an ESM study. Recruitment took place between February and July 2019. Of 163 participants initially enrolled in the study, 3 had to be excluded during the baseline session for not meeting all inclusion criteria. In addition, for 2 participants the time on the study phone was off by more than 1 hour at follow up and for one participant the experimental condition was wrongly recorded[1] and these participants were therefore excluded from the current analyses. Further, 2 participants experienced technical problems that led to missing more than one assessment in a row on one day and data during these technical issues was coded as missing instead of non-compliant in the analyses. In addition, 4 participants missed more than a full day of the ESM period due to technical issues and all data after the start of the technical problem was marked as missing in the analyses. Finally, 3 participants received only 13 days of ESM beeps due to scheduling issues and experimenter error. The study was powered for its main hypothesis covered in Eisele et al. (2022). We address the

[1] This exclusion represents a deviation from the post-registration, as we were not aware of this issue at the time of registering the analyses.

topic of statistical power in relation to the current analyses in the discussion section.

Procedure

The study protocol was approved by the Social and Societal Ethics Committee of KU Leuven (G-2018 07 1285). Interested participants were invited to the lab for a baseline session. After signing informed consent, participants filled in a number of baseline questionnaires (see supplementary materials for a full list), received the study phone (Motorola DEFY+) with mobileQ (Meers et al., 2020) installed on it, and were briefed about the ESM period (see supplementary materials for a detailed description of the briefing session). On the day after the baseline session, the 14-day ESM period started. During this time, participants received either 30 or 60 item-long questionnaires 3, 6, or 9 times per day, depending on the experimental condition to which they had been randomly assigned at baseline. Each day, the assessments were spread semi-randomly in equal time windows between 9 am and 10:30 pm. At each assessment moment, participants had 90s to start the questionnaire and 90s until time-out per question. After the 14 days, participants returned to the lab for a follow-up session. During this session, they returned the study phone, filled in a number of additional questionnaires, and were debriefed about the study (see supplementary materials). A randomly chosen subsample of 50 participants additionally took part in a 10min-long qualitative interview about their experience in the study. At the end of the follow-up session, participants were rewarded with 40, 60, or 80€ gift vouchers depending on the assigned sampling frequency. Participants were informed during the briefing that the reward could be lowered based on their compliance but not about the exact threshold (compliance >33% for full reward).

Measures

A full overview of all questionnaires in the study can be found in the supplementary materials.

Screen times. The log median screen time per item was calculated for every assessment moment based on the 30 items that were common to both the short and long questionnaire versions. Since the ESM questionnaire contained a branching logic (different items were presented depending on whether the participant indicated being alone or in company and based on whether they indicated having missed a beep since the last assessment), not all 30 items could be included in the screen time calculations. Instead, we compared three different operationalizations of the median screen time to investigate the robustness of our conclusions against these variations. First, we used only the screen times of items up to the first branching in the questionnaire, i.e. all 21 items up to and including the item "Are you alone?". The results using this operationalization are described in detail in the main text. Second, we used only screen times of the 27 items that were not branched and thus shown at every beep irrespective of the reported social company and missed beeps. Third, we used screen times of all non-branched items and included branched items that had a similar format across the two branches (when in company: "I feel comfortable in company"; when alone: "I feel comfortable alone") resulting in a total of 28 items. The detailed results of these two additional operationalizations can be found in the appendix.

Self-reported attentiveness. Beep-level measures of self-reported attentiveness were obtained with the item "I filled in the questions attentively" (original Dutch item: "Ik heb de vragen aandachtig ingevuld."), which was the second-to-last non-branched item in the questionnaire. Participants could respond on a 7-point Likert scale with labeled endpoints ("Not at all" 1-2-3-4-5-6-7 "Very much"; original Dutch labels were Niet – Zeer). The item was developed for the purpose of this study and refined before data collection in focus groups with ESM researchers. Participants were informed during the briefing session that responses to this item would not influence their reward.

Attention check item. An attention check item ("Think back about what you were doing just before the beep. Please select <Not at all>."; orginial Dutch item: "Denk terug aan wat je aan het doen was net voor de beep. Gelieve voor deze vraag <Niet> te

kiezen.") was added to one questionnaire on day 3, 6, 9, and 12 of the study. Participants could respond on a 7-point Likert scale with labeled endpoints ("Not at all" 1-2-3-4-5-6-7 "Very much"; original Dutch labels were Niet – Zeer). Responses to the attention check item were coded as 0 when the correct answer was selected and 1 when a wrong answer was selected. The item was developed for the purpose of this study and refined before data collection in focus groups with ESM researchers. This item was not introduced during the briefing session.

Compliance. Compliance was defined as having responded to the last, non-branched item of the ESM questionnaire. Compliance was coded as 1 if a response was given and 0 if the item was not answered.

Analysis Strategy

To get a general understanding of attentive and C/IER behavior in the ESM study, we first applied the model by Ulitzsch, Nestler, et al. (2024) without covariates. Next, all hypotheses and explorations were addressed using model extensions.

Investigating the Effects of ESM Study Design Choices on C/IER

To address our hypotheses and explorations concerning the effects of ESM study design choices, we expanded the model by condition-specific logistic intercepts $\gamma_{0,g}$ as well as within- and between day effect parameters $\gamma_{W,g}$ and $\gamma_{B,g}$, with $g \in \{SQ, LQ\}$ when investigating the effects of questionnaire length and $g \in \{3B, 6B, 9B\}$ when investigating the effects of sampling frequency. All remaining parameters were assumed to be invariant across conditions. This corresponds to the assumption that ESM study design choices may affect C/IER occurrence, but not parameters related to attentive and C/IER behavior, e.g., the time respondents—with sufficient practice and habituation—minimally require to provide attentive responses as determined by the lower asymptote μ_β or the mean C/IER log median screen time μ_0. For the sake of simplicity, we studied the effects of questionnaire length and sampling frequency with separate models.

To test H1 and H2, we obtained condition-specific average attentiveness probabilities $\pi_{\dots,g}$ (with dots denoting the average across respondents, beeps, and days) as well as their pairwise differences as derived parameters. Likewise, to address E1.1 and E2.1, we obtained pairwise differences in condition-specific within- and between day effect parameters $\gamma_{W,g}$ and $\gamma_{B,g}$ as derived parameters.

In our robustness analyses (E1.2 and E2.2), we fully crossed the screen time measure employed for model implementation with different specifications of both habitual decay and linear condition-specific within-day changes in attentiveness difficulty, modeling these either as a function of completed ESM surveys, respectively, administered beeps or as a function of the time since the start of the study, respectively, the time of the day.

To explore potential quadratic within- and/or between-day changes in attentiveness difficulties (E1.3 and E2.3), we again focused on screen times of items up to the first branching and modeled habitual decay as a function of completed ESM surveys. We fully crossed the inclusion of condition-specific quadratic terms for within- and between-day changes in attentiveness difficulties, varying whether within-day changes were modeled as a function of administered beeps or time of the day.

Investigating the Relationship Between Model-Implied C/IER and Other Engagement Measures

In investigating the relationship between model-implied C/IER and other engagement measures, we did not include the experimental conditions as covariates in the model. Therefore, the obtained results should be interpreted as marginal effects across experimental conditions.

