A Mixed-Method Investigation Into Measurement Reactivity to the Experience Sampling Method: The Role of Sampling Protocol and Individual Characteristics

Psychological Assessment

A Mixed-Method Investigation Into Measurement Reactivity to the Experience Sampling Method: The Role of Sampling Protocol and Individual Characteristics

Gudrun Eisele, Hugo Vachon, Ginette Lafit, Daphne Tuyaerts, Marlies Houben, Peter Kuppens, Inez Myin-Germeys, and Wolfgang Viechtbauer

Online First Publication, September 29, 2022. http://dx.doi.org/10.1037/pas0001177

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

ISSN: 1040-3590

https://doi.org/10.1037/pas0001177

A Mixed-Method Investigation Into Measurement Reactivity to the Experience Sampling Method: The Role of Sampling Protocol and Individual Characteristics

Gudrun Eisele[1], Hugo Vachon[1], Ginette Lafit[1, 2], Daphne Tuyaerts[1], Marlies Houben[1], Peter Kuppens[2], Inez Myin-Germeys[1], and Wolfgang Viechtbauer[1, 3]

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

[2] Research Group for Quantitative Psychology and Individual Differences, KU Leuven

[3] Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University

Since the introduction of the experience sampling method (ESM), there have been concerns that the repeated assessments typically related to this method may alter the behavior, thoughts, or feelings of participants. Previous studies have offered mixed results with some studies reporting reactive changes, while others failed to find such effects. Our aim was to investigate under which circumstances ESM induces reactive effects. Students ($N = 151$) were randomly assigned to receive a questionnaire containing 30 or 60 items three, six, or nine times per day for 14 days. A random sample of 50 participants took part in qualitative interviews after the end of the data collection. We investigated changes over time in the data, while taking into account the sampling protocol and characteristics of participants, and analyzed qualitative reports of measurement reactivity. Decreases in completion time, within-person variance of ratings and subjective reports of habituation point toward the existence of a habituation period. While participants reported increases in emotional awareness in interviews, ESM measures indicated a decrease in emotional awareness over time. Changes in behavior were rare in quantitative and qualitative reports. Positive affect was decreasing over time in the ESM data, and various changes in affect, emotion regulation, and thoughts were reported in interviews. Individual characteristics and sampling protocol had inconsistent effects on changes over time. The results suggest that ESM induces changes in within-person variability, completion times, affect, or emotional awareness over time. Further research is needed to explore whether observed changes affect the validity of ESM data.

Public Significance Statement

Increasingly, researchers use frequently repeated self-report measures to assess individuals' experiences in the context of their daily lives. We find signs of changes in response behavior, affect, and emotional awareness that are triggered by these frequent assessments. The possibility of such reactive changes over time is important to consider when collecting repeated self-report measures in daily life.

Gudrun Eisele https://orcid.org/0000-0002-4466-3733
Hugo Vachon https://orcid.org/0000-0003-1259-649X
Ginette Lafit https://orcid.org/0000-0002-8227-128X
Daphne Tuyaerts https://orcid.org/0000-0002-8238-3329
Peter Kuppens https://orcid.org/0000-0002-2363-2356
Inez Myin-Germeys https://orcid.org/0000-0002-3731-4930
Wolfgang Viechtbauer https://orcid.org/0000-0003-3463-4063

This research was funded by an Odysseus grant (Grant GOF8416N) allocated to Inez Myin-Germeys by Fonds voor Wetenschappelijk Onderzoek. The authors are extremely grateful for the assistance of Tessa Biesemans and Mariam Chichua during the data collection; and Beau Reusens, Aleksandra Nowak, and Amine Zerrouk during the transcription of the interviews. Further, the authors would like to express our gratitude toward our colleagues for their contributions during the design and setup of the study. Finally, the authors would like to thank the participants.

Gudrun Eisele played a lead role in conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, and writing of original draft. Hugo Vachon played an equal role in conceptualization, methodology, supervision, and writing of review and editing. Ginette Lafit played an equal role in conceptualization, methodology, supervision, and writing of review and editing. Daphne Tuyaerts played a supporting role in writing of review and editing and an equal role in formal analysis. Marlies Houben played a supporting role in conceptualization and an equal role in writing of review and editing. Peter Kuppens played a supporting role in conceptualization and an equal role in writing of review and editing. Inez Myin-Germeys played a lead role in funding acquisition and an equal role in conceptualization and writing of review and editing. Wolfgang Viechtbauer played an equal role in conceptualization, methodology, supervision, and writing of review and editing.

Only measures used in the current article are described in detail. For a full overview of the questionnaires administered at baseline and follow-up assessment, the authors refer the reader to the Open Science Framework (OSF) webpage of the project (https://osf.io/pzx8r/). Data on which study conclusions are based are available from the authors upon request. The analysis code is available in the Supplemental Material. All data exclusions and manipulations are reported. The study was not preregistered. However, the reported analyses were preregistered on the OSF page.

Correspondence concerning this article should be addressed to Gudrun Eisele, Department of Neurosciences, Center for Contextual Psychiatry, KU Leuven, Kapucijnenvoer 33 bus 7001 (blok h), 3000 Leuven, Belgium. Email: gudrunvera.eisele@kuleuven.be

Keywords: ecological momentary assessment, ambulatory assessment, measurement reactivity, response behavior

Supplemental materials: https://doi.org/10.1037/pas0001177.supp

In recent years, the experience sampling method (ESM, also referred to as ecological momentary assessment; Larson & Csikszentmihalyi, 1983; Myin-Germeys et al., 2018) has made its way to the standard toolbox of the psychology researcher. In ESM studies, participants are asked to fill in multiple, short questionnaires per day, typically over several days. Concerns that the frequent assessments in ESM studies may induce changes in the behavior, feelings, or thoughts of participants have already been voiced in the first reports of the method (Larson & Csikszentmihalyi, 1983). Yet, almost 40 years later, the circumstances under which ESM induces measurement reactivity remain poorly understood.

