# The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population

*Article*

Assessment
1–16

DOI: 10.1177/1073191120957102
journals.sagepub.com/home/asm
SAGE

# The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population

**Gudrun Eisele[1], Hugo Vachon[1], Ginette Lafit[1], Peter Kuppens[1], Marlies Houben[1], Inez Myin-Germeys[1], and Wolfgang Viechtbauer[1,2]**

## Abstract

Currently, little is known about the association between assessment intensity, burden, data quantity, and data quality in experience sampling method (ESM) studies. Researchers therefore have insufficient information to make informed decisions about the design of their ESM study. Our aim was to investigate the effects of different sampling frequencies and questionnaire lengths on burden, compliance, and careless responding. Students ($n = 163$) received either a 30- or 60-item questionnaire three, six, or nine times per day for 14 days. Preregistered multilevel regression analyses and analyses of variance were used to analyze the effect of design condition on momentary outcomes, changes in those outcomes over time, and retrospective outcomes. Our findings offer support for increased burden and compromised data quantity and quality with longer questionnaires, but not with increased sampling frequency. We therefore advise against the use of long ESM questionnaires, while high-sampling frequencies do not seem to be associated with negative consequences.

## Keywords

experience sampling, ecological momentary assessment, sampling frequency, questionnaire length, data quality

The use of momentary self-report measures in daily life is appealing to researchers for various reasons. The wish to maximize ecological validity, reduce recall biases, and look at dynamic processes in daily life has led to an increase in studies using ambulatory self-report methods, such as the experience sampling method (ESM; Larson & Csikszentmihalyi, 1983; Myin-Germeys et al., 2018). In an ESM study, participants are typically asked to fill in several short questionnaires at random times during the day over several days. Setting up an ESM study requires making numerous design choices (Janssens et al., 2018). Many of these choices are thought to be associated with a trade-off between the aim to collect as much information as possible on the one hand, and the risk of compromising data quantity and quality as a result of the intensive assessment on the other (Arslan et al., 2019; May et al., 2018). Yet methodological research into the effects of different design choices on data quantity and quality is sparse, which means that there is currently insufficient empirical evidence to base design decisions on (Himmelstein et al., 2019).

Two central design choices in every ESM study are the sampling frequency (i.e., the number of assessments or "beeps" per day) and the questionnaire length, which has varied from 1 to more than 50 beeps per day and 1 to 135 items per beep in previous studies (Ebner-Priemer & Sawitzki, 2007; Morren et al., 2009; Ono et al., 2019; Vachon et al., 2019). The unknown time course of many phenomena calls for high-sampling frequencies (Collins & Graham, 2002). Likewise, reliability concerns with scales of less than three items (Shrout & Lane, 2012) and the wish to gather as much potentially relevant information as possible call for long questionnaires. However, both high-sampling frequencies and longer questionnaires are thought to increase the perceived burden experienced by study participants (Shiffman et al., 2008; Trull & Ebner-Priemer, 2020; Piasecki et al., 2007), which in turn is expected to lead to

[1]KU Leuven, Leuven, Belgium
[2]Maastricht University, Maastricht, Limburg, Netherlands

**Corresponding Author:**
Gudrun Eisele, Department of Neurosciences, Research Group Psychiatry/Center for Contextual Psychiatry, KU Leuven, Kapucijnenvoer 33 bus 7001 (blok h), Leuven 3000, Belgium.
Email: gudrunvera.eisele@kuleuven.be

reduced data quantity and quality (Fuller-Tyszkiewicz et al., 2013; Santangelo et al., 2013). In particular, the quantity and quality of the data collected in ESM studies depends on the participants' compliance to the protocol. When experiencing the protocol as overly burdensome, participants might refrain from responding to the assessments leading to a reduction in data quantity (Stone et al., 2003). If this noncompliance is systematic (i.e., associated with certain states), it can lead to biases in the data that cannot always be corrected for post hoc, thereby compromising data quality (Courvoisier et al., 2012).

Besides not responding, participants might also engage in more subtle behaviors to reduce burden, such as responding carelessly (Jones et al., 2018). Careless responding is defined as responding without paying sufficient attention to the questions (Meade & Craig, 2012). While this is a well-established problem in cross-sectional research (Meade & Craig, 2012), this threat to data quality has received relatively little attention in the ESM literature which has long focused on noncompliance as the sole measure of data quality (for recent exceptions, see van Berkel et al., 2018, 2019). Since careless responding can lead to bias and/or increased measurement error, it is important to also examine its prevalence and to consider steps to reduce or eliminate its occurrence.

## *Sampling Frequency*

The relationship between sampling frequency, perceived burden, data quantity, and data quality has been the topic of several investigations. Surprisingly, studies that manipulated sampling frequencies have not found support for the expected decrease in compliance with higher frequencies. In one study, 85 students responded to 20 beeps (six items) either within 1 day or within 2 days (Walsh & Brinker, 2016). The groups did not differ in terms of compliance, completeness of responses, or response delays. In a study that included measures of perceived burden, 91 chronic pain patients were randomly assigned to receive either 3, 6, or 12 beeps per day (up to 19 items depending on the branching) for 14 days or to a no ESM control group (Stone et al., 2003). Higher frequencies were not associated with changes in compliance, but with increases in perceived burden. Finally, two studies reported no differences in compliance but found that higher frequencies were associated with lower response latencies. In one of these studies, 162 participants were randomly assigned to report on their happiness with three items either one, three, or six times daily for 13 days (Conner & Reid, 2012). In a second study, McCarthy et al. (2015) had 110 smokers, who were trying to quit, respond to either one or six daily questionnaires (containing a maximum of 28 items) for 4 weeks. Meta-analyses and a pooled data analysis have also not found support for the influence of sampling frequency on compliance (Jones et al., 2018; Morren et al., 2009; Ono et al., 2019; Soyster et al., 2019). However, a recent meta-analysis of ESM studies in the field of psychopathology did observe a significant negative association between assessment frequency and compliance (Vachon et al., 2019).

## *Questionnaire Length*

In traditional cross-sectional assessments, longer questionnaires have been found to be associated with increased perceived burden (Rolstad et al., 2011), lower participation, and changes in data quality (Galesic & Bosnjak, 2009). To our knowledge, only one pilot study has compared different questionnaire lengths in ESM (Intille et al., 2016). In this investigation, 33 participants completed either ESM assessments on a phone six times a day (six items) or micro-interactions (one item) on a smartwatch for 4 weeks. Participants receiving micro-interactions had higher compliance and reported fewer distractions, but also experienced the assessments as more disruptive. This suggests that a high-assessment frequency with a short questionnaire might lead to higher data quantity and possibly quality than a longer questionnaire with a lower frequency. However, since frequency, length, answer options of questions, and assessment mode (phone vs. smartwatch) were all manipulated simultaneously, it is not possible to disentangle what exactly caused the group differences. Moreover, meta-analyses on predictors of compliance have come to varying conclusions. While Morren et al.(2009) found that shorter questionnaires were associated with better compliance, two pooled data analyses (Rintala et al., 2018; Soyster et al., 2019) and other meta-analyses have not detected an association between questionnaire length and compliance (Jones et al., 2018; Ono et al., 2019; Vachon et al., 2019). However, findings from these analyses have to be interpreted cautiously, since they faced limitations such as low variability in included questionnaire lengths (Rintala et al., 2018) and lack of sufficient information on questionnaire length or compliance in reviewed articles (Jones et al., 2018; Ono et al., 2019; Vachon et al., 2019).