To address H3.1, we obtained average attentiveness levels for participants who failed attention check items and for those who did not as well as the difference in these averages as derived parameters. In analogy, for addressing H3.2, we obtained average attentiveness probabilities for observations with failed and passed attention checks as well as the

difference in these averages as derived parameters.

To investigate agreement between self-reported and model-implied respondent attentiveness (H4), we expanded the model and used self-reported momentary attentiveness r_{ibd} as a metric indicator for an additional latent attentiveness variable ψ_{SR}. In doing so, we mirrored the component model for attentiveness probabilities given in Equation 3, that is

$$r_{ibd} \sim \mathcal{N}(\gamma_{0,SR} + \gamma_{W,SR} \cdot (b-1) + \gamma_{B,SR} \cdot (d-1) + \psi_{i,SR}, \sigma_s^2). \tag{4}$$

We took a positive correlation between latent attentiveness measured with self-reports ψ_{SR} and latent attentiveness based on attentiveness probabilities ψ as supporting evidence for H4. Further, we compared the direction of the within- and between-day effect parameters γ_W and γ_B with their self-report counterparts $\gamma_{W,SR}$ and $\gamma_{B,SR}$ to explore similarities in the overall pattern of attentiveness trajectories across the course of the study.

In analogy, to investigate agreement between non-compliance and model-implied respondent attentiveness (H5), we employed the binary compliance indicator c_{ibd} for an additional latent compliance variable that mirrored the attentiveness measurement model, that is

$$p(c_{ibd} = 1) = \frac{\exp(\psi_{i,C} - [\gamma_{0,C} + \gamma_{W,C} \cdot (b-1) + \gamma_{B,C} \cdot (d-1)])}{1 + \exp(\psi_{i,C} - [\gamma_{0,C} + \gamma_{W,C} \cdot (b-1) + \gamma_{B,C} \cdot (d-1)])}. \tag{5}$$

We took a positive correlation between latent compliance ψ_C and attentiveness ψ as supporting evidence for H5. Further, we explored similarities in C/IER and non-compliance trajectories by comparing the direction of the within- and between-day effect parameters γ_W and γ_B with their compliance counterparts $\gamma_{W,C}$ and $\gamma_{B,C}$. Although the hypotheses were registered, the exact analytical strategy to investigate the relationship between model-implied C/IER and other engagement measures was developed after accessing the data.

For Bayesian estimation of all models, we used Stan version 2.19 (Carpenter et al., 2017) employing the rstan package version 2.19.3 (Guo et al., 2018), adhering to the prior

settings suggested in Ulitzsch, Nestler, et al. (2024). We ran two Markov chain Monte Carlo (MCMC) chains with 6,000 iterations each, with the first half being employed as warm-up. The sampling procedure was assessed on the basis of potential scale reduction factor (PSRF) values, with PSRF values below 1.10 for all parameters being considered as satisfactory (Gelman & Rubin, 1992; Gelman & Shirley, 2011). We employed the posterior mean (EAP) as a Bayesian point estimate and obtained 95% credibility intervals for statistical inference.

Exemplary Stan code for our analyses addressing our main hypotheses and explorations (H1, H2, E1.1, E2.1) as well as code illustrating model expansion by additional measurement models (H4) is provided in the OSF repository accompanying this study. Data and data processing scripts are available upon reasonable request.

Results

Figure 1 displays smoothed trajectories of observed median screen times in milliseconds based on all items up to the first branching. Overall, these trajectories tend to follow patterns that align with the assumed law-of-practice effect: after initially longer median screen times, many respondents accelerated, especially during the first measurement occasions.

General Pattern of Model-Implied Attentiveness

Across all experimental conditions, respondents, and measurement occasions, the average attentiveness probability was .95 [.93; .95], corresponding to an expected overall C/IER rate of 5%. Mean C/IER log median screen time was rather short ($\mu_0 = 7.40$ [7.37; 7.42], corresponding to $\exp(7.40) = 1{,}635$ milliseconds. We obtained a lower asymptote close to the mean C/IER log screen time ($\mu_\beta = 7.40$ [7.37; 7.43]), indicating that the mean log median screen time associated with careless responses was hardly distinguishable from the lower asymptote of attentive log median screen times.

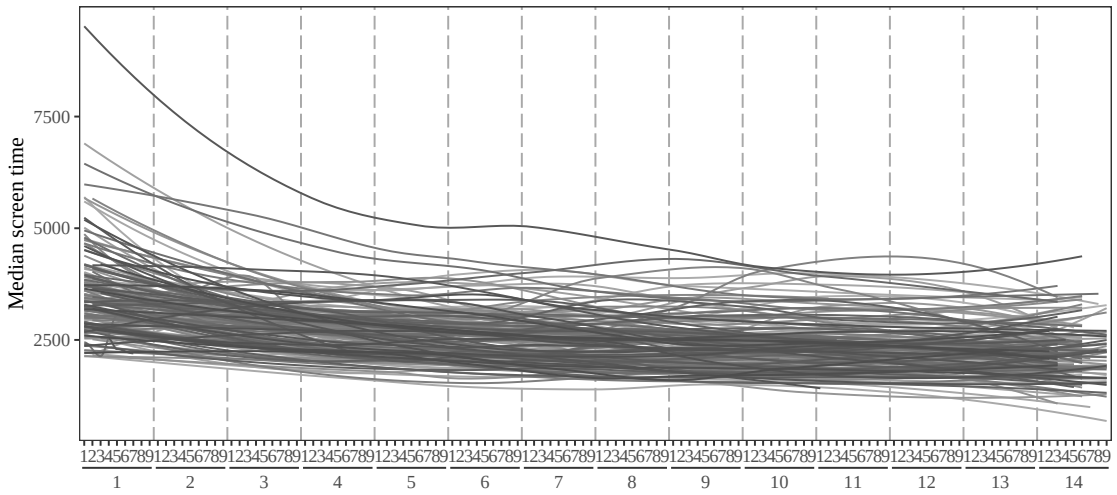


Figure 1. Smoothed trajectories of observed median screen times in milliseconds based on all items up up to the first branching across beeps and days. Note that trajectories for all participants are displayed and that the number of administered beeps differs across participants.

Figure 2 displays observed log median screen times across beeps and days color-coded by attentiveness probability π_{ibd}. The gray dashed line marks the mean of the C/IER log median screen time distribution μ_0. As can be seen, over time, a larger proportion of (predominantly overly short) median screen times was associated with lower attentiveness probabilities.

Table 1 displays means, standard deviations, and correlations of person attentiveness, log initial time expenditure, and log habitual decay.[2] From the mean log initial time expenditure and log habitual decay parameters, it can be concluded that, at the beginning of the study, respondents typically required markedly longer to provide attentive responses ($\exp(\mu_{\ln(\tau_1)}) = 0.65$), but sped-up across measurement occasions ($\exp(\mu_{\ln(\tau_2)}) = 0.97$). At the same time, both initial time expenditure and habitual decay varied across respondents. EAPs of individual habitual decay parameters ranged from 0.94 to 0.99, indicating that for none of the respondents attentive screen times increased across

[2] Note that a multivariate normal distribution is assumed for attentiveness, log initial time expenditure and log habitual decay, with the mean of the latent attentiveness factor set to zero for model identification.