Measurement Reactivity in ESM Research

Measurement reactivity has been defined as any change in the participant that is caused by the measurements (French & Sutton, 2010). Sometimes, it is additionally specified that these changes need to have a biasing effect on the data collected (e.g., Barta et al., 2012). This broad definition includes changes in the underlying construct, which may have long-lasting effects on the participant (e.g., a participant becomes more aware of their feelings), changes in the participant's behavior (e.g., a participant avoids certain activities during the study period), and changes in the participants' response behavior (e.g., a participant uses the response scale differently over time). Reactive changes are a possible problem in any psychological research. However, the intensive nature of ESM, which typically involves multiple assessments per day and over several days in participant's daily lives, has been suggested to be particularly prone to induce changes in response behavior or the underlying construct (Barta et al., 2012). Measurement reactivity could bias ESM findings that rely on measuring life as it is experienced, in other words, undermine the ecological validity of the assessment (Ram et al., 2017). Therefore, it is important to understand if measurement reactivity arises in ESM studies, under which circumstances it is more likely to arise, and to what extent it affects the validity of the data collected. In the following paragraphs, we will review previous studies that have offered inconclusive answers to those questions.

Mixed Evidence for Changes in the Underlying Construct and Behavior

In the absence of an intervention or an event that affects the whole sample, ESM or diary data are not expected to change systematically over time. If such changes are detected, they can therefore be interpreted as signs of measurement reactivity. Previous studies that observed systematic shifts in ESM or diary data over time have mostly interpreted them in terms of changes in the underlying construct. This interpretation has been based on theoretical accounts that suggest that the frequent reporting of internal states may increase the participants' self-awareness, induce rumination or other changes in emotion regulation, and subsequently lead to changes in their affective states (e.g., Conner & Reid, 2012; Johar & Sackett, 2018; Scollon et al., 2003). Indeed, individual studies have detected increases in emotional awareness (Kauer et al., 2012; Ludwigs et al., 2018) and the ability to differentiate emotions (Hoemann et al., 2021; Widdershoven et al., 2019) over the course of an ESM study. For changes in affect, reports have been mixed, with some studies detecting changes in individual affective variables (Ludwigs et al., 2018; Rowan et al., 2007; Zawadzki et al., 2019), while others have not (Aaron et al., 2005; Cruise et al., 1996; De Vuyst et al., 2019; Husky et al., 2010). It has also been suggested that taking part in an ESM study could lead to changes in the participants' behavior, which has been observed in some studies (Husky et al., 2010; Johnson et al., 2009), but not in others (e.g., Csikszentmihalyi & Larson, 1987). In addition, studies that asked participants to report on experienced changes over time have reported low to moderate levels of subjective reactivity (Ebner-Priemer et al., 2007; Palmier-Claus et al., 2012). In clinical samples, studies have further investigated systematic shifts in reported symptoms such as pain, substance use, suicidal thoughts, or depression levels. Generally, the majority of studies have not detected changes in mean levels of symptoms over time (Cruise et al., 1996; Law et al., 2015; Stone et al., 2003; von Baeyer, 1994), yet in some cases, symptoms were found to decrease (Broderick & Vikingstad, 2008; Kramer et al., 2014; Shiffman et al., 1997).

Detected Changes in Response Behavior

Aside from these changes in the mean level of variables, previous research has repeatedly detected a decrease in the variability of ESM responses over time (Csikszentmihalyi & Larson, 1987; Fuller-Tyszkiewicz et al., 2013; Vachon et al., 2016). This decrease in variability has not been interpreted as a change in the underlying construct (i.e., the state of participants is not thought to become less variable over time), but as a change in the way participants use the response scale. Two possible explanations have been suggested. One hypothesis is that the variability of responses decreases because of a habituation effect (also referred to as calibration). This means that by repeatedly using the scale, participants may become better at indicating how they feel, as they develop more stable conceptualizations of the different scale points. It could also mean that participants are overusing extremes in the beginning, but do that less over time, which could lead to increases in data quality over time. Alternatively, a decrease in motivation to provide high-quality responses may explain this pattern and lead to a more uniform response behavior over time (i.e., fatigue effect, also referred to as satisficing or boredom effect; Fuller-Tyszkiewicz et al., 2013). Specifically, participants may become increasingly annoyed by the assessments and consequently revert to heuristic ways of responding, which could manifest itself as more and more homogeneous responses over the duration of the study. In the case of a fatigue effect, such decreases in variance are expected to be accompanied by a weakening of the associations between variables, while associations between variables are not expected to weaken in the case of habituation. Studies that investigated changes in associations between variables have not detected them (Csikszentmihalyi & Larson, 1987; Fuller-Tyszkiewicz et al., 2013;

Johnson et al., 2009). However, other evidence does point toward decreases in motivation and data quality over time. For instance, one recent study on the reporting of social media use has found decreases in the convergent validity between reported and objective social media use over time (Verbeij et al., 2021), supporting a fatigue effect. While having similar impacts on the collected data at first sight, a habituation effect is not expected to undermine the validity of the collected data, while lower data quality related to a decrease in motivation over time could. Therefore, it is important to distinguish these two types of changes in response behavior.