## *The Current Study*

The aim of the current investigation was to test how sampling frequency and questionnaire length influence (a) perceived burden, (b) compliance, and (c) careless responding. We expected that a higher frequency and longer questionnaire would be associated with higher reported burden, lower compliance, and increased careless responding. Additionally, we expected that questionnaire length would moderate the effect of sampling frequency on perceived burden, compliance, and careless responding (i.e., the effects of high-sampling frequencies become stronger with longer questionnaires). Specifically, we thought that participants would be able to tolerate frequent measures

relatively well when the questionnaire was short, but that the combination of frequent sampling and a long questionnaire would overburden participants and lead to a stronger drop in data quality and quantity. Furthermore, the effects of more intensive measurements on perceived burden, compliance, and careless responding were expected to strengthen over time. Specifically, participants were expected to initially try to adhere even to intensive protocols, but that as the study progresses this initial motivation would wear off, participants would feel increasingly burdened by too intensive protocols and data quality and quantity would decline. This hypothesis was based on previous research where declines in compliance over time have been repeatedly observed (e.g., Courvoisier et al., 2012; Forkmann et al., 2018; Rintala et al., 2018; Silvia et al., 2013). There has been less research into changes in momentary burden or data quality over time but reported momentary burden has been found to increase over time (Rintala et al., 2020) and proxies of data quality have been found to decrease over time in some cases (e.g., Reynolds et al., 2016; van Berkel et al., 2019). In line with most previous studies, we also expected that there would be a decline in all measures of data quality and quantity in all groups, however, this hypothesis was not preregistered. The remaining hypotheses and analyses were preregistered.

## Method

### Sample

Participants were required to be Dutch speaking, between 18 and 30 years old, and currently enrolled as a student at a university. Additionally, they could not have previously taken part in an ESM study. Participants were recruited by advertisements on social media and on campus sites. Based on power calculations, a sample of 25 participants per group was deemed sufficient, adding up to a total of 150 participants. We estimated power using a Monte Carlo-based simulation approach to compare compliance between the groups (code is included on the OSF page https://osf.io/pzx8r/?view_only=7afaed46a9b24ebcbf4e8947644015e8). The power analysis was focused on the two main effects of interest (beep frequency and questionnaire length) and assumed average compliance rates of 80%, 70%, and 60% for the three, six, and nine-beep conditions for the short version of the questionnaire and 70%, 60%, and 50% for the long version of the questionnaire (and standard deviations of 1.0 and 0.6 for the between-person and between-day logit-transformed compliance proportions; these SD values were based on prior studies conducted by our research group). In the absence of previous work to retrieve an expected effect size, these differences in compliance rates were considered sufficiently large to be practically important. Based on the simulated conditions, power was 98% for the beep frequency and 76% for the questionnaire length effect. Power might be lower for other analyses.

### Procedure

All study procedures were approved by the Social and Societal Ethics Committee at the University of Leuven (KU Leuven), Belgium. After providing informed consent, participants were randomly assigned to one of the six experimental conditions (30- or 60-item questionnaire, three, six, or nine times per day) and asked to fill out a number of baseline questionnaires (not relevant for the purpose of this study). During this visit, participants were briefed individually for approximately 20 minutes about the ESM procedure following a standardized protocol which is included in the preregistration. As determined by the randomization procedure, participants were told how many times per day they would be signaled to complete the questionnaire and how much money they would receive for their participation. The time frames and answer options of items of the questionnaire (short or long depending on condition) were explained and items with higher complexity were paraphrased in a standardized way. Participants were asked to fill in the questionnaire on the study phone once to practice. An ESM testing period of 14 days was then started on the next day. On the day following the end of the ESM testing period, a second session was conducted during which participants were asked to fill out the posttest questionnaires (more information below). A qualitative interview lasting approximately 10 minutes was conducted with a randomly chosen subsample of approximately 10 participants per condition. After the second session, participants received 40, 60, or 80 Euros in form of gift vouchers depending on the number of beeps per day. We reduced the incentive to 25, 30, and 40 euros if they responded to less than a third of the beeps (this amount corresponds to the full payment for both lab sessions plus roughly a third of the payment for the ESM period) to avoid free riding. Participants were informed about a minimum compliance during the briefing, but never about the exact requirement, to avoid that this would lower their compliance. This choice was based on the experiences of researchers involved in setting up the current study and corresponds to the standard procedure implemented in ESM studies in our center. The exact instructions given to participants can be found on the OSF page of this project. Testing was performed between February and July 2019.

### ESM Protocol and Experimental Manipulation

The ESM period lasted 14 days and was signal contingent. Participants received three, six, or nine beeps per day, depending on the condition they had been assigned to during the baseline session. Beeps were semirandomly spread

between 9:00 a.m. and 10:30 p.m. Participants were randomly assigned to receive either a short (30 items with full branching) or long questionnaire (60 items with full branching) during the study (see online supplementary materials). Both questionnaires were designed to resemble typically employed ESM questionnaires, with items about cognitions, emotions, and context. Items were always presented in the same order and it was not possible to skip items. Participants received a smartphone (Motorola DEFY+) with the app mobileQ (Meers et al., 2020) installed on it for the duration of the study. The phone beeped and/or vibrated (this could be customized by the participant depending on circumstances) for 90 seconds or until the participant started filling in the questionnaire. If the questionnaire was not started within this time frame, it became unavailable. Once participants started filling in the questionnaire, they had a maximum of 90 seconds to answer each question. Again, if no response was received within this time frame, the questionnaire was terminated.

## Measures[1]

### Time Invariant Variables

*Demographics.* Gender and age were assessed at baseline.

*Questionnaire length.* A factor was used to indicate the questionnaire length, with the Value 1 for the short questionnaire and 2 for the long questionnaire.

*Sampling frequency.* A factor was used to indicate the sampling frequency, with the Values 1, 2, and 3 corresponding to the three, six, and nine beeps per day conditions, respectively.

*Retrospective perceived burden and related aspects.* We assessed the experience of taking part in the study with 16 items that were partly based on the experience sampling feasibility questionnaire (Edwards et al., 2016). The items were divided into four thematic subscales: burden (e.g., "I found it stressful to use the app."), ease of use (e.g., "The questionnaires on the phone were easy to complete."), instructions (e.g., "The training I received at the beginning of the study was adequate to use the app for two weeks."), and reward ("How much money do you find appropriate to receive for the participation in this study?"). We calculated means of the items for subscales 1 to 3 ($\omega$ = .76, .60, and .32, respectively). The full questionnaire in Dutch and an English translation are provided in the online supplementary materials.

*Retrospective reported careless responding.* Retrospective careless responding was measured with three items ("I didn't pay much attention to what the questions actually meant"; "I filled out the questions without thinking about myself"; and "I responded carelessly to the questions"; adapted from Huang et al., 2012) of which we calculated a mean ($\omega$ = .71). The answer options of these questions ranged from 1 = *never* to 7 = *always* on a 7-point Likert-type scale. The Dutch items can be found in the online supplementary materials.

### Time Variant Variables

*Day number.* Study day was coded numerically, ranging from 0 to 13.

*Beep.* A "beep block" variable was coded, ranging from 0 to 8. In the nine-beep condition, the beeps within a day were simply numbered from 0 to 8. In the three- and six-beep conditions, beeps were assigned the number of the nine-beep condition interval that they fell in (e.g., a beep in the three-beep condition delivered at 12:05 was assigned the number 2 because it falls into the time window where the third beep of the nine-beep condition would be triggered).

*Momentary perceived burden.* Burden was measured by taking the mean of the two items assessing burden at every beep ("Filling in the questionnaire took effort" and "This beep disturbed me"; Cronbach's $\alpha$ was .78 for person means and 0.53 for deviations from these means). Both items were rated on a 7-point Likert-type scale (ranging from *not at all* to *very much*).

*Compliance.* Compliance at the beep level was defined as having responded to the last item of the ESM questionnaire (coded 1 for "yes" and 0 for "no"). In cases of pervasive technical problems, missing data was coded as missing instead of noncompliant, as described in the analysis section.

*Momentary reported careless responding.* Momentary reported careless responding was assessed with one item ("I filled in the questions attentively"). This question was rated on a 7-point Likert-type scale.