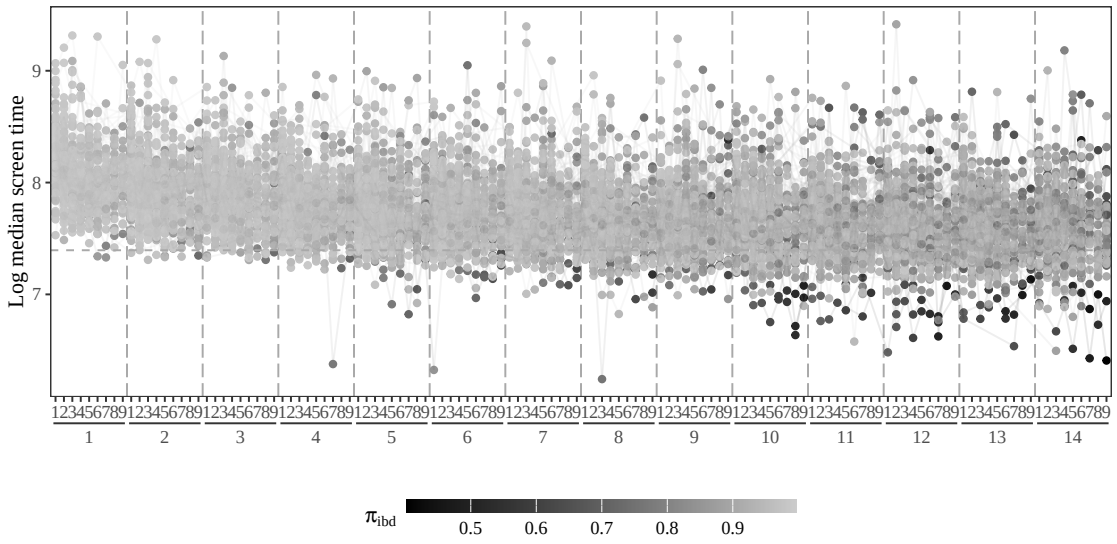


Figure 2. Log median screen times based on all items up to the first branching across beeps and days color-coded by model-implied attentiveness probability π_{ibd}. The horizontal gray dashed line marks the mean of the careless log median screen time distribution μ_0.

the study.

Table 1
Means, standard deviations, and correlations of person parameters obtained from the baseline model without covariates

	ψ	$\ln(\tau_1)$	$\ln(\tau_2)$	$\mu_{\mathcal{P}}$
ψ	1.16			0
	[0.92; 1.42]			
$\ln(\tau_1)$	.09	0.29		-0.43
	[-.11; .31]	[0.25; 0.33]		[-0.49; -0.37]
$\ln(\tau_2)$	.11	.31	0.02	-0.03
	[-.09; .31]	[.14; .46]	[0.01; 0.02]	[-0.03; -0.02]

Notes: 95% Bayesian credibility intervals are given in squared brackets. ψ: attentiveness; τ_1: initial time expenditure; τ_2: habitual decay; $\mu_{\mathcal{P}} = \left(\mu_{\psi}, \mu_{\ln(\tau_1)}, \mu_{\ln(\tau_2)}\right)$.

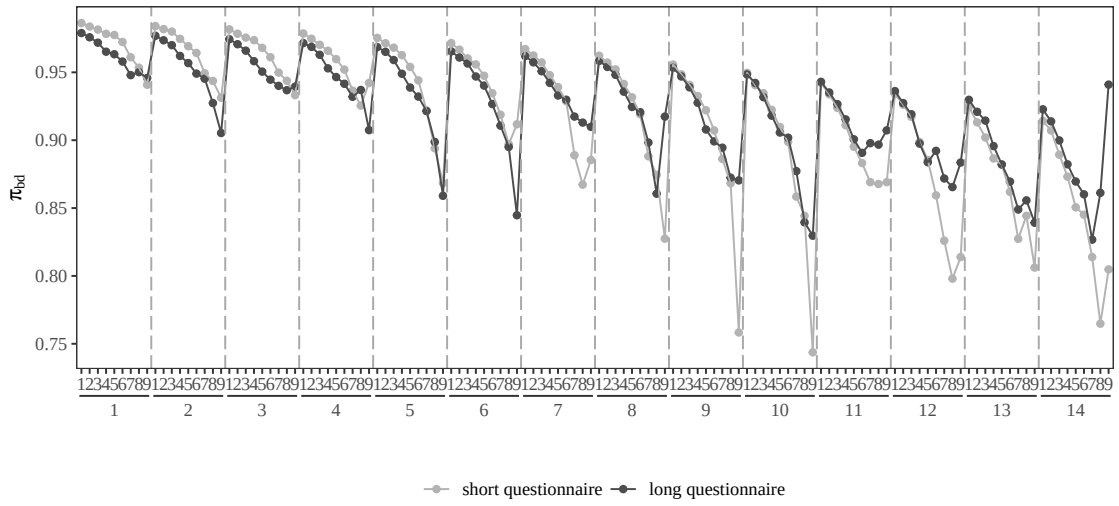
Investigating the Effects of ESM Study Design Choices on C/IER

Investigating the effects of questionnaire length. Contrary to our hypothesis (H1), we obtained virtually the same average attentiveness probabilities for short and long questionnaire conditions ($\pi_{\ldots,SQ} = \pi_{\ldots,LQ} = .94$ [.93; .95]; $\Delta_{\pi\ldots,LQ-SQ} = .00$ [−.01; .01]).

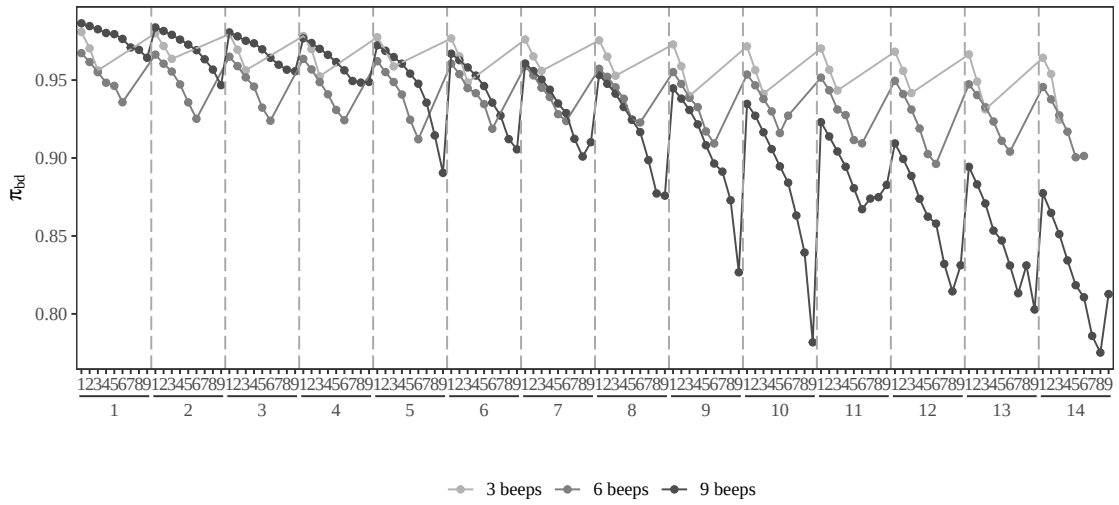
Figure 3a displays model-implied average attentiveness probabilities by questionnaire length condition across the course of the study. As can be seen, both questionnaire length conditions exhibited comparable patterns of declines in attentiveness probabilities within and between ESM study days (E1.1).[3] For both questionnaire conditions, the linear within- and between-day effects on attentiveness difficulty were credibly different from zero ($\gamma_{W,SQ} = 0.17\ [0.09; 0.25]$; $\gamma_{W,LQ} = 0.14\ [0.06; 0.23]$; $\gamma_{B,SQ} = 0.16\ [0.11; 0.21]$; $\gamma_{B,LQ} = 0.11\ [0.06; 0.16]$), but not from each other ($\Delta_{\gamma_W, LQ-SQ} = -0.03\ [-0.16; 0.07]$; $\Delta_{\gamma_B, LQ-SQ} = -0.05\ [-0.11; 0.01]$). This indicates that in both conditions attentiveness difficulties increased within and across days to a comparable extent.