Initial Elevation Bias

Shrout et al. (2018) introduced a new type of change over time to the intensive longitudinal data literature, the initial elevation bias. The initial elevation bias refers to situations in which the first data point is higher than subsequent measures, an observation that has repeatedly been made in non-ESM, longitudinal studies (e.g., Knowles et al., 1996; Patrick & Gilbert, 1998). The mechanism underlying this change has not been identified, meaning that it could be caused by both changes in the underlying construct, as well as by changes in response behavior. However, a recent study that investigated the initial elevation bias in diary data has not detected changes consistent with such an effect (Arslan et al., 2021).

The Role of Individual Characteristics, Study Design, and the Operationalization of Measurement Reactivity

Although a few studies have suggested different types of reactive effects in ESM or diary data, other studies have thus failed to observe any changes over time or detected inconsistent changes. It has been argued that the characteristics of the sampling protocol (e.g., how many assessments per day) and of the participants may lead to differential reactive effects (Barta et al., 2012; Conner & Reid, 2012; Hoemann et al., 2021; McCarthy et al., 2015; Stone et al., 2003) and thus possibly cause these diverging findings. This notion is supported by a study on the frequent reporting of happiness, where researchers detected decreases in happiness over time with more frequent reporting only for individuals high in neuroticism and depression, while other individuals showed increases in happiness with more frequent reporting (Conner & Reid, 2012). However, the sampling protocol and the individual characteristics of participants have typically not been considered when measurement reactivity was investigated in the past. Another factor that could contribute to the diverging findings is the way that reactivity has been operationalized. In the few studies that investigated reactive changes over time, researchers have mostly investigated linear changes in mean levels of variables (for exceptions, see Cruise et al., 1996; Zawadzki et al., 2019). However, this form may be inadequate, as it is possible that change manifests itself in an early stage of a study and flattens out over time (as suggested, e.g., by Paterson et al., 2020 and Shrout et al., 2018) or that it appears only after a longer period of ESM. Additionally, there is indication that the perception of participants and the changes that can be detected in the data do not always converge (Aaron et al., 2005; Litt et al., 1998). This underlines the need to assess both perceived and objectively measurable measurement reactivity.

The Present Study

In the present study, we aim to further our understanding of measurement reactivity to ESM assessments and address some of the gaps in the literature outlined before. We use the previously described broad definition of reactivity (French & Sutton, 2010). This means that any changes in participants' responses or experiences are treated as reactive effects in the present study and, in a first instance, we do not distinguish between possibly beneficial or biasing forms of reactivity. Specifically, we first investigate quantitative changes in mean levels of assessed variables over time. To gather information on the underlying mechanism of the reactive changes, we also investigate changes in the within-person variance of these ratings and in the associations between different variables over time. We focus on changes in affect, rumination, and emotional awareness, as these variables have been previously suggested to be particularly prone to reactive changes over time (Conner & Reid, 2012; Kauer et al., 2012; Widdershoven et al., 2019) and are frequently assessed in ESM studies. Subsequently, we investigate whether changes over time are moderated by study design factors (sampling frequency and questionnaire length) and/or individual characteristics of the participants (neuroticism and depression, based on Conner & Reid, 2012). Based on previous findings (Kauer et al., 2012; Widdershoven et al., 2019), we expect to observe increases in emotional awareness and clarity over time. Additionally, we expect that a higher sampling frequency and higher baseline levels of depression and neuroticism will be associated with more reactivity, that is, larger changes in mean, within-person variance, and associations between different variables over time (based on Conner & Reid, 2012). Finally, we analyze qualitative reports of measurement reactivity that were provided by a subsample of the participants during interviews at the end of the data collection. The conducted analyses were preregistered (https://osf.io/xdws2/?view_only=7407ee92cd994dbc961d726300795441; https://osf.io/r5w48/?view_only=ccdbce7ff60245fa8789643fb4b7f81f), and deviations from the preregistration are noted in the text.

Method

Sample

A sample of 163 students were recruited for the study. Students were required to be between 18 and 30 years old and to have never taken part in an ESM study before. The study was powered for hypotheses discussed in a previous article (Eisele et al., 2020). Power may be lower for the current analyses. In addition, the study is not powered for cross-level interactions, which should therefore be interpreted as exploratory. A random subsample of 51 participants were interviewed after the end of the data collection. The sample size for the subsample that took part in the interviews was determined based on practical considerations and considered to be sufficiently large to cover diverging experiences of participants during the study. The study was approved by the Social and Societal Ethics Committee of KU Leuven.

Procedure

Interested participants were invited to the lab. After providing informed consent, participants completed several baseline questionnaires, received instructions for the ESM period, and were

randomly assigned to receive an either 30 or 60 item-long questionnaire, three, six, or nine times per day for 14 days. The ESM assessments started on the day after the baseline session. Participants could take part in baseline sessions on Mondays, Tuesdays, Wednesdays, or Thursdays, and it was assured that an equal number of participants from each condition started on each day, to avoid that systematic differences between days would lead to spurious time trends in the ESM data. ESM questionnaires were delivered using the app MobileQ (Meers et al., 2020) on smartphones (Motorola DEFY+ model) that were lent to participants for the time of the study. ESM questionnaires were delivered at random times in fixed time windows that lasted from 9 a.m. to 10:30 p.m. Participants had 90 s to react to each ESM questionnaire. After the ESM period, participants returned to the lab to fill in several follow-up questionnaires. Directly after finishing the follow-up questionnaires, an approximately 10–15 min long semistructured interview was conducted with a randomly chosen subsample of participants. As a compensation, participants received vouchers of 40, 60, or 80 euros, depending on the sampling frequency they were assigned to.