*Momentary objective careless responding/instructed response item.* Momentary objective careless responding was assessed once on Day 3, 6, 9, and 12 by including an additional item in the questionnaire that simply instructed participants to choose the 1 (*not at all*) option of a 7-point Likert-type scale ("Think back about what you were doing just before the beep. Please select answer option 1"). This item was scored dichotomously (i.e., 1 if the "not at all" option was selected, 0 otherwise). To ensure that participants would not anticipate this data quality check, the item was not introduced during the briefing session and presented infrequently.

*Reasons for missing a beep.* Reasons for missing a beep were assessed by asking participants whether they had missed the previous beep and to select all reasons that applied ("Did you, since the last questionnaire that you filled in, knowingly not react to a beep? (Yes/No) Why did you not react? Select all that applies: I could not respond (on time); I did not feel like responding; I was too stressed to respond; I felt disturbed by the beep; Other").

### Analyses

Analyses were performed in R (version 3.6.1; R Core Team, 2019), with the lme4 (version 1.1-21; Bates et al., 2015), psych (version 1.8.12; Revelle, 2018), data.table (version 1.12.2; Dowle & Srinivasan, 2019), and car (version 3.0-3; Fox & Weisberg, 2019) packages. The analyses were preregistered (https://osf.io/pzx8r/?view_only=7afaed46a9b24ebcbf4e8947644015e8). Minor deviations from the preregistered R code are documented in the online supplementary materials.

Unless otherwise stated, three-level (logistic) regression models were used for momentary outcomes with measurements at Level 1, day at Level 2, and persons at Level 3. Random intercepts were included at the person level, for day number nested in person, and for beep block nested in person (i.e., as crossed random effects for the latter two). To test the main effect of questionnaire length and sampling frequency, both were added as factors to the model. Then, an interaction between questionnaire length and sampling frequency was added to the model. Finally, the model (excluding the interaction between questionnaire length and sampling frequency) was extended by adding day number as a time variant predictor and allowing it to interact with the questionnaire length and sampling frequency factors. A random slope for day number was also added to the model. We also conducted a test of the main effect of day number in a model excluding the interaction between day number and length and day number and frequency (not preregistered). Wald-type (chi-square or F) tests were conducted for the fixed effects. Retrospective outcomes were analyzed with two-way ANOVAs which included questionnaire length and sampling frequency as factors. In a second step, the model was extended with the interaction term of questionnaire length and sampling frequency. In the multilevel logistic regression model for compliance, gender was included as an additional control variable in all steps. Gender has repeatedly been found to predict compliance (Rintala et al., 2018; Silvia et al., 2013; Sokolovsky et al., 2014). It was therefore added to the model to avoid that unequal gender distributions across groups would distort the effects of the sampling protocol. Objective momentary careless responding was analyzed with a multilevel regression model as described above, with the exception that the effect of day number was not investigated. Day number was rescaled by dividing it by 13 prior to the analyses to avoid very small coefficients that can lead to model nonconvergence. Rescaling variables does not influence the significance level of coefficients in the models. The coefficients and standard errors reported were transformed back to the original scale of the day number variable to facilitate their interpretation.

### Missing Data

Participants that missed more than one but not more than all data points on one day due to a communicated technical problem were included in all analyses. The data missing due to the technical problem was indicated as missing (as opposed to noncompliant) in the compliance analysis. Participants who missed more than 1 full day of data due to a technical problem were included in all analyses of momentary outcomes until the day of the technical problem that led to more than one day of data loss. All data beyond this point was indicated as missing. Participants for whom the time on the phone was off by at least 1 hour were excluded from the analyses. These choices were made during data collection and were admittedly arbitrary. The reasoning was that more than 1 day of missing data or a disruptive timing of the beeps would distort the experience of taking part in the ESM study. Less intrusive technical problems and other forms of missing data were considered an integral part of an ESM study and therefore handled as noncompliant in the analyses.

## Results

### Sample Characteristics

Of the 163 participants that took part in the baseline session, three participants had to be excluded because they were not currently enrolled as students and two dropped out during the ESM period. Both drop outs had been randomized into the highest intensity condition (nine beeps & long questionnaire). In both cases, the high time investment was given as reason for early termination. Furthermore, two participants were excluded from all analyses because the time on their study phone was found to be off by more than 1 hour at follow-up, which led to the beeps being delivered at wrong times. Demographic characteristics of the remaining sample are reported in Table 1. Four participants missed more than a full day of data each because of technical problems and were therefore excluded from the analyses of follow-up outcomes (for two participants the battery of the phone broke and for two participants the time of the phone was reset for unknown reasons, which led to beeps not being delivered). One participant failed to meet the minimum compliance requirement (less than one third of the scheduled ESM questionnaires were started) but was included in the analyses. Analyses were repeated excluding this outlier (conclusions remained

**Table 1.** Sample Characteristics and Descriptives (for Level 1 Outcomes, Descriptives of Person-Means Are Reported).

| Questionnaire length | | Short | | | Long | | |
|---|---|---|---|---|---|---|---|
| Frequency | | Three beeps | Six beeps | Nine beeps | Three beeps | Six beeps | Nine beeps |
| *N* | | 27 | 25 | 25 | 26 | 25 | 28 |
| % Female | | 63% | 72% | 80% | 85% | 84% | 89% |
| Age[a] | *M* (*SD*) | 21.90 (1.98) | 21.75 (1.97) | 21.31 (1.59) | 21.97 (1.54) | 21.93 (1.93) | 21.51 (1.68) |
| Compliance | *M* (*SD*) | 0.80 (0.17) | 0.84 (0.11) | 0.85 (0.08) | 0.81 (0.14) | 0.81 (0.12) | 0.78 (0.12) |
| | Median | 0.86 | 0.88 | 0.86 | 0.80 | 0.87 | 0.78 |
| Partially compliant beeps | *M* (*SD*) | 0.03 (0.03) | 0.02 (0.02) | 0.02 (0.02) | 0.05 (0.06) | 0.03 (0.03) | 0.04 (0.04) |
| Momentary burden | *M* (*SD*) | 2.37 (0.72) | 2.19 (0.66) | 2.18 (0.74) | 2.99 (0.84) | 2.89 (1.09) | 2.72 (0.83) |
| | Median | 2.21 | 2.24 | 2.04 | 2.79 | 2.74 | 2.65 |
| Momentary reported careless responding | *M* (*SD*) | 6.03 (0.57) | 6.04 (0.67) | 6.00 (0.56) | 5.80 (0.78) | 5.63 (0.49) | 5.89 (0.54) |
| | Median | 6.07 | 6.07 | 5.96 | 5.81 | 5.74 | 5.88 |
| Momentary objective careless responding | *M* (*SD*) | 1.00 (0.00) | 0.95 (0.20) | 0.96 (0.12) | 0.99 (0.07) | 1.00 (0.00) | 0.91 (0.25) |
| Retrospective burden[b] | *M* (*SD*) | 1.58 (0.78) | 1.62 (0.82) | 1.94 (0.59) | 2.11 (0.71) | 2.13 (0.76) | 2.27 (0.98) |
| | Median | 1.33 | 1.44 | 1.89 | 1.89 | 2.22 | 2.33 |
| Retrospective ease of use[c] | *M* (*SD*) | 2.24 (1.30) | 1.88 (1.19) | 2.36 (1.30) | 2.42 (1.23) | 2.48 (1.28) | 2.22 (1.19) |
| | Median | 2.17 | 1.67 | 2 | 2.17 | 2.33 | 2.17 |
| Retrospective instructions[a] | *M* (*SD*) | 1.91 (0.88) | 1.48 (0.59) | 1.64 (0.66) | 1.77 (0.61) | 1.97 (1.04) | 1.35 (0.82) |
| | Median | 1.67 | 1.33 | 1.67 | 1.67 | 1.67 | 1.17 |
| Retrospective reward[d] | *M* (*SD*) | 48.74 (11.94) | 58.0 (12.48) | 66.91 (17.83) | 48.48 (13.60) | 56.88 (14.36) | 75.63 (22.28) |
| Retrospective careless responding[a] | *M* (*SD*) | 0.87 (0.57) | 0.84 (0.56) | 1.17 (0.80) | 0.87 (0.69) | 1.16 (0.87) | 0.96 (0.64) |
| | Median | 1 | 1 | 1 | 0.67 | 1 | 1 |

[a]$N = 152$. [b]$N = 150$. [c]$N = 151$. [d]$N = 141$.

unchanged, except where noted). Due to missing values on individual variables the sample sizes varied slightly per analysis and are stated in the results tables.