In our robustness checks (E1.2; full results given in the appendix), when habitual decay was modeled as a function of the measurement occasion, all models yielded comparable condition-specific average attentiveness probabilities (range: [.94; .95]) as well as linear within- and between-day effects on attentiveness difficulty that were credibly different from zero but not credibly different across questionnaire length conditions. Note that within-day effects cannot be compared across models where these are modeled as a function of the measurement occasion and the time of the day, because the obtained parameters are on different scales. However, when habitual decay was modeled as a function of the time passed since the start of the study, we encountered severe convergence issues, as indicated by large PSRF values and strong trends in the traceplots. Here, only one out of six specified models converged, using only screen times up to the first branching and modeling within-day changes in attentiveness difficulty as a function of the number of beeps. Nevertheless, the converged model yielded condition-specific average attentiveness

[3] Note that the descriptive differences in average attentiveness probabilities for the last 3 beeps of the day occurring after the first study week are based on attentiveness probabilities of respondents in the 9-beep conditions only. Due to the small sample sizes for these measurement occasions (roughly 30 respondents in each questionnaire length condition), estimates of average attentiveness probabilities for these last 3 beeps of the day are rather unstable, and descriptive comparisons between the questionnaire length conditions are not trustworthy.



(a) Questionnaire length



(b) Sampling frequency

Figure 3. Average attentiveness probabilities across beeps and days by experimental condition.

probabilities comparable to the other model specifications and supported the conclusion that in both conditions, attentiveness difficulties increased within and across days to a comparable extent.

There was no evidence for quadratic within- and/or between-day effects on attentiveness difficulties in any of the conditions across all considered model specifications (E1.3). We encountered convergence issues when quadratic effects were included for both

within- and between-day effects in the same model.

Investigating the effects of sampling frequency. In line with our hypothesis (H2), we found the average attentiveness probability to be lower in conditions with higher sampling frequencies ($\pi_{...,3B} = .96$ [.95; .98]; $\pi_{...,6B} = .94$ [.93; .95]; $\pi_{...,9B} = .92$ [.91; .94]). All pairwise differences in average attentiveness probabilities were credibly different from zero ($\Delta_{\pi...,6B-3B} = -.02$ [−.04; −.01]; $\Delta_{\pi...,9B-3B} = -.04$ [−.06; −.02]; $\Delta_{\pi...,9B-6B} = -.02$ [−.03; −.00]).

Figure 3b displays model-implied average attentiveness probabilities by sampling frequency condition across the course of the study. As can be seen, the sampling frequency conditions exhibited different patterns of declines in attentiveness probabilities between ESM study days (E2.1). All frequency conditions exhibited comparable attentiveness levels at the beginning of the ESM study ($\gamma_{0,3B} = -4.75$ [−5.80; −3.57]; $\gamma_{0,6B} = -3.92$ [−4.61; −3.29]; $\gamma_{0,9B} = -4.95$ [−5.63; −4.28]; $\Delta_{\gamma_0,6B-3B} = 0.83$ [−0.35; 2.04]; $\Delta_{\gamma_0,9B-3B} = -0.21$ [−1.43; 1.08]; $\Delta_{\gamma_0,9B-6B} = -1.04$ [−1.76; −0.07]). Further, there were no credible differences in within-day effects on attentiveness difficulty across conditions ($\gamma_{W,3B} = 0.49$ [−0.06; 1.02]; $\gamma_{W,6B} = 0.18$ [0.06; 0.31]; $\gamma_{W,9B} = 0.13$ [0.07; 0.20]; $\Delta_{\gamma_W,6B-3B} = -0.30$ [−0.87; 0.24]; $\Delta_{\gamma_W,9B-3B} = -0.35$ [−0.90; 0.18]; $\Delta_{\gamma_W,9B-6B} = -0.05$ [−0.21; 0.09]).[4] There, were, however, pronounced differences in between-day effects on attentiveness difficulty across conditions. In the 3-beep condition, there was no evidence for linear between-day changes in attentiveness difficulty ($\gamma_{B,3B} = 0.05$ [−0.04; 0.13]), such that, across days, the average attentiveness probability remained relatively stable around .96, corresponding to an expected C/IER rate of 4%. For the 6-beep condition, there was a

[4] Note that for the 3-beep condition, the within-day effect parameter $\gamma_{W,3B}$ was hard to estimate when within-day changes in attentiveness difficulty were inferred from three measurement occasions per day, resulting in rather broad credibility intervals. When within-day changes were modeled as a function of the time of the day, because respondents were beeped at random times of the day, estimation was more stable, credibility intervals were, accordingly, more narrow, and there was evidence for substantial within-day declines in attentiveness difficulty across all conditions ($\gamma_{W,3B} = 0.12$ [0.04; 0.21]; $\gamma_{W,6B} = 0.09$ [0.04; 0.15]; $\gamma_{W,9B} = 0.08$ [0.04; 0.12]). Nevertheless, the conclusion that there were no credible differences in within-day effects on attentiveness difficulty across conditions remained ($\Delta_{\gamma_W,6B-3B} = -0.04$ [−0.14; 0.07]; $\Delta_{\gamma_W,9B-3B} = -0.05$ [−0.15; 0.06]; $\Delta_{\gamma_W,9B-6B} = -0.01$ [−0.07; 0.05]).

mild increase in attentiveness difficulty between ESM study days ($\gamma_{B,6B} = 0.04$ $[0.00; 0.10]$), translating into an average attentiveness probability of .95 (corresponding to an expected C/IER rate of 5%) on the first day of the ESM study to .93 (corresponding to an expected C/IER rate of 7%) on day 14. In the 9-beep condition, there was a strong increase in attentiveness difficulty between ESM study days ($\gamma_{B,9B} = 0.19$ $[0.15; 0.24]$) that was credibly different from the parameters obtained for the remaining conditions ($\Delta_{\gamma_B,6B-3B} = -0.01$ $[-0.11; 0.10]$; $\Delta_{\gamma_B,9B-3B} = -0.14$ $[0.04; 0.24]$; $\Delta_{\gamma_B,9B-6B} = -0.15$ $[0.08; 0.22]$). This translates into a steep decline of an average attentiveness probability of .98 (corresponding to an expected C/IER rate of 2%) on the first day of the ESM study to .83 (corresponding to an expected C/IER rate of 17%) on day 14. Descriptively, trajectories of average attentiveness probabilities remained comparable across frequency conditions within the first study week and diverged from each other only in the second.