Measures

Baseline Measures

Depression was assessed with the Center for Epidemiologic Studies Depression Scale (Radloff, 2012; validated in Dutch by van de Velde et al., 2011). Neuroticism was measured with the Neuroticism/Negative Emotionality subscale of the Big Five Inventory-2 (Soto & John, 2017; translation and validation in Dutch; Denissen et al., 2008). For both depression and neuroticism, a sum score of the items was calculated ($\omega = 0.92$ for depression; $\omega = 0.92$ for neuroticism).

ESM Measures

The full ESM questionnaire can be found in the Supplemental Materials. Questions were always presented in the same order. Answer options ranged from 1 to 7 on a 7-point Likert-type scale, unless otherwise stated. Positive affect was measured with four items ("Right now, I feel happy/relaxed/energetic/satisfied") of which we calculated a mean for every assessment moment ($\omega = 0.86$ within persons; $\omega = 0.97$ between persons). Negative affect was measured with four items ("Right now, I feel stressed/anxious/irritated/down") of which we also calculated a mean for every assessment moment ($\omega = 0.85$ within persons; $\omega = 0.96$ between persons). Momentary rumination was assessed with the item "I am ruminating," momentary emotional awareness with the item "At the moment, I am aware of my emotions," and momentary emotional clarity with the item "I found it difficult to indicate in a number how I am feeling." The item measuring momentary emotional clarity was developed for the purposes of this study. It was reviewed by other ESM researchers from the Center for Contextual Psychiatry at KU Leuven and pilot tested before the study. Event pleasantness was assessed with the item "Think of the most important event that happened since the last beep. This event was: Very unpleasant −3 −2 −1 0 1 2 3 very pleasant," which was only present in the long questionnaire version. Therefore, the analyses of the associations between event valence and affect could only be conducted in the long questionnaire group.[1] Behavioral reactivity was assessed with the item "I changed my daily routine because I was anticipating this beep." Completion time in seconds was calculated for each assessment moment, by adding up the time needed to fill in each of the nonbranched items that were common to the short and long questionnaire versions (see Footnote 1).

Interviews

Semistructured interviews were administered by the researchers who conducted the data collection (GE, a master student, and a research assistant). Some participants had been briefed by the same researcher who also interviewed them, but this was not always the case. All researchers conducted an interview together to assure an equal approach. Interview questions can be found in the Supplemental Materials and covered reactivity but also other methodological topics that were considered relevant based on a review of the literature (e.g., Beal, 2015). The interview questions were pilot tested with ESM experts and refined according to the resulting feedback. Interviews lasted on average 10–15 min, were recorded, and transcribed verbatim before analysis.

Analysis

Quantitative Analysis

Quantitative analyses were conducted in R (Version 4.1.1; R Core Team, 2021) with the packages rms (Harrell, 2021), nlme (Pinheiro et al., 2021), and car (Fox & Weisberg, 2019). Analyses consisted of three-level multilevel regression models with ESM assessments at Level 1, nested in days at Level 2, nested in persons at Level 3. Random intercepts were added at the person and day level, and a random slope for day was nested in persons. Separate models were run for each of the outcome variables. A cubic spline transformation was applied to the day variable to model nonlinear changes over time. For this transformation, knots were placed at Day 3, 6, and 9, since changes were a priori expected to be more likely to occur early during the data collection. However, it is important to note that the cubic spline allows flexible modeling of changes over time and therefore changes do not need to occur at the knot points (see Harrell, 2015). To test for changes in mean levels over time in the whole sample, day and the cubic spline transformed day variable were entered as predictors. The significance of both the original and transformed day variables was tested together with a Wald-type (chi-square) test, which allows testing the significance of composite hypotheses about fixed effects in mixed-effects models (Singer & Willett, 2003). To investigate linear changes in the within-person variance of outcomes over time, this model was extended by allowing the within-person variance of the outcome to change as a function of the day variable, that is, by fitting a location-scale model (Hedeker et al., 2012). The within-person variance of ratings for a person i at assessment moment j was modeled as

$$\sigma_{ij}^2 = \sigma^2 \exp(\beta_1 D_{ij}), \quad (1)$$

where σ^2 gives the estimated within-person variance on Day 0, D_{ij} stands for the day number for a participant i at assessment moment j, and β_1 indicates the multiplicative factor showing the change in

[1] This represents a deviation from the preregistration.

within-person variance per day. The resulting model was compared to a model excluding the heterogeneous within-person variance with a likelihood-ratio test. This corresponds to the standard procedure for assessing the significance of the variance components in a mixed-effects model (Singer & Willett, 2003). To investigate changes in relationships between variables, a model was fit with affect as outcome, and day, event valence, and their interaction as predictors. The interaction between day and event valence was tested for significance with a Wald-type test.