## *Momentary Burden*

As hypothesized, the longer questionnaire was associated with significantly higher average momentary reported burden than the short version, $\chi^2(1) = 22.33$, $p < .001$. When all frequency groups are pooled together, the 60-item questionnaire was estimated to lead to a 0.62 point increase (on a 7-point scale) in momentary burden compared with the 30 items questionnaire. However, in contrast to our hypotheses, momentary burden did not vary as a function of sampling frequency, $\chi^2(2) = 2.19$, $p = .33$. There was also no significant interaction between sampling frequency and questionnaire length, $\chi^2(2) = 0.25$, $p = .88$. Additionally, momentary burden was not found to increase over time, $\chi^2(1) = 0.14$, $p = .70$, and the main effects of questionnaire length, $\chi^2(1) = 0.22$, $p = .64$, and sampling frequency did not change over time either, $\chi^2(2) = 5.10$, $p = .08$. For more details, see Table 2.

## *Retrospective Burden and Related Aspects*

A similar pattern emerged for retrospective burden which was found to be higher in participants who had received the long questionnaire, $F(1, 146) = 12.92$, $p < .001$. When all frequency groups were pooled together, this resulted in a mean burden of 2.17 ($SD = 0.82$) in the long questionnaire group as opposed to 1.71 ($SD = 0.74$) in the short questionnaire group. There were no significant differences based on different sampling frequencies, $F(2, 146) = 1.76$, $p = .18$, nor was there a significant interaction between sampling frequency and questionnaire length, $F(2, 144) = 0.25$, $p = .78$. No significant group differences based on questionnaire length, sampling frequency, or their interaction could be detected for the ease of use subscale, $F(1, 147) = 1.04$, $p = .31$; $F(2, 147) = 0.18$, $p = .83$; $F(2, 145) = 1.12$, $p = .33$. For the instructions subscale, there were no significant group differences based on questionnaire length, $F(1, 148) = 0.02$, $p = .89$, or sampling frequency, $F(2, 148) = 2.60$, $p = .08$. However, there was a significant interaction effect between sampling frequency and questionnaire length on the instructions subscale, $F(1, 146) = 3.57$, $p = .031$. Post hoc tests indicated that the effect of length was significantly different in the six-beep compared with the three-beep condition, $F(1, 146) = 4.09$, $p = .045$[2] and to the nine-beep condition, $F(1,146) = 6.38$, $p = .013$. No significant difference between the three- and the nine-beep condition was found, $F(1, 146) = 0.26$, $p = .61$. Participants in the six-beep condition with the long questionnaire reported less satisfaction with the instructions than participants who had received the short questionnaire. For three and nine beeps,

**Table 2.** Momentary Burden.

| | Model 1 | | | | Model 2 | | | | Model 3 | | | | Model 4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predictor | β | *SE* | $\chi^2(1)$ | *p* | β | *SE* | $\chi^2(1)$ | *p* | β | *SE* | $\chi^2(1)$ | *p* | β | *SE* | $\chi^2(1)$ | *p* |
| *Fixed effects* | | | | | | | | | | | | | | | | |
| Intercept | 2.37 | 0.13 | 329.85 | **<.001** | 2.37 | 0.16 | 218.33 | **<.001** | 2.33 | 0.13 | 314.45 | **<.001** | 2.28 | 0.14 | 265.54 | **<.001** |
| Length[a] | 0.62 | 0.13 | 22.33 | **<.001** | 0.63 | 0.23 | 7.55 | **.006** | 0.61 | 0.13 | 22.21 | **<.001** | 0.59 | 0.14 | 17.99 | **<.001** |
| Frequency 2[b] | −0.14 | 0.16 | 0.72 | .40 | −0.17 | 0.23 | 0.57 | .45 | −0.12 | 0.16 | 0.59 | .44 | −0.08 | 0.17 | 0.24 | .62 |
| Frequency 3[b] | −0.23 | 0.16 | 2.17 | .14 | −0.19 | 0.23 | 0.67 | .41 | −0.17 | 0.16 | 1.23 | .27 | −0.05 | 0.17 | 0.07 | .79 |
| Length * Frequency 2 | | | | | 0.07 | 0.33 | 0.05 | .82 | | | | | | | | |
| Length * Frequency 3 | | | | | −0.09 | 0.32 | 0.08 | .78 | | | | | | | | |
| Day number | | | | | | | | | 0.00 | 0.01 | 0.14 | .70 | 0.01 | 0.01 | 1.26 | .26 |
| Day * Length | | | | | | | | | | | | | 0.01 | 0.01 | 0.22 | .64 |
| Day * Frequency 2 | | | | | | | | | | | | | −0.01 | 0.01 | 0.30 | .58 |
| Day * Frequency 3 | | | | | | | | | | | | | −0.03 | 0.01 | 4.52 | **.034** |
| *Random effects* | $\sigma^2$ | | | | $\sigma^2$ | | | | $\sigma^2$ | | | | $\sigma^2$ | | | |
| Id: Day number | 0.20 | | | | 0.20 | | | | 0.14 | | | | 0.14 | | | |
| Id: Beep block | 0.04 | | | | 0.04 | | | | 0.04 | | | | 0.04 | | | |
| Id | 0.63 | | | | 0.64 | | | | 0.64 | | | | 0.64 | | | |
| Day number | | | | | | | | | 0.04 | | | | 0.04 | | | |
| Residual | 1.04 | | | | 1.04 | | | | 1.04 | | | | 1.04 | | | |

*Note.* $N = 156$; Number of observations = 10,633. Statistically significant ($\alpha = .05$). *p* Values are marked in bold. Frequencies 2 and 3 stand for six and nine beeps, respectively. *SE* = standard error.
[a]Reference = short questionnaire. [b]Reference = three beeps.

**Table 3.** Results of ANOVAs for Retrospective Outcomes.

| | Length | | Frequency | | Length* Frequency | | |
|---|---|---|---|---|---|---|---|
| Outcome | $F(df_{Length}, df_{Error})$ | $p$ | $F(df_{Frequency}, df_{Error})$ | $p$ | $F(df_{Length * Frequency}, df_{Error})$ | $p$ | Significant group differences |
| Retrospective burden | 12.90 (1, 146) | **<.001** | 1.76 (2, 146) | .18 | 0.25 (2, 144) | .78 | Short < long |
| Retrospective ease of use | 1.04 (1, 147) | .31 | 0.18 (2, 147) | .83 | 1.12 (2, 145) | .33 | |
| Retrospective instructions | 0.02 (1, 148) | .89 | 2.60 (2, 148) | .08 | 3.57 (2, 146) | **.031** | Six beeps * long > three beeps * long<br>Six beeps * long > nine beeps * long |
| Retrospective reward | 0.78 (1, 137) | .38 | 24.29 (2, 137) | **<.001** | 1.38 (2, 135) | .26 | Three beeps < six beeps < nine beeps |
| Retrospective careless responding | 0.08 (1, 148) | .78 | 1.04 (2, 148) | .36 | 1.87 (2, 146) | .16 | |

*Note*. See Table 1 for the group means. Statistically significant (= .05) $p$ values are marked in bold. ANOVA = analysis of variance.

the opposite pattern was found, with participants receiving the long questionnaire reporting higher satisfaction with the instructions (see Table 1 for the means per group). The reward that was considered appropriate differed in expected ways with participants in high-sampling frequency conditions reporting a higher reward, $F(2, 137) = 24.29$, $p < .001$. Post hoc tests indicated that all group differences were significant, 3 versus 6: $F(1, 137) = 7.41$, $p = 0.007$; 3 versus 9: $F(1, 137) = 47.50$, $p < .001$; 6 versus 9: $F(1,137) = 18.29$, $p < .001$. Participants in the three beep group indicated on average that a reward of 48.61 euro ($SD = 12.66$) would be appropriate, as opposed to 57.45 ($SD = 13.30$) and 71.46 ($SD = 20.53$) in the six- and nine-beep groups, respectively. No significant effect of questionnaire length was observable, indicating that participants receiving the longer questionnaire did not indicate a higher appropriate reward than participants in the short questionnaire conditions, $F(1, 137) = 0.78$, $p = .38$. There was also no significant interaction effect between questionnaire length and sampling frequency, $F(2, 135) = 1.38$, $p = .25$. The distribution of responses per group are reported in Table 1, the results of all analyses of retrospective outcomes are reported in Table 3.