Our conclusions remained robust across model specifications (E2.2; full results given in the appendix), with the only exception being that results concerning the presence of between-day effects in the 6-beep condition were inconclusive. Again, specifications where habitual decay was modeled as a function of the time passed since the start of the study were susceptible to convergence issues, and only the specification using screen times up to the first branching and modeling within-day changes in attentiveness difficulty as a function of the number of beeps converged.

There was no evidence for quadratic within- and/or between-day effects on attentiveness difficulties in any of the conditions across all considered model specifications (E2.3). Again, we encountered convergence issues when quadratic effects were included for both within- and between-day effects in the same model.

Investigating the Relationship Between Model-Implied C/IER and Other Engagement Measures

We did not find any relationships that were credibly different from zero between model-implied C/IER with attention checks, momentary self-reported attentiveness, and non-compliance. Contrary to H3.1, the average attentiveness of the 10 respondents who failed attention check items ($\bar{\psi}_{fAC} = -0.17$ $[-0.60; 0.27]$) was not credibly different from the average attentiveness of the remaining respondents who did not ($\bar{\psi}_{pAC} = 0.02$ $[-0.15; 0.21]$; $\Delta_{\bar{\psi}, fAC-pAC} = -0.19$ $[-0.65; 0.21]$). Likewise, contrary to H3.2, the 19 observations with failed attention checks ($\pi_{\dots,fAC} = .96$ $[.93; .98]$) exhibited comparable attentiveness probabilities than the 503 with passed checks ($\pi_{\dots,pAC} = .94$ $[.94; .95]$; $\Delta_{\pi\dots,fAC-pAC} = .01$ $[-.01; .03]$). Contrary to H4, model-implied attentiveness and attentiveness measured with momentary self-reports were unrelated ($cor(\psi, \psi_{SR}) = .01$ $[-.15; .24]$). Further, within- and between-day declines of model-implied attentiveness probabilities were not mirrored by the momentary self-reports (E4). Instead, the measurement model for momentary self-reports yielded within- and between-day effect parameters of essentially zero ($\gamma_{W,SR} = 0.00$ $[-0.01; 0.00]$; $\gamma_{B,SR} = 0.00$ $[-0.01; 0.00]$).

Contrary to H5, the small positive correlation between latent compliance and attentiveness ($.17$ $[-.12; .38]$) was not credibly different from zero. Compliance difficulty increased across ESM study days ($\gamma_{B,C} = 0.04$ $[0.03; 0.07]$), implying increasing non-compliance across days and mirroring between-day attentiveness difficulty trajectories (E5). Within days, however, contrasting linearly increasing within-day attentiveness difficulties, there was a minor decrease in compliance difficulties ($\gamma_{B,C} = -0.02$ $[-0.04; -0.00]$), implying slightly decreasing non-compliance within ESM study days.

Discussion

As researchers increasingly rely on ESM data to answer questions about individuals' daily lives, understanding how ESM study design influences data quality has become more

important than ever. In the current investigation, the recently developed screen-time-based mixture model by Ulitzsch, Nestler, et al. (2024) allowed us to detect design-related increases in C/IER in ESM data. Specifically, our findings suggest detrimental effects of increases in sampling frequency but not questionnaire length, particularly when a study extends beyond a week. Aside from the practical implications for ESM study design, our findings underline the need to consider different forms of disengagement when investigating the effects of study design on the obtained data.

Base Rate and Time Course of C/IER

The overall C/IER rate of 5% in the current sample matches the previous application of the screen-time-based mixture modeling approach in ESM data with 7 assessments of ca 20 items per day for 7 days (Ulitzsch, Nestler, et al., 2024) and is comparable to the base rate of 8% of observations in a study with 6 assessments of 7 items per day for 14 days flagged with a mixture modeling approach leveraging item responses (Vogelsmeier et al., 2024). Our results further suggest that the distributions of careless and attentive screen times indeed overlap and would not be distinguishable with fixed thresholds (e.g. Jaso et al., 2022, as in), which highlights the advantage of the employed model-based approach. General increases in inattentiveness over study days are in line with our expectations, previous findings (Ulitzsch, Nestler, et al., 2024), and match the time course of compliance observed in other ESM studies (Forkmann et al., 2018; Ono et al., 2019; Rintala et al., 2020; Silvia et al., 2013). Notably, within-day changes in C/IER showed a different trend from within-day changes typically observed for compliance. Instead of decreasing (as typically observed for compliance), C/IER was found to increase over the course of an individual day. The underlying changes in attentiveness difficulty were linear; there was no evidence for quadratic within-day effects on attentiveness difficulty. Participants' wake times have been suggested to explain lower rates of compliance in early hours of the day (Eisele et al., 2022; Rintala et al., 2020), implying that the circadian cycles of compliance

may indeed not be driven by respondent disengagement, which could explain the different within-day changes of compliance and C/IER.

ESM Study Design and Data Quality and Quantity

Higher sampling frequencies led to subtle decreases in data quality that became increasingly large after approximately one week of data collection, while we could not detect changes in C/IER based on the questionnaire length. This contrasts previous findings from the same dataset concerning compliance, self-reported attentive responding, and burden, which were all found to be affected by increases in questionnaire length but not sampling frequency (Eisele et al., 2022). A possible explanation for these diverging findings could be that increases in sampling frequency and questionnaire length result in different types of burden, that may invite participants to apply different burden-reduction strategies. For increases in questionnaire length, the actual length of the interruption (approximately 215s for the long questionnaire version compared to 108s for the short questionnaire) may be perceived as the most burdensome aspect of the sampling. As a result, longer questionnaires may trigger more non-compliance in situations that can not be interrupted for an extended time (e.g., a social interaction), as well as more respondent burden and lowered self-reported attentiveness. On the other hand, when participants are faced with a high sampling frequency for a sustained time, the repetitiveness of the assessments may represent a key problem and may invite participants to engage in fast forms of C/IER that will not be flagged with self-report items. Future research could try to disentangle such effects and explore different types of respondent burden in more detail. Whatever mechanism explains the diverging findings, the present results underline the need to evaluate design choices comprehensively, with non-compliance being only one of several problematic consequences of more intensive study designs.

A related salient finding in the current study was the lack of correspondence between different measures of respondent engagement. Neither self-reported attentiveness,

nor the attention check item, nor non-compliance did match the model-implied C/IER. Given the well-known limitations of other measures of C/IER, the lack of correspondence is perhaps not surprising. Specifically, the attention check item was only presented four times in this study and as a result, it is a coarse measure that likely misses many instances of C/IER. In addition, as noted previously, it is unlikely that participants rush through the questionnaire but respond carefully to the self-report measure of attentive responding at the end of the questionnaire. Therefore, the self-report item likely taps into a different form of inattentiveness than the screen-time-based model.