All models were then extended by separately including the possible moderator variables sampling frequency, questionnaire length, neuroticism, and depression, and their interactions with the day and cubic spline transformed day variables. Interactions between the respective moderator and the day and cubic spline transformed day variables were tested for significance together with a Wald-type test. Then, these models were further extended by letting the within-person variance depend on day, the moderator variable, and their interaction. To assess the significance of the interaction term, this model was compared to a model without the interaction term with a likelihood-ratio test. Finally, the models predicting affect based on event valence, day, and their interaction were extended by including each of the moderator variables separately. The moderator variables were also allowed to interact with event valence, day, and the interaction between event valence and day. The significance of the three-way interaction term was assessed with a Wald-type test. For analyses involving sampling frequency, significant omnibus tests were followed by testing the pairwise contrasts between all sampling frequencies individually. The adequacy of fitted models was tested by visually inspecting Q–Q plots of the residuals at each level (normality assumption) and plotting the residuals against the predictors (homoscedasticity assumption). Further, distributions of variables were visually inspected for univariate outliers. Baseline depression and neuroticism were centered around the sample mean to facilitate the interpretation of coefficients. Further, the day number variable was rescaled by dividing it by 100 to avoid overly small coefficients that led to convergence problems. The quantitative analyses were preregistered (https://osf.io/xdws2/?view_only=7407ee92cd994dbc961d726300795441), and the analysis code can be found in the Supplemental Materials. Deviations from the preregistration are marked with footnotes in the text.

Qualitative Analysis

The interviews were analyzed using NVivo (QSR International Pty Ltd, 2020). GE and DT independently familiarized themselves with the transcripts and assigned initial topic codes. They then independently reviewed the codes, organized them into broader themes and subthemes in a data-driven way, then reviewed the entire data to ensure that it was adequately covered. By discussing differences in themes and codes, the researchers then developed a refined coding scheme that was subsequently reviewed by HV, GL, and WV. Then, a second round of coding was conducted by GE. Finally, themes and codes were checked against all transcripts to ensure that the entirety of the data was adequately covered. The qualitative analysis was preregistered (https://osf.io/r5w48/?view_only=ccdbce7ff60245fa8789643fb4b7f81f), and deviations from the preregistration are marked with footnotes in the text.

Transparency and Openness

Only measures used in the current article are described in detail. For a full overview of the questionnaires administered at baseline and follow-up assessment, we refer the reader to the Open Science Framework (OSF) webpage of the project (https://osf.io/pzx8r/). Data on which study conclusions are based are available from the authors upon request. The analysis code is available in the Supplemental Materials. All data exclusions and manipulations are reported. The study was not preregistered. However, the reported analyses were preregistered on the OSF page (https://osf.io/xdws2/?view_only=7407ee92cd994dbc961d726300795441; https://osf.io/r5w48/?view_only=ccdbce7ff60245fa8789643fb4b7f81f).

Results

Sample Characteristics

A sample of 163 students was initially enrolled in the study. Three participants were excluded after the baseline session because they did not fulfill the inclusion criteria, two participants dropped out of the study, two participants received beeps at wrong times due to a technical problem, and one participant responded to less than one third of the scheduled beeps. These participants were therefore excluded from the current analyses. Further, four participants were identified as careless responders in a previous analysis of the data (Eisele et al., 2020) and were also excluded. After these exclusions, a sample of 151 participants remained for the quantitative analyses and a subsample of 50 participants for the qualitative analyses.[2] Additionally, four participants who experienced a technical problem that led to missing more than one full day of the ESM period were excluded from the quantitative analyses after the appearance of the technical problem. This led to the exclusion of 3–7 days from the analysis for these participants. All data exclusions were specified in the preregistration of the quantitative analyses. One participant who had received beeps at wrong times was excluded from the qualitative interviews. The mean age of the remaining sample was 21.73 years (SD = 1.78) and 79% of the sample was females.

Quantitative Analyses

Descriptive statistics of all variables are reported in Supplemental Table 1. The average compliance was 81%. Notably, participants reported only low levels of behavioral reactivity, which represents an interesting finding in itself. Due to the resulting skewed distribution of responses and model misfit, the behavioral reactivity item was excluded from the originally planned analyses of changes over time. The significance of the conducted tests can be found in Table 1, coefficients of all fitted models can be found in Supplemental Tables 2–15, and changes in mean and within-person variance of variables over days are further depicted in Figures 1 and 2.

[2] The exclusion of one participant who had received beeps at wrong times had not been specified in the preregistration for the qualitative analyses but only for the quantitative analyses.

Table 1
Results of the Quantitative Analyses

	Time effect in whole group			Interaction between sampling frequency and time effect			Interaction between questionnaire length and time effect			Interaction between depression and time effect			Interaction between neuroticism and time effect		
Outcome	Mean	Within-person variance	Association with event valence	Mean	Within-person variance	Association with event valence	Mean	Within-person variance		Mean	Within-person variance	Association with event valence	Mean	Within-person variance	Association with event valence
Test statistic	$\chi^2(df = 2)$	Z	$\chi^2(df = 1)$	$\chi^2(df = 4)$	Z	$\chi^2(df = 2)$	$\chi^2(df = 2)$	Z		$\chi^2(df = 2)$	Z	$\chi^2(df = 1)$	$\chi^2(df = 2)$	Z	$\chi^2(df = 1)$
Emotional awareness	16.184***	63.813***		4.342	13.119** (6 < 3; 6 < 9)		2.255	18.315***		0.309	0.001		0.3	0.001	
Clarity	0.701	21.979***		3.559	6.937* (6 < 9)		5.095	8.847**		3.06	0.578		3.439	0.047	
Rumination	0.422	108.244***		6.215	10.419** (3 > 6; 3 > 9)		2.169	0.007		2.734	10.749**		0.29	7.203**	
Completion time	210.852***	78.65***		11.104* (6 > 3; 6 > 9)	31.068*** (9 < 3; 9 < 6)		7.834*	19.362***		2.915	0.555		6.124*	16.175***	
Negative affect	0.074	108.237***	0.837	7.101	14.19*** (6 < 3; 6 < 9)	0.243	3.024	3.413		0.033	0.185	4.912*	0.26	1.673	4.785*
Positive affect	8.464*	159.306***	10.78**	1.997	2.644	1.506	5.339	2.919		0.189	2.52	0.514	1.637	0.603	3.248

*$p < .5$. **$p < .01$. ***$p < .001$.