## Compliance

We found that, in line with our expectations, compliance was significantly lower in the long questionnaire conditions, $\chi^2(1) = 6.06$, $p = .014$. Based on the log odds ratios reported in Table 4, predicted probabilities of compliance can be calculated for different combinations of gender, frequency, and length. For example, the estimated probability of compliance for a female participant in the short questionnaire three beep group was 89% (= exp[2.05] / [1 + exp(2.05)]). The estimated probability of compliance of a female participant with three beeps but the long questionnaire was lower, namely 84% (= exp[2.05 − 0.36] / [1 + exp(2.05 − 0.36)]). Contrary to our expectations, no significant effect of sampling frequency could be detected, $\chi^2(2) = 1.07$, $p = .59$, and there was no significant interaction between length and frequency, $\chi^2(2) = 1.98$, $p = .37$. Compliance was found to decrease over time, $\chi^2(1) = 40.30$, $p < .001$. To illustrate, the estimated probability of compliance of a female participant receiving the short questionnaire three times per day was 91% (= exp[2.34 − 0 * 0.05] / [1 + exp(2.34 − 0 * 0.05)]) on Day 1 of the study which dropped to 84% (= exp[2.34 − 13 * 0.05] / [1 + exp(2.34 − 13 * 0.05)]) on the 14th day. The effects of length, $\chi^2(1) = 1.02$, $p = .31$, and frequency did not change over time, $\chi^2(2) = 0.42$, $p = .81$. Results can be found in Table 4. As noted in the method section, all questionnaires that were not completed until the last item were counted as noncompliant. The average prevalence of partially completed questionnaires (i.e., when a participant responded to the first item, but not to the last item) varied from 2% to 5% between the experimental groups (see Table 1).

## Reported Momentary Careless Responding

Sampling frequency did not significantly predict momentary reported careless responding, $\chi^2(2) = 0.91$, $p = .64$). However, in line with our hypotheses, participants receiving the long questionnaire reported more careless responding than participants receiving the short version, $\chi^2(1) = 6.23$, $p = .013$). Irrespective of sampling frequency, the 60-item questionnaire was associated with a decrease of 0.24 (on a 7-point scale) in attention paid to the questions compared with the 30-item questionnaire. There was also no significant interaction between sampling frequency and questionnaire length in predicting careless responding, $\chi^2(2) = 1.48$, $p = .48$. There was no main effect of time, indicating that participants did not report higher careless responding over time, $\chi^2(1) = 2.42$, $p = .12$. We did

**Table 4.** Compliance.

| | Model 1 | | | | Model 2 | | | | Model 3 | | | | Model 4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predictor | Log *OR* | *SE* | $\chi^2(1)$ | *p* | Log *OR* | *SE* | $\chi^2(1)$ | *p* | Log *OR* | *SE* | $\chi^2(1)$ | *p* | Log *OR* | *SE* | $\chi^2(1)$ | *p* |
| *Fixed effects* | | | | | | | | | | | | | | | | |
| Intercept | 2.05 | 0.16 | 162.46 | **<.001** | 1.94 | 0.19 | 99.78 | **<.001** | 2.34 | 0.17 | 188.07 | **<.001** | 2.26 | 0.20 | 132.25 | **<.001** |
| Length[a] | −0.36 | 0.15 | 6.06 | **.014** | −0.13 | 0.26 | 0.27 | .61 | −0.35 | 0.15 | 5.76 | **.016** | −0.26 | 0.17 | 2.42 | .12 |
| Frequency 2[b] | 0.13 | 0.18 | 0.52 | .47 | 0.20 | 0.25 | 0.66 | .42 | 0.14 | 0.18 | 0.58 | .44 | 0.22 | 0.22 | 0.97 | .33 |
| Frequency 3[b] | −0.04 | 0.18 | 0.06 | .81 | 0.20 | 0.25 | 0.67 | .41 | −0.04 | 0.18 | 0.05 | .82 | −0.01 | 0.21 | 0.00 | .97 |
| Gender[c] | −0.45 | 0.18 | 6.27 | **.012** | −0.43 | 0.18 | 5.87 | **.015** | −0.44 | 0.18 | 5.96 | **.015** | −0.44 | 0.18 | 5.97 | **.015** |
| Length * Frequency 2 | | | | | −0.15 | 0.36 | 0.17 | .68 | | | | | | | | |
| Length * Frequency 3 | | | | | −0.47 | 0.35 | 1.85 | .17 | | | | | | | | |
| Day number | | | | | | | | | −0.05 | 0.01 | 40.30 | **<.001** | −0.03 | 0.02 | 3.51 | **.06** |
| Day * Length | | | | | | | | | | | | | −0.01 | 0.01 | 1.02 | .31 |
| Day * Frequency 2 | | | | | | | | | | | | | −0.01 | 0.02 | 0.39 | .53 |
| Day * Frequency 3 | | | | | | | | | | | | | −0.01 | 0.02 | 0.08 | .78 |
| Random effects | $\sigma^2$ | | | | $\sigma^2$ | | | | $\sigma^2$ | | | | $\sigma^2$ | | | |
| Id: Day number | 0.28 | | | | 0.28 | | | | 0.25 | | | | 0.25 | | | |
| Id: Beep block | 0.14 | | | | 0.14 | | | | 0.14 | | | | 0.14 | | | |
| Id | 0.64 | | | | 0.63 | | | | 0.61 | | | | 0.61 | | | |
| Day number | | | | | | | | | 0.00 | | | | 0.00 | | | |

*Note*. $N = 156$; Number of observations = 12,991. Statistically significant ($\alpha = .05$). p Values are marked in bold. Frequencies 2 and 3 stand for six and nine beeps, respectively. *SE* = standard error; *OR* = odds ratio.

[a]Reference = short questionnaire. [b]Reference = three beeps. [c]Reference = female.

however observe a significant interaction between sampling frequency and time, $\chi^2(2) = 7.09$, $p = .029$. Post hoc tests indicated that the effect of receiving nine beeps as opposed to three beeps became significantly less negative over time, $\chi^2(1) = 6.27$, $p = .012$). In other words, on Day 0, the nine-beep group was associated with an expected decrease of 0.14 in attention paid to the questions compared with the three-beep group, but this difference decreased with 0.03 every day in the study. After 5 days in the study, the nine-beep group was associated with paying more attention compared with the three-beep group. The effect of receiving a longer questionnaire did not change over time, $\chi^2(1) = 1.43$, $p = .23$. For more details, see Table 5.

### *Objective Momentary Careless Responding*

Ten participants failed the directed response item at least once. Of these 10 participants, 4 provided a wrong response more than once, suggesting pervasive careless responding (two had received the long questionnaire nine times per day, one the short questionnaire nine times per day, and the other one had received the short questionnaire six times per day). There were no significant group differences in objective careless responding. Specifically, careless responding did not depend on sampling frequency, $\chi^2(2) = .57$, $p = .75$, or questionnaire length, $\chi^2(1) = 0.00$, $p = .96$. However, it needs to be noted that the prevalence of objective careless responding was so low that we had low power to detect group differences. Results can be found in Table 6. In the model with the interaction between length and frequency, the absence of objective careless responding in some groups led to perfect separation. As a result, this model failed to converge and its parameters are therefore not reported.