Future research is needed to evaluate the practical significance of the observed rates of C/IER and their implications for ESM study design, as the rate of C/IER is merely one of many factors that influence the choice of an optimal ESM study design. Therefore, the advantages and disadvantages of different sampling frequencies need to be weighed carefully against each other to determine the optimal design for a particular research question. To illustrate, employing a design with 9 assessments per day for 14 days is likely to result in substantially more attentive data points than a design with 3 assessments per day, even when considering the higher rate of C/IER associated with the higher sampling frequency observed in the current study (9 beeps $\times$ 14 days $\times$ .80 compliance $\times$ .92 attentiveness probability $\approx$ 93 attentive data points compared to 3$\times$ 14 $\times$ .80 $\times$ .96 $\approx$ 32 attentive data points). This suggests that higher sampling frequencies may still allow researchers to obtain more accurate estimates than lower sampling frequencies in many cases. Yet, such calculations rest on the assumption that C/IER instances can reliably be detected and filtered or down-weighted (see Ulitzsch, Domingue, et al., 2023; Ulitzsch, Shin, & Lüdtke, 2023) in subsequent analyses. If this is not the case, undetected C/IER instances may bias conclusions in unpredictable ways (Huang et al., 2015). Further, we point out that attentive responses may not necessarily be valid and that C/IER is not the only type of response bias that may occur due to repeated administration of the same measures and that may be impacted by ESM study design choices. For instance, attentive

item responses may be confounded with response styles (i.e., idiosyncrasies in how respondents use rating scales; Deng et al., 2018; Hasselhorn et al., 2024) or other types of measurement reactivity bias (Arslan et al., 2021). Indeed, Hasselhorn et al. (2024) found longer questionnaires in ESM studies to be associated with higher reliance on midpoint and extreme response styles. Likewise, item evaluation may be altered by familiarity, learning, and increased and repeated reflection on the item content.

Possible effects of higher frequencies on recruitment rates (Smyth et al., 2021) and higher payments needed to motivate participants to comply with a higher sampling frequency are other factors to consider when selecting a sampling frequency. In addition, the recent literature has stressed the need to align sampling strategy and research question (e.g., Dejonckheere & Erbas, 2021). Increases in C/IER rates over time in our data suggest that when a high sampling frequency is needed in a study (e.g., when a researcher wants to model fast-paced lagged associations), shortening the study duration may help maintain high data quality. Thereby, our study provides evidence to support existing practices in ESM research, where sampling frequency and study duration are typically adjusted to each other (Kaurin et al., 2023; Wrzus & Neubauer, 2023). When it is not possible to shorten the study duration, researchers may draw on planned missingness designs and randomization of item display and order to keep participants engaged (as discussed in Arslan et al., 2021; Silvia et al., 2014). Alternatively, researchers may also consider using measurement burst designs (Nesselroade, 1991) with multiple waves of high-frequency assessments to safeguard data quality, although the consequences of such complex sampling strategies still need to be evaluated.

Limitations and Future Research

The lacking agreement among different measures of C/IER highlights the challenge to evaluate the accuracy of any individual C/IER detection method. The screen-time-based mixture modeling approach poses a promising tool for investigating C/IER and offers many

advantages compared to other methods (e.g., drawing on subject-matter considerations for formulating attentive and inattentive component models and communicating C/IER classification uncertainty; Ulitzsch, Nestler, et al., 2024). Yet, the employed method rests on important and, even though theoretically plausible, untested assumptions, for instance regarding the temporal form of changes in attentive response times, and the independence of inattentive screen times from person and occasion characteristics (as described in Ulitzsch, Nestler, et al., 2024). The tenability of these assumptions is difficult to evaluate but crucial to the accurate detection of C/IER, and therefore represents a topic that should be addressed in the future. However, in the context of the current study, we expect model misspecifications to affect overall rates of C/IER more than differences between groups receiving different designs, and the influence on our findings should therefore be limited. In addition, the detection of design-related effects in the expected direction offers evidence supporting the validity of the method (as discussed in Ulitzsch, Nestler, et al., 2024).

Previous investigations of the current data observed that next to a non-linear decrease in average screen times over study days aligning with the law-of-practice effect, the within-person(/error) variance of mean screen times increased linearly over time (Eisele et al., 2023). It is unclear to what extent the increase in variance can be attributed to a higher rate of C/IER towards the end of the study. In this context, it is interesting to consider that deviating screen times that are higher than expected according to the law-of-practice effect would be associated with a higher probability of C/IER in the model-based approach, as can be visually observed in Figure 2. These deviating, slow responses contradict the consideration that inattentive responses are generally faster. Yet, it could be argued that the outlying, slow median screen times also indicate the presence of distraction in the environment and multitasking, and therefore point towards a suboptimal response situation. It may be interesting to model slow forms of C/IER in ESM data separately in the future.

The high complexity of the modeling approach combined with the relatively small

sample size resulted in non-convergence of some models of our robustness analyses. Overall, our analyses indicated that the results were not influenced by small variations in the operationalizations of the screen times and the way time was modeled. However, convergence issues arose more frequently in some of the specifications. Notably, higher non-convergence rates were observed in models in which the median screen time was based on both non-branched and branched items. These convergence issues could be due to these screen time measures being less reliable because different items per branch are used. We therefore advice against this approach in future investigations, a finding that is particularly relevant for studies with adaptive sampling strategies. In addition, high rates of non-convergence were observed in models that included a continuous measure of time for the habitual decay function. Arguably, modeling the habitual decay parameter as continuous, even though one of our post-registered goals, may also be less sensible, as it is difficult to imagine how respondents would, for instance, become more proficient and thus faster at filling in the questionnaire simply as a function of time passing (e.g., also during the night). This may explain why non-convergence rates were higher in models that included this specification. For within-day changes, it is easier to imagine changes in attentiveness that follow a circadian rhythm (either due to natural fluctuations in human performance or activities) that can be captured with a continuous modeling approach. Our results further indicated that when only 3 beeps per day were administered, within-day changes in attentiveness were even easier to estimate using continuous time of the day. Based on our results, we suggest that future applications of the screen-time-based mixture model should use measurement occasion to model the habitual decay function, and explore with-day changes of attentiveness difficulty as a function of both the number of administered beeps and time of the day.

For some of the investigated hypotheses, the observed point estimates were pointing in the expected direction but the credibility intervals were large and included zero (e.g., for the link between C/IER and compliance). These large credibility intervals and the

resultant reduction in statistical power go back to the combination of a comparably small sample and the employed model's complexity, with indicators for measuring respondent attentiveness being latent (see Ulitzsch et al., 2020, for further discussions). Repeating similar analyses in larger samples and/or with priors informed by the current results would be desirable to obtain more precise estimates.