Changes Over Time in the Whole Group

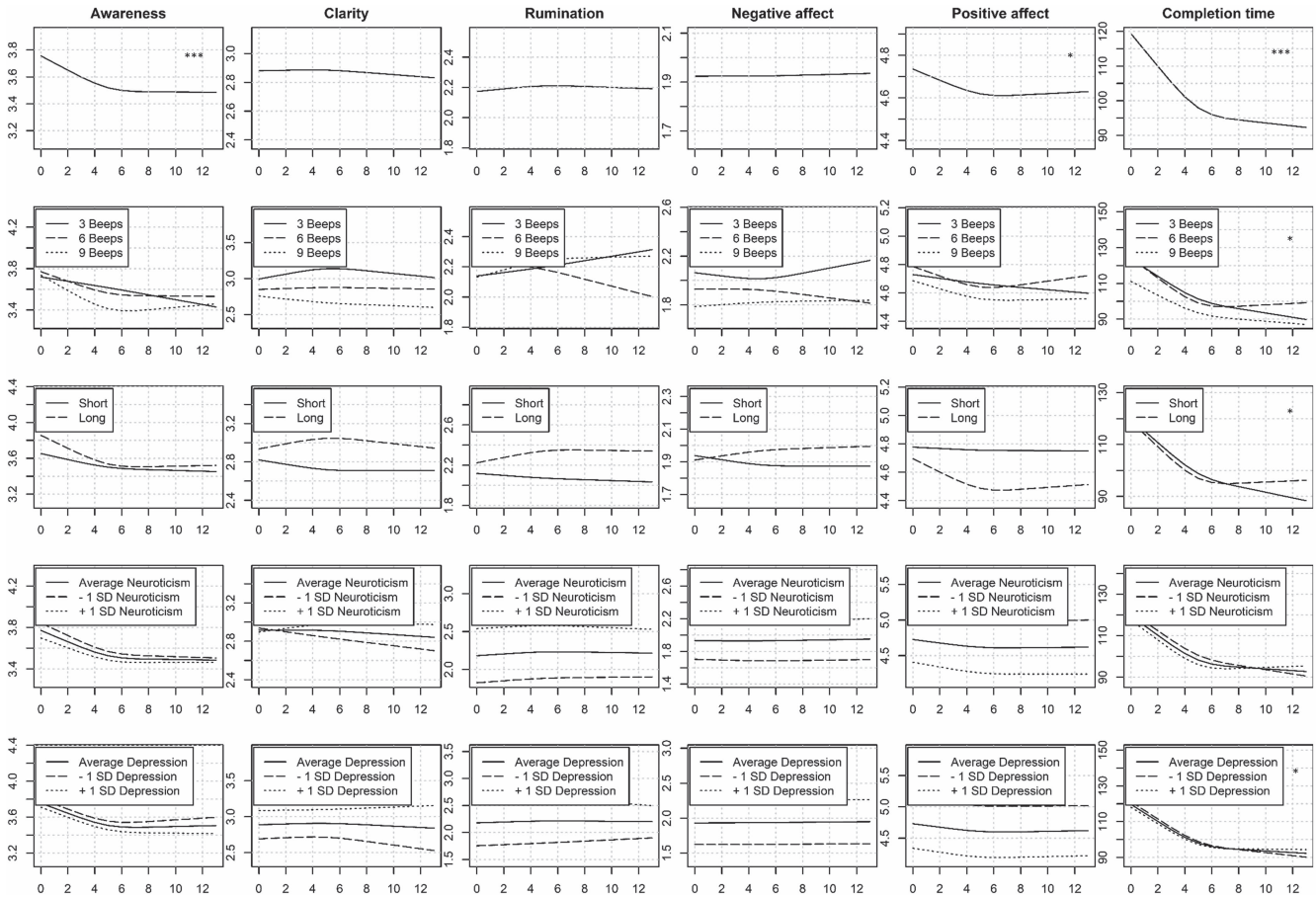
When considering the whole group together, significant decreases over days were observed in emotional awareness, positive affect, and completion time (see Figure 1). We did not detect significant changes over days in mean levels of clarity, rumination, and negative affect. The within-person variance was found to decrease significantly over days for all variables but the completion time, for which the opposite pattern, namely a significant increase in within-person variance over days, could be observed (see Figure 2). The positive coefficient of event valence in predicting positive affect was found to become significantly smaller over days, while no such change was evident for the prediction of negative affect. As suggested by a reviewer, we also explored the significance of the random effects of time in the models to test whether participants differ significantly in the changes over time. Likelihood-ratio tests comparing models with and without the random effect of time indicated that there was a significant amount of heterogeneity in the temporal changes in variables (awareness: $z = 224.739$, $p < .001$; clarity: $z = 250.987$, $p < .001$; rumination: $z = 61.200$, $p < .001$; negative affect: $z = 59.732$, $p < .001$; positive affect: $z = 46.904$, $p < .001$; completion time: $z = 85.551$, $p < .001$). These results highlight that participants followed various different trajectories over time in the study.