### *Retrospective Careless Responding*

The prevalence of retrospectively reported careless responding did not differ between different experimental groups, as indicated by the nonsignificant main effects of frequency, length, and their interaction, $F(2, 148) = 1.04$, $p = .36$; $F(1, 148) = .08$, $p = .78$; $F(2, 146) = 1.87$, $p = .16$. The distribution of responses per group are reported in Table 1.

### *Reasons for Missing Beeps*

Out of 2,366 unanswered beeps, a reason for missing the beep was indicated at only 273 of the subsequent beeps. In 70% of these cases, participants indicated not being able to respond. The options of not wanting to respond, being too stressed to respond, perceiving the beep as burdensome, and not responding for another reason were endorsed less frequently (in 5%, 7%, 14%, and 21% of the cases, respectively). The distribution of reasons was similar in the different groups and can be found in Table 7.

## Discussion

The aim of the current study was to investigate the effects of sampling frequency and questionnaire length on perceived burden and measures of data quality and quantity. Our main finding was that a longer ESM questionnaire was associated with more momentary and retrospective perceived burden, lower compliance, and higher momentary reported careless responding compared with a shorter questionnaire, whereas sampling frequency was not associated with differences in the outcomes. To our knowledge, this is the first experimental study that isolated the effect of questionnaire length on burden and different aspects of data quality and quantity in ESM. Questionnaires in ESM studies show a wide variation in the number of items (Morren et al., 2009; Ono et al., 2019; Vachon et al., 2019). The current results suggest that these variations in questionnaire lengths can have a measurable impact on burden, data quality, and quantity. Most previous analyses have failed to find an association between questionnaire length and compliance (Jones et al., 2018; Ono et al., 2019; Soyster et al., 2019; Vachon et al., 2019), however, some of these findings were limited by the availability of information on full questionnaire lengths. Moreover, previous nonexperimental findings were limited by the fact that researchers typically adapt design characteristics to each other (e.g., longer questionnaire combined with lower frequency). While these analyses did control for the effects of some other design characteristics of the included studies, more subtle aspects (such as the detail of instructions, which might be adapted to the protocol intensity in individual studies) were not included. The current results indicate that when all other study parameters are kept constant, increasing the questionnaire length can have detrimental effects on the collected data. Interestingly, only one of three measures of careless responding was found to be affected by questionnaire length. The prevalence of objective careless responding was low, with 19 out of 528 cases. The power to detect group differences in this analysis was therefore low, which could explain the absence of a significant effect. It is also possible that the three measures tap into different underlying constructs. In line with this, previous research has found only low to moderate correlations between objective and self-report measures of careless responding (Meade & Craig, 2012). Compared with the retrospective measure of reported careless responding, the momentary measure allowed reducing measurement error through repeated assessments and the assessment in the moment limited distortions by recall bias. This may also explain the different results obtained with these two self-report measures.

Contrary to our hypotheses, no consistent relationships between sampling frequency and measures of burden, data quantity, and quality could be detected. Our study thereby replicated previous findings indicating the absence of an

**Table 5.** Momentary Reported Careless Responding.

| | Model 1 | | | | Model 2 | | | | Model 3 | | | | Model 4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Predictor | β | *SE* | $\chi^2(1)$ | *p* | β | *SE* | $\chi^2(1)$ | *p* | β | *SE* | $\chi^2(1)$ | *p* | β | *SE* | $\chi^2(1)$ | *p* |
| Fixed effects | | | | | | | | | | | | | | | | |
| Intercept | 6.03 | 0.10 | 3852.6 | **<.001** | 6.03 | 0.12 | 2572.2 | **<.001** | 6.00 | 0.10 | 3619.4 | **<.001** | 6.04 | 0.11 | 3115.1 | **<.001** |
| Length[a] | −0.24 | 0.10 | 6.23 | **.013** | −0.23 | 0.17 | 1.89 | .17 | −0.23 | 0.10 | 5.73 | **.017** | −0.18 | 0.11 | 2.87 | .09 |
| Frequency 2[b] | −0.08 | 0.12 | 0.42 | .52 | −0.00 | 0.17 | 0.00 | .98 | −0.08 | 0.12 | 0.48 | .49 | −0.12 | 0.13 | 0.85 | .36 |
| Frequency 3[b] | 0.03 | 0.12 | 0.08 | .78 | −0.03 | 0.17 | 0.04 | .84 | −0.00 | 0.12 | 0.00 | .97 | −0.14 | 0.13 | 1.17 | .28 |
| Length * Frequency 2 | | | | | −0.17 | 0.24 | 0.47 | .49 | | | | | | | | |
| Length * Frequency 3 | | | | | 0.12 | 0.24 | 0.28 | .60 | | | | | | | | |
| Day number | | | | | | | | | 0.01 | 0.00 | 2.42 | 0.12 | 0.00 | 0.01 | 0.01 | .94 |
| Day * Length | | | | | | | | | | | | | −0.01 | 0.01 | 1.43 | .23 |
| Day * Frequency 2 | | | | | | | | | | | | | 0.01 | 0.01 | 0.44 | .51 |
| Day * Frequency 3 | | | | | | | | | | | | | 0.03 | 0.01 | 6.27 | **.012** |
| Random effects | $\sigma^2$ | | | | $\sigma^2$ | | | | $\sigma^2$ | | | | $\sigma^2$ | | | |
| Id: Day number | 0.09 | | | | 0.09 | | | | 0.06 | | | | 0.06 | | | |
| Id: Beep block | 0.00 | | | | 0.00 | | | | 0.00 | | | | 0.00 | | | |
| Id | 0.35 | | | | 0.35 | | | | 0.38 | | | | 0.38 | | | |
| Day number | | | | | | | | | 0.03 | | | | 0.02 | | | |
| Residual | 0.73 | | | | 0.73 | | | | 0.73 | | | | 0.73 | | | |

*Note*. $N = 156$; Number of observations = 10,627. Statistically significant ($\alpha = .05$). *p* Values are marked in bold. Frequencies 2 and 3 stand for six and nine beeps, respectively. *SE* = standard error.
[a]Reference = short questionnaire. [b]Reference = three beeps.

**Table 6.** Objective Careless Responding.

| | Model 1 | | | | Model 2: NA | | | |
|---|---|---|---|---|---|---|---|---|
| Predictor | Log *OR* | *SE* | $\chi^2(1)$ | *p* | Log *OR* | *SE* | $\chi^2(1)$ | *p* |
| *Fixed effects* | | | | | | | | |
| Intercept | 11.33 | 3.35 | 11.43 | **<.001** | | | | |
| Length[a] | −0.09 | 1.91 | 0.00 | .96 | | | | |
| Frequency 2[b] | −1.04 | 3.51 | 0.09 | .77 | | | | |
| Frequency 3[b] | −2.11 | 3.17 | 0.44 | .51 | | | | |
| Length * Frequency 2 | | | | | | | | |
| Length * Frequency 3 | | | | | | | | |
| *Random effects* | $\sigma^2$ | | | | | | | |
| Id | 94.71 | | | | | | | |

*Note.* $N = 156$; Number of observations = 528. Model 2 failed to converge. Statistically significant ($\alpha = .05$). *p* Values are marked in bold. Frequencies 2 and 3 stand for six and nine beeps, respectively. *SE* = standard error; *OR* = odds ratio.
[a]Reference = short questionnaire. [b]Reference = three beeps.