Larger samples would also facilitate even more comprehensive analyses of the effects of ESM study design factors on C/IER occurrence and trajectories. Future studies with larger samples may investigate interaction effects among questionnaire length and sampling frequency. It may, for instance, well be that shorter questionnaires buffer declines in attentiveness due to a high sampling frequency. Further, while the employed model allows for attentiveness to vary across occasions, it is assumed to be constant within occasions. However, increases in C/IER across lengthy questionnaires are a well-documented phenomenon (Baer et al., 1997; Berry et al., 1992; Bowling et al., 2020; Galesic & Bosnjak, 2009; Ulitzsch, Buchholz, et al., 2024; Ulitzsch, Shin, & Lüdtke, 2023; Ulitzsch, Yildirim-Erbasli, et al., 2022) and especially for the long questionnaire condition administering 60 items per assessment, it is plausible that some respondents were attentive at the beginning of an administrations' questionnaire, but rushed through it towards the end. It may well be that not accounting for such within-occasion changes in C/IER underestimates C/IER occurrence. Future research may further investigate how within-occasion C/IER trajectories are impacted by ESM study design.

Finally, the current results were obtained in a student sample, and rates of C/IER and design-related effects may be different in other (clinical) populations which may differ in their motivation and the level of environmental distraction they encounter, both of which have been suggested to relate to C/IER (Meade & Craig, 2012). As ESM is increasingly used in clinical populations, future research on the effects of design choices beyond student populations is warranted.

Conclusion

Our findings indicate that higher sampling frequencies in ESM studies lead to drops in data quality, particularly when data collection extends beyond one week. The larger amount of data obtained through a high-frequency sampling comes at the cost of lower-quality data. Increases in questionnaire length were not associated with increases in C/IER in the current sample. Combined with previous results on decrease in compliance associated with longer questionnaires, this finding suggests that different design choices may be associated with different types of participant burden and have distinct effects on the collected data. Future research into design-related changes in ESM data quality is warranted. We believe that the employed screen-time-based approach poses a versatile tool to this end.

References

Arias, V. B., Garrido, L., Jenaro, C., Martinez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*, 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2021). Routinely randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports. *Psychological Methods*, *26*(2), 175–185. https://doi.org/10.1037/met0000294

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality aAssessment*, *68*(1), 139–151. https://doi.org/10.1207/s15327752jpa6801_11

Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340. https://doi.org/10.1037/1040-3590.4.3.340

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, 1–21. https://doi.org/10.1177/1094428120947794

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2021). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*. https://doi.org/10.1177/10944281211056520

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

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

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00102

Dejonckheere, E., & Erbas, Y. (2021). Designing an experience sampling study. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing esm studies* (pp. 33–70). Center for Research on Experience Sampling; Ambulatory Methods, Leuven.

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

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

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.

Forkmann, T., Spangenberg, L., Rath, D., Hallensleben, N., Hegerl, U., Kersting, A., & Glaesmer, H. (2018). Assessing suicidality in real time: A psychometric evaluation of self-report items for the assessment of suicidal ideation and its proximal risk factors

using ecological momentary assessments. *Journal of Abnormal Psychology*, *127*(8), 758–769. https://doi.org/10.1037/abn0000381

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

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 163–174). Chapman Hall.

Guo, J., Gabry, J., & Goodrich, B. (2018). *Rstan: R interface to Stan* [R package version 2.18.2]. https://CRAN.R-project.org/package=rstan

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

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

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

Hasselhorn, K., Ottenstein, C., Meiser, T., & Lischetzke, T. (2024). The effects of questionnaire length on the relative impact of response styles in ambulatory assessment. *Multivariate Behavioral Research*, 1–15. https://doi.org/10.1080/00273171.2024.2354233

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. (2022). 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

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H., 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

Kaurin, A., King, K. M., & Wright, A. G. (2023). Studying personality pathology with ecological momentary assessment: Harmonizing theory and method. *Personality Disorders: Theory, Research, and Treatment*, *14*(1), 62–72. https://doi.org/10.1037/per0000596

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. Reis (Ed.), *New directions for methodology of social and behavioral science*. Jossey-Bass.

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. *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

Meers, K., Dejonckheere, E., Kalokerinos, E. K., Rummens, K., & Kuppens, P. (2020). Mobileq: A free user-friendly application for collecting experience sampling data. *Behavior Research Methods*, *52*, 1510–1515. https://doi.org/10.3758/s13428-019-01330-1

Mestdagh, M., & Dejonckheere, E. (2021). Ambulatory assessment in psychopathology research: Current achievements and future ambitions. *Current Opinion in Psychology*, *41*, 1–8. https://doi.org/10.1016/j.copsyc.2021.01.004

Morren, M., van 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

Myin-Germeys, I., & Kuppens, P. (Eds.). (2021). *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies*. Center for Research on Experience Sampling; Ambulatory Methods Leuven; Leuven.

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry*, *17*(2), 123–132. https://doi.org/10.1002/wps.20513

Nesselroade, J. R. (1991). The warp and the woof of the developmental fabric. In R. M. Downs, L. S. Liben, & D. S. Palermo (Eds.), *Visions of aesthetics, the environment & development—The legacy of Joachim F. Wohlwill* (pp. 213–240). Psychology Press.

Neubauer, A. B., Scott, S. B., Sliwinski, M. J., & Smyth, J. M. (2020). How was your day? Convergence of aggregated momentary and retrospective end-of-day affect ratings across the adult life span. *Journal of Personality and Social Psychology*, *119*(1), 185. https://doi.org/10.1037/pspp0000248

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. (2022). Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors. *Assessment*, *29*(8), 1765–1776. https://doi.org/10.1177/10731911211032718

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

Roman, Z. J., Schmidt, P., Miller, J. M., & Brandt, H. (2023). Identifying dynamic shifts to non-compliant behavior in questionnaire responses: A novel approach and experimental validation. https://doi.org/10.31234/osf.io/ydgqz

Schwarz, N. (2012). Why researchers should think “real-time": A cognitive rationale. In M. R. Mehl & T. Conner (Eds.), *Handbook of research methods for studying daily life* (Vol. 22). Guilford.

Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review*, *31*(4), 471–481. https://doi.org/10.1177/0894439313479902

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods*, *46*, 41–54. https://doi.org/10.3758%2Fs13428-013-0353-y

Smyth, J. M., Jones, D. R., Wen, C. K., Materia, F. T., Schneider, S., & Stone, A. (2021). Influence of ecological momentary assessment study design features on reported willingness to participate and perceptions of potential research studies: An experimental study. *BMJ open*, *11*(7), e049154. https://doi.org/10.1136/bmjopen-2021-049154

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

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*(1-2), 343–351. https://doi.org/10.1016/s0304-3959(03)00040-x

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, *16*(3), 199–202. https://doi.org/10.1093/abm/16.3.199

Ulitzsch, E., Buchholz, J., Shin, H. J., Bertling, J., & Lüdtke, O. (2024). Using a novel multiple-source indicator to investigate the effect of scale format on careless and insufficient effort responding in a large-scale survey experiment. *Large-scale Assessments in Education*, *12*(18). https://doi.org/10.1186/s40536-024-00205-y

Ulitzsch, E., Domingue, B. W., Kapoor, R., Kanopka, K., & Rios, J. (2023). A probabilistic filtering approach to non-effortful responding. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12567

Ulitzsch, E., Nestler, S., Lüdtke, O., & Nagy, G. (2024). A screen-time-based mixture model for identifying and monitoring careless and insufficient effort responding in ecological momentary assessment data. *Psychological Methods*. https://doi.org/10.1037/met0000636