The Moderating Role of the Sampling Protocol for Changes Over Time

In the second part of the analysis, we investigated how changes in outcomes over days are influenced by the sampling frequency and the questionnaire length. Changes over days in the mean level of completion time were found to be moderated by both the sampling frequency and the questionnaire length (see Figure 1). Follow-up tests indicated that the decrease in completion time over days in the six beeps group flattened out after about 6 days, which was not the case in the other groups. For participants receiving the long questionnaire version, the decrease in completion time also flattened out after approximately 6 days, while no such flattening of the effect was apparent in the short questionnaire group. Mean levels of the other outcomes were not moderated by sampling frequency or questionnaire length. However, the sampling frequency was found to significantly influence the changes over days in the within-person variance of emotional awareness, clarity, negative affect, rumination, and completion time. Follow-up tests indicated that for emotional awareness, the within-person variance showed a stronger decrease in the six beeps group than in the other groups. For clarity, the decrease in within-person variance was significantly larger in the six beeps compared to the nine beeps group. For rumination, the decrease in within-person variance in the six beeps and nine beeps groups was significantly larger than in the three beeps group. For negative affect, the decrease in within-person variance was significantly larger in the six beeps compared to the three and nine beeps groups. Finally, the increase in within-person variance in completion time was significantly larger in the nine beeps group compared to the three and six beeps groups. In addition, the within-person variance of emotional awareness and clarity showed a stronger decrease in the group who received the long questionnaire. For completion time, the increase in within-person variance was stronger in the long questionnaire group compared to the short

Figure 1

Estimated Changes in Mean Levels of Variables Over Days



Note. Significance level for sampling frequency refers to omnibus test.
* $p < .5$. *** $p < .001$.

questionnaire group. The associations between event pleasantness and affect were not found to change depending on the sampling frequency.

The Moderating Role of Individual Characteristics for Changes Over Time

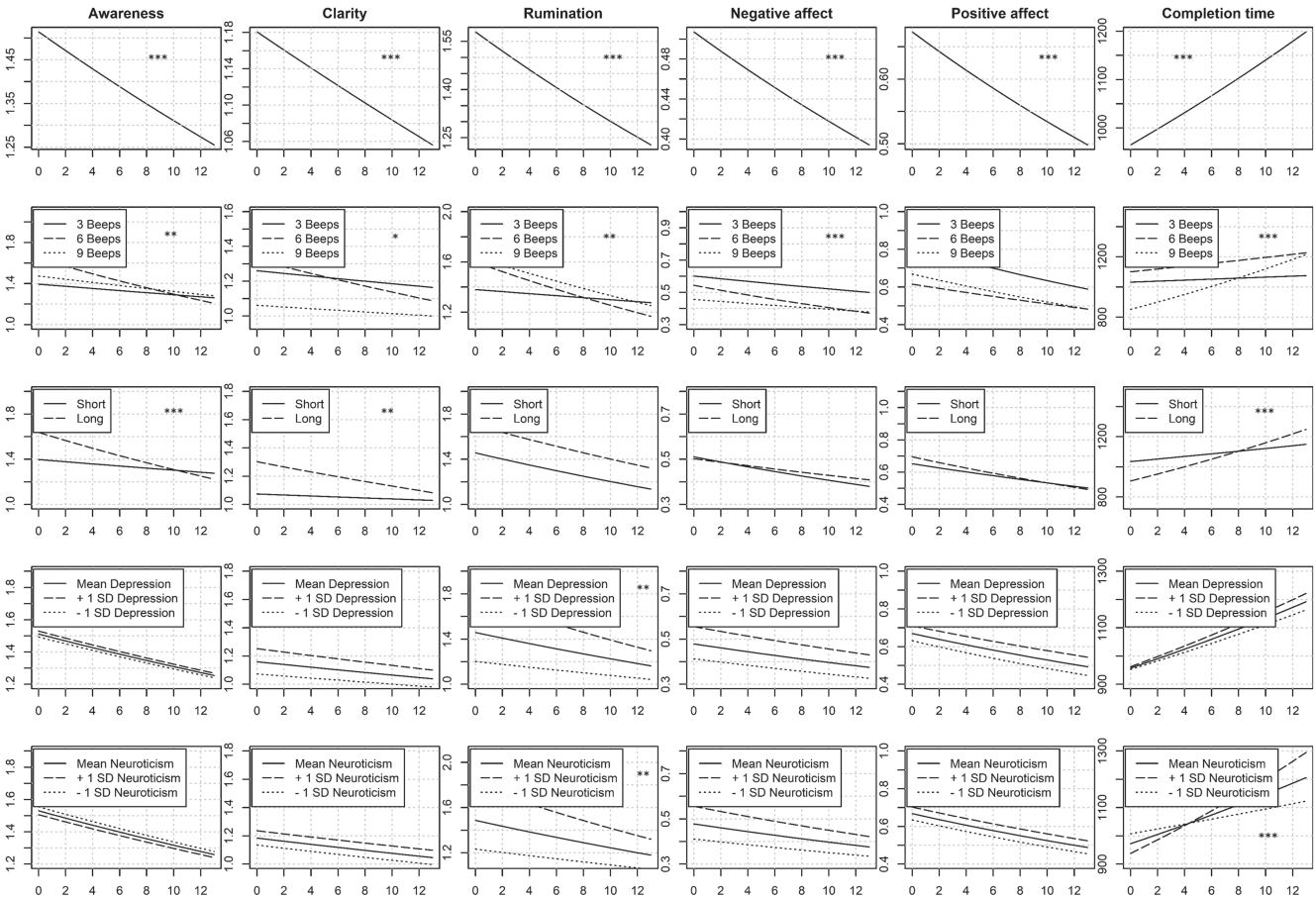
Next, we investigated how changes in outcomes over days are influenced by neuroticism and depression levels at baseline. Neuroticism was found to significantly moderate the decrease in completion time over days (see Figure 1). For individuals scoring higher on neuroticism, the decrease in completion time was flattening out more compared to individuals with lower neuroticism. Changes in mean levels of none of the other variables over days were found to depend on neuroticism or depression. The decrease in the within-person variance of rumination was bigger over days for individuals scoring higher on depression or neuroticism and the increase in the within-person variance of completion time over days was smaller for individuals higher in neuroticism (Figure 2). None of the other changes in within-person variance were significantly influenced by neuroticism or depression level. The increase in the negative coefficient of the event valence variable in predicting negative affect over days was less strong for individuals with a higher score on the baseline measure of neuroticism or depression, while no such changes were observed for positive affect.