**Table 7.** Reasons for Missing Beeps per Experimental Condition.

| Questionnaire length | Short | | | Long | | |
|---|---|---|---|---|---|---|
| Frequency | Three beeps | Six beeps | Nine beeps | Three beeps | Six beeps | Nine beeps |
| Number of observations | 27 | 41 | 39 | 16 | 62 | 88 |
| I could not respond (on time) | 0.78 | 0.59 | 0.64 | 0.88 | 0.65 | 0.77 |
| I did not feel like responding | 0 | 0 | 0.03 | 0.06 | 0.11 | 0.06 |
| I was too stressed to respond | 0.04 | 0.07 | 0.08 | 0.06 | 0.13 | 0.02 |
| I felt disturbed by the beep | 0 | 0.27 | 0.05 | 0.06 | 0.18 | 0.14 |
| Other reason | 0.30 | 0.27 | 0.33 | 0.06 | 0.15 | 0.17 |

effect of sampling frequency on compliance (Conner & Reid, 2012; Jones et al., 2018; McCarthy et al., 2015; Morren et al., 2009; Ono et al., 2019; Soyster et al., 2019; Stone et al., 2003; Walsh & Brinker, 2016). In the short questionnaire conditions, compliance was higher with high-sampling frequencies. Although these differences in compliance did not reach statistical significance, this pattern is in line with a previous study that found lower compliance to be associated with longer inter beep intervals (Sokolovsky et al., 2014). In contrast with the study by Stone et al. (2003), we did not detect differences in perceived burden between the different sampling frequency groups. Interestingly, momentary burden was the highest in the three-beep conditions. This could indicate that frequent beeps might even be perceived as less burdensome in some cases, possibly because the higher predictability allows participants to better integrate the measures into their daily routine (as has previously been suggested in Janssens et al., 2018). This strongly contrasts with previous studies, where participants expressed a preference for lower sampling frequencies (Rosenkranz et al, 2020; Spook et al., 2013). It is possible that the effect of sampling frequency was canceled out by the increased motivation due to the higher incentive. However, the payment in our study was adapted to the time investment for participants, which is typically required by ethical committees. It is therefore usually not possible to manipulate sampling frequency without simultaneously adapting the payment. It is also possible that participants were not disturbed by the high-sampling frequencies because they are used to technology which relies on frequent short interruptions in the form of notifications. This is supported by high rates of social networking app use reported in young adults (Perrin & Anderson, 2019).

The effect of sampling frequency was also not found to change depending on questionnaire length. All previous experimental studies examining different sampling frequencies employed relatively short questionnaires (3 to 28 items), which left the possibility that sampling frequency would affect data quality and quantity when combined with longer questionnaires, as has previously been suggested (Intille et al., 2016; Walsh & Brinker, 2016; Piasecki et al., 2007). While our results do not support this hypothesis, it needs to be noted that our study had low power for tests of interactions. Although no clear effects of frequency were apparent, the two dropouts had received the most intensive ESM protocol. Additionally, compliance was lowest in this condition and below the 80% threshold, which has previously been suggested as a minimum mark for representativeness of data (Jones et al., 2018). Two of the four consistent careless

responders were also in the nine-beep, long questionnaire condition. Taken together, there is some indication that the combination of a high-sampling frequency with a long questionnaire may be detrimental for at least part of the participants, while no such signs were apparent when the short questionnaire was administered nine times per day.

The effects of frequency and length were not found to change over time. However, the observed decrease in compliance over time across all groups is in line with previous work (Forkmann et al., 2018; Ono et al., 2019; Rintala et al., 2018; Silvia et al., 2013). Surprisingly, the included measures of burden and momentary reported careless responding were not changing over the time of the study. Similar results were found in a previous study, where decreases in compliance and variances over time were not associated with changes in associations between variables, suggesting that quantity, but not quality of data declines over time (Fuller-Tyszkiewicz et al., 2013). We expected that as the study progressed, participants would experience the protocol as increasingly burdensome. In follow-up interviews, some participants indeed mentioned a calibration period of several days after which their experience of taking part in the study stabilized. Interestingly, some participants reported an increase in burden after this initial period (i.e., the first excitement wore off and questions became boring), while others reported a decrease (i.e., they got used to the beeps). These heterogeneous experiences might explain why no overall change in burden could be detected over time. In line with this assumption, a model including a random effect of day number fitted significantly better than a model without this term, indicating significant heterogeneity in the changes in momentary burden over time, $\chi^2(2) = 103.51, p < .001$.

The prevalence of careless responding was low, both when relying on participants' reports and on an objective test. We found that the median response to the momentary subjective careless responding item was 6 out of 7 (with 7 indicating paying full attention) and 1 or below out of 6 (with 6 indicating always responding carelessly) for retrospective careless responding in all the groups. Only 3.6% of all responses to the objective test were wrong, which is at the lower end of base rates of careless responding found in cross-sectional studies (Meade & Craig, 2012). However, we did detect that four participants repeatedly failed these basic attention checks, even though participants were warned about data quality checks during the briefing. Previous research has indicated that a small part of participants may misunderstand such objective measures of careless responding and that a wrong answer on such an item may not always indicate actual carelessness (Kim et al., 2018). Indeed, several participants in the current study noted during follow-up interviews that they were confused/surprised by the objective measure of careless responding that had not been introduced during the briefing session. Nevertheless, all these participants reported selecting the right response, which offers some support that the item functioned as intended in the current study. The item was only included four times per participant, and it is possible that the prevalence of careless responding was underestimated. However, more frequent presentations of such an instructed response item may bring their own problems, as participants may start looking out for these items. Another consideration is that attention checks might make participants aware that accuracy is monitored and therefore increase data quality. It might therefore be advisable to include data quality checks in future studies, despite the relatively low prevalence detected in the current sample.

Even though the long questionnaire was perceived as more burdensome than the shorter version, participants did not indicate a higher appropriate reward. Participants seemed to base their ratings on the reward they actually received, which was adapted to the sampling frequency. The average rewards indicated by participants in the six- and nine-beep conditions were lower than what they received, indicating that participants might have been overpaid.

This study was the first to our knowledge to test an ESM item about reasons for missing beeps. A first interesting finding was that the use of the item was relatively infrequent, indicating that either most beeps were missed without the participant noticing, or that participants typically missed several beeps in a row. Both cases were described by participants during follow-up interviews. Another possibility is that participants were not fully honest when answering whether they had heard the missed beep. The prevalence of partially completed questionnaires may also partly explain the discrepancy between the number of noncompliant beeps and the number of times reasons for not responding were given. Since in cases of partial compliance participants had started the questionnaire, it seems unlikely that they would indicate not having responded to these beeps. When looking at the reasons, we found that participants mostly felt that they were not able to answer (on time). Few participants indicated not wanting to respond. However, it is possible that these results were influenced by social desirability and that actual prevalence of not wanting to respond was higher. A future study could assess the activity at the time of missing the beep, which could further inform about whether noncompliance is systematic. It may also be helpful to include a function in the app that informs participants about missed beeps and then inquires about the circumstances.

### *Limitations*

Several limitations need to be considered when interpreting the current findings. First, our aim was to include the same topics in the short and long questionnaire. However, this led

to more repetitiveness in the long questionnaire, since individual constructs were assessed with more items than in the short version (e.g., positive affect was assessed with four items in the short questionnaire and nine items in the long questionnaire). This might have amplified the effect of questionnaire length, since the questionnaire was not only longer but also more repetitive. More variation in item topics or presentation may reduce the effect of questionnaire length. Another limitation is the composition of our sample, which was restricted to young students and it is unclear to what extent the results will generalize to other populations. It is for instance possible that the influence of the sampling protocol on burden and related changes in data quality and quantity is stronger in some clinical populations. However, since many ESM studies are conducted in comparable populations, we believe that our findings can be applied to these groups. Yet the derived guidelines cannot replace the evaluation of measurements for every new study, as we expect burden and data quantity and quality to largely depend on the specific sample. Additionally, our study was powered for the main effects of frequency and length on compliance and power was lower for interactions. It also needs to be noted that the internal consistency of the instruction subscale was relatively low ($\omega = .32$), suggesting that the items were assessing distinct aspects related to the instructions and were only weakly correlated. Finally, previous studies have detected signs of increased measurement reactivity with high-sampling frequencies (Conner & Reid, 2012; McCarthy et al., 2015). This is, next to noncompliance and careless responding, another way by which ESM data could be biased that we did not examine in the current analysis.