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2023). Using response times for joint modeling of careless responding and attentive response styles. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986231173607

Ulitzsch, E., Shin, H.-J., & Lüdtke, O. (2023). Accounting for careless and insufficient effort responding in large-scale survey data—Development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-02053-6

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level nonresponse. *British Journal of Mathematical and Statistical Psychology*, *73*(1), 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in survey data. *British Journal of Mathematical and Statistical Psychology*, *75*, 668–698. https://doi.org/10.1111/bmsp.12272

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 Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12317

Vogelsmeier, L. V. D. E., Uglanova, I., Rein, M. T., & Ulitzsch, E. (2024). Investigating contextual correlates of inattentive responding in ecological momentary assessment data with a confirmatory mixture IRT model. https://doi.org/10.31234/osf.io/p9cfm

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

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538

Appendix

Robustness Evaluations

Table A1
Sampling frequency effects for different model specifications

ST	WD	HD	$\pi_{...,3B}$	$\pi_{...,6B}$	$\pi_{...,9B}$	$\gamma_{W,3B}$	$\gamma_{W,6B}$	$\gamma_{W,9B}$	$\gamma_{B,3B}$	$\gamma_{B,6B}$	$\gamma_{B,9B}$
all	mo	mo	0.97 [0.96; 0.98]	0.95 [0.94; 0.96]	0.93 [0.93; 0.94]	0.59 [0.05; 1.15]	0.23 [0.07; 0.35]	0.12 [0.05; 0.20]	-0.02 [-0.12; 0.08]	0.06 [0.02; 0.11]	0.17 [0.13; 0.21]
to branch	mo	mo	0.97 [0.95; 0.98]	0.94 [0.93; 0.95]	0.93 [0.91; 0.93]	0.51 [0.00; 1.01]	0.18 [0.04; 0.30]	0.13 [0.06; 0.21]	0.05 [-0.03; 0.15]	0.05 [0.00; 0.11]	0.19 [0.16; 0.24]
with branch	mo	mo	0.97 [0.96; 0.98]	0.95 [0.94; 0.96]	0.93 [0.92; 0.94]	0.61 [0.08; 1.08]	0.20 [0.08; 0.35]	0.10 [0.02; 0.16]	-0.00 [-0.10; 0.09]	0.06 [0.00; 0.13]	0.15 [0.11; 0.20]
all	hr	mo	0.97 [0.96; 0.98]	0.95 [0.94; 0.96]	0.94 [0.92; 0.95]	0.16 [0.05; 0.29]	0.10 [0.05; 0.16]	0.08 [0.03; 0.12]	0.00 [-0.10; 0.11]	0.06 [-0.00; 0.12]	0.16 [0.11; 0.20]
to branch	hr	mo	0.97 [0.95; 0.98]	0.94 [0.93; 0.95]	0.92 [0.91; 0.94]	0.11 [0.02; 0.20]	0.09 [0.03; 0.14]	0.08 [0.04; 0.12]	0.06 [-0.02; 0.14]	0.04 [-0.01; 0.10]	0.19 [0.15; 0.23]
with branch	hr	mo	0.97 [0.96; 0.98]	0.95 [0.94; 0.96]	0.93 [0.92; 0.94]	0.15 [0.05; 0.28]	0.09 [0.04; 0.15]	0.07 [0.03; 0.11]	0.00 [-0.08; 0.10]	0.06 [0.00; 0.12]	0.15 [0.11; 0.20]
all	mo	hr									
to branch	mo	hr	0.97 [0.96; 0.98]	0.95 [0.94; 0.96]	0.94 [0.93; 0.95]	0.45 [-0.05; 1.07]	0.16 [-0.00; 0.29]	0.12 [0.04; 0.18]	0.04 [-0.07; 0.16]	0.03 [-0.02; 0.07]	0.19 [0.14; 0.23]
with branch	mo	hr									
all	hr	hr									
to branch	hr	hr									
with branch	hr	hr									

Notes: An empty line indicates non-convergence. ST: screen-time measure; WD: within-day effect on attentiveness modeled as a function of the measurement occasion (mo) or time of the day (hr); HD: habitual decay modeled as a function of the measurement occasion (mo) or time since the start of the study (hr); $\pi_{...,3B}$, $\pi_{...,6B}$, $\pi_{...,9B}$ give average attentiveness probabilities for conditions with 3, 6, and 9 beeps per day; $\gamma_{W,3B}$, $\gamma_{W,6B}$, and $\gamma_{W,9B}$ give within-day effects on attentiveness difficulties for conditions with 3, 6, and 9 beeps per day; $\gamma_{B,3B}$, $\gamma_{B,6B}$, and $\gamma_{B,9B}$ give between-day effects on attentiveness difficulties for conditions with 3, 6, and 9 beeps per day; 95% credibility intervals are given in squared brackets.

Table A2

Questionnaire length effects for different model specifications

ST	WD	HD	$\pi_{...,SQ}$	$\pi_{...,LQ}$	$\gamma_{W,SQ}$	$\gamma_{W,LQ}$	$\gamma_{B,SQ}$	$\gamma_{B,LQ}$
all	mo	mo	0.94 [0.93; 0.95]	0.95 [0.94; 0.96]	0.14 [0.06; 0.22]	0.18 [0.07; 0.28]	0.12 [0.08; 0.16]	0.10 [0.06; 0.16]
to branch	mo	mo	0.94 [0.93; 0.95]	0.94 [0.93; 0.95]	0.17 [0.09; 0.27]	0.13 [0.04; 0.22]	0.16 [0.11; 0.21]	0.11 [0.07; 0.16]
with branch	mo	mo						
all	hr	mo	0.94 [0.93; 0.95]	0.95 [0.94; 0.96]	0.10 [0.05; 0.14]	0.10 [0.05; 0.16]	0.12 [0.08; 0.17]	0.10 [0.06; 0.14]
to branch	hr	mo	0.94 [0.93; 0.95]	0.94 [0.93; 0.95]	0.10 [0.05; 0.14]	0.08 [0.04; 0.12]	0.15 [0.11; 0.21]	0.10 [0.06; 0.15]
with branch	hr	mo						
all	mo	hr						
to branch	mo	hr	0.95 [0.94; 0.96]	0.95 [0.94; 0.96]	0.15 [0.06; 0.25]	0.11 [0.03; 0.20]	0.15 [0.10; 0.21]	0.10 [0.06; 0.14]
with branch	mo	hr						
all	hr	hr						
to branch	hr	hr						
with branch	hr	hr						

Notes: An empty line indicates non-convergence. ST: screen-time measure; WD: within-day effect on attentiveness modeled as a function of the measurement occasion (mo) or time of the day (hr); HD: habitual decay modeled as a function of the measurement occasion (mo) or time since the start of the study (hr); $\pi_{...,SQ}$ and $\pi_{...,LQ}$ give average attentiveness probabilities for conditions with short and long questionnaires; $\gamma_{W,SQ}$ and $\gamma_{W,LQ}$ give within-day effects on attentiveness difficulties for conditions with short and long questionnaires; $\gamma_{B,SQ}$ and $\gamma_{B,LQ}$ give between-day effects on attentiveness difficulties for conditions with short and long questionnaires; 95% credibility intervals are given in squared brackets.