Explorative Analyses of the Initial Elevation Bias

Based on reviewer comments, we further explored initial elevation of the data in the current sample. To this end, the analyses of the means for the whole group were repeated by adding a dummy variable for the first day as a predictor. The dummy variable for the first day did not indicate significant differences between first and later days for any of the outcomes (see Supplemental Materials for the models). To further explore the initial elevation bias, we repeated the mean-level analyses for the whole group but added a dummy that indicated only the very first measurement moment that was responded to by a participant. Again, no significant effect of the dummy for the first assessment moment appeared except for the model estimating completion time (see Supplemental Materials). Here, the very first assessment moment had a significantly higher completion time than subsequent measures.

Figure 2

Estimated Changes in Within-Person Variance of Variables Over Days



Note. Significance level for sampling frequency refers to omnibus test.
* $p < .5$. ** $p < 0.01$. *** $p < .001$.

Exploratory Analyses of the Role of Other Personality Variables

As suggested by a reviewer, we explored the role of the other personality variables that were assessed in the study. Changes in mean levels of variables were not affected by the baseline level of extraversion, agreeableness, conscientiousness, or openness to experience. Extraversion and conscientiousness did, however, have a significant effect on changes in the average response speed over time (see Supplemental Figures 1 and 2). The response speed showed a more pronounced decrease for more extraverted individuals, while more conscientious individuals showed a stronger decrease in response time in the first days of the study, which flattened out more toward the end of the study than in individuals lower on conscientiousness. Changes in the within-person variance of variables were also found to be influenced by personality variables. In general, the within-person variance of responses tended to decrease more slowly over time for individuals higher on extraversion, agreeableness, and openness to experience. For the response speed, higher levels of conscientiousness, agreeableness, and extraversion were associated with a lower increase in the within-person variance of response times, while the opposite effect was observed for individuals higher on openness to experience. The detailed results can be consulted in the Supplemental Materials. Finally, the baseline level of conscientiousness was found to impact the changes in the association between event valence and negative affect over time. The negative association between event valence and negative affect was found to become more negative over time for more conscientious individuals.

Qualitative Analysis

We organized the qualitative data under the overarching themes compliance, response process, changes in the person, representativeness of the data, and suggestions for improvement. Identified themes and subthemes that are considered relevant with respect to our research questions are discussed in more detail below, while a full hierarchical overview of all identified themes can be found in the Supplemental Materials. Example quotes for relevant subthemes are reported in Supplemental Table 16.

Changes in the Response Process

Twenty-eight participants (56%) reported an increase in habitual responding over time. This included reports of learning the order of

questions and increased familiarity with the questions over time, which led to easier and hence faster responding over time. For example, Participant 89 described the following evolution:

> I noticed that I could fill it in faster at the end. Aehm not that I thought about it less, but I didn't have to read every question every time. Well, I knew that if it said 'happy', the question was about happy. I had to spend less time on the questions and I knew as well: 'Ah 1 is not at all and 7 is very much'. These things, and also with the answer options, I knew what all the options were and what I was doing, stuff like that.

Some participants described that they developed an automatism in answering the questions. For example, Participant 56 indicated:

> After some time it was less exciting. Because of course I already knew, these questions will come. After some time it became a bit automatic. Like okay. I know roughly what I usually respond and that's correct, some things always come back. Because I hardly ever talk to someone about my emotions, so I knew that it was ... no no no.

However, one participant also described the increased difficulty of remembering when the last beep occurred as the ESM assessments became more and more of a routine (2%). Two participants reported changes in response behavior that were the opposite of habituation, namely an increase in difficulty in responding to questions over time (4%). One participant further stated that they used more extreme numbers toward the end of the study (2%).

Changes in the Underlying Construct

Changes in the underlying construct were further divided into changes in affect, emotional awareness, emotion regulation, and behavior, as well as reports on the absence of changes.

Changes in Affect. Nineteen participants (37%) noted positive affective reactions to the study, such as a positive user experience or a general interest in the study. However, all participants also reported some degree of a negative affective reaction to the study. Many participants ($N = 48$; 96%) described situations in which the assessments were most disturbing for them. Most commonly named themes were disturbance in social situations ($N = 27$; 54%), during class ($N = 20$; 40%), in busy moments ($N = 11$; 22%), and when sleeping ($N = 9$; 18%). For instance, Participant 90 noted:

> I had expected that the beeps would disturb me less, but I did find it quite annoying actually. Like ... well, these kinds of moments, like, 'Is it beeping again?', when I was just doing something for example or just talking to someone. I found that difficult. I had not expected that I would feel this annoyed by it.

Besides disturbance in specific situations, 23 participants (46%) also reported other factors that contributed to their negative experience during the study, such as the loud sound of the device or the fact that they had to pay attention to the phone. Participant 69 for instance noted:

> What did bother me