### *Recommendations and Future Research*

Based on the current findings, we recommend that future ESM studies should aim to limit the number of items in ESM questionnaires. Thirty items achieved high compliance even with high-sampling rates, while 60 items led to a significant increase in burden and decrease in data quality and quantity. It is not possible to tell, based on the current findings, after which number of items these changes occur exactly. Given the absence of consistent effects of high-sampling frequencies on the outcomes, high-sampling frequencies seem preferable, since they offer more fine grained information. This is especially relevant for researchers planning to conduct lagged analyses, where effects might not be detectable when the temporal resolution of the ESM data is too low. Researchers could also consider the use of planned missing-data designs to shorten their questionnaires (Silvia et al., 2013). Another interesting avenue for future investigations is to look at alternative indices of careless responding or other aspects of data quality (e.g., within-person variance [Fuller-Tyszkiewicz et al., 2013; Vachon et al., 2016] and response times [McCabe et al., 2012; van Berkel et al., 2019]). In the current study, we manipulated two central design aspects of ESM studies. There are, however, many other design choices that could possibly affect burden, data quality, and quantity (e.g., the order of questions, the duration of the study, the briefing). Investigating the effects of even shorter questionnaires and other sampling frequencies than the ones included in the current study would also be an interesting avenue for future research. More systematic investigations into design characteristics influencing data quality and quantity in ESM studies are needed. It also remains to be investigated how different rewards affect participation (both who originally signs up and who completes the study), perceived burden, and data quality.

## Conclusion

The current study offered support for increased perceived burden, lower compliance, and increases in reported careless responding with longer ESM questionnaires. Sampling frequency did not have a consistent effect on burden and the included measures of data quantity and quality. The current results suggest that questionnaire length is an important design parameter to consider when setting up an ESM study.

### Acknowledgments

We would like to thank Mariam Chichua and Tessa Biesemans for their help during the data collection, colleagues for their help during the design and set up of the study, and finally, the participants without whom this study would not have been possible.

### Declaration of Conflicting Interests

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

### Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article : This research was funded by an Odysseus grant (Grant GOF8416N) allocated to Inez Myin-Germeys by Fonds voor Wetenschappelijk Onderzoek (FWO).

### ORCID iD

Gudrun Eisele https://orcid.org/0000-0002-4466-3733

### Supplemental Material

Supplemental material for this article is available online.

### Notes

1. Only measures used in the current investigation are discussed. For an overview of the full questionnaire battery, we refer the reader to the OSF page of this project (https://osf.io/pzx8r/?view_only=7afaed46a9b24ebcbf4e8947644015e8).

2. Excluding the outlier (the one participant who failed to meet the minimum compliance requirement) resulted in a nonsignificant difference between the six and the three beep groups.

## References


Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2019). Routinely randomize the display and order of items to estimate and adjust for biases in subjective reports. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000294

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Collins, L. M., & Graham, J. W. (2002). The effect of the timing and spacing of observations in longitudinal studies of tobacco and other drug use: Temporal design considerations. *Drug and Alcohol Dependence*, *68*(Suppl. 1), 85-96. https://doi.org/10.1016/S0376-8716(02)00217-X

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

Dowle, M., & Srinivasan, A. (2019). *data.table: Extension of 'data.frame'* (1.12.2) [Computer software]. https://CRAN.R-project.org/package=data.table

Ebner-Priemer, U. W., & Sawitzki, G. (2007). Ambulatory assessment of affective instability in borderline personality disorder. *European Journal of Psychological Assessment*, *23*(4), 238-247. https://doi.org/10.1027/1015-5759.23.4.238

Edwards, C. J., Cella, M., Tarrier, N., & Wykes, T. (2016). The optimisation of experience sampling protocols in people with schizophrenia. *Psychiatry Research*, *244*(October), 289-293. https://doi.org/10.1016/j.psychres.2016.07.048

Forkmann, T., Spangenberg, L., Hallensleben, N., Hegerl, U., & Kersting, A. (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

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

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

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. http://dx.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

Intille, S. S., Haynes, C., Maniar, D., Ponnada, A., & Manjourides, J. (2016). µEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *UbiComp '16: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 1124-1135). ACM. https://doi.org/10.1145/2971648.2971717

Janssens, K. A. M., Bos, E. H., Rosmalen, J. G. M., Wichers, M., & Riese, H. (2018). A qualitative approach to guide choices for diary design. *BMC Medical Research Methodology*, *18*(1), Article 140. https://doi.org/10.1186/s12874-018-0579-6

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Fred Wen, C. K., & Field, M. (2018). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction*, *114*(4), 609-619. https://doi.org/10.1111/add.14503

Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A., & King, K. M. (2018). Detecting random responders with infrequency scales using an error-balancing threshold. *Behavior Research Methods*, *50*(5), 1960-1970. https://doi.org/10.3758/s13428-017-0964-9

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. T. Reis (Ed.), *New directions for methodology of social and behavioral science* (pp. 41-56). Jossey-Bass.

May, M., Junghaenel, D. U., Ono, M., & Stone, A. A. (2018). Ecological momentary assessment methodology in chronic pain research: A systematic review. *Journal of Pain*, *19*(7), 699-716. https://doi.org/10.1016/j.jpain.2018.01.006

McCabe, K. O., Mack, L., & Fleeson, W. (2012). A guide to data cleaning in experience-sampling studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of Research methods for studying daily life* (pp. 321-338). Guilford Press.

McCarthy, D. E., Minami, H., Yeh, V. M., & Bold, K. (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

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., 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

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), Article e11398. https://doi.org/10.2196/11398

Perrin, A., & Anderson, M. (2019, April 10). *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018*. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

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

R Core Team. (2019). *R: A language and environment for statistical computing [Computer software]*. https://www.r-project.org/

Revelle, W. (2018). *Psych: Procedures for personality and psychological research [Computer software]*. Northwestern University. https://CRAN.R-project.org/package=psych

Reynolds, B. M., Robles, T. F., & Repetti, R. L. (2016). Measurement reactivity and fatigue effects in daily diary research with families. *Developmental Psychology*, *52*(3), 442-456. https://doi.org/10.1037/dev0000081

Rintala, A., Wampers, M., Lafit, G., Myin-Germeys, I., & Viechtbauer, W. (2020). *Perceived disturbance of the diary signal and predictors thereof in studies using the experience sampling method* [Manuscript in preparation].

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2018). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226-235. https://doi.org/10.1037/pas0000662

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101-1108. https://doi.org/10.1016/j.jval.2011.06.003

Rosenkranz, T., Takano, K., Watkins, E. R., & Ehring, T. (2020). Assessing repetitive negative thinking in daily life: Development of an ecological momentary assessment paradigm. *PLOS ONE*, *15*(4), Article e0231783. https://doi.org/10.1371/journal.pone.0231783

Santangelo, P. S., Ebner-Priemer, U. W., & Trull, T. J. (2013). Experience sampling methods in clinical psychology. In J. Comer, & P. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 188-210). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793549.013.0011

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1-32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shrout, P., & Lane, S. (2012). Psychometrics. In M. Mehl & T. Conner (Ed.), *Handbook of research methods for studying daily life* (pp. 302-320). Guilford Press.

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

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

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

Spook, J. E., Paulussen, T., Kok, G., & Van Empelen, P. (2013). Monitoring dietary intake and physical activity electronically: Feasibility, usability, and ecological validity of a mobile-based Ecological Momentary Assessment tool. *Journal of Medical Internet Research, 15*(9), e214. https://doi.org/10.2196/jmir.2617

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

Trull, T., & Ebner-Priemer, U. (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/10.31234/osf.io/eakyj

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*(November), 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: A systematic review and meta-analysis. *Journal of Medical Internet Research*, *21*(12), Article 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. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-51). ACM.

van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., & Kostakos, V. (2018). Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies*, *125*(May), 118-128. https://doi.org/10.1016/j.ijhcs.2018.12.002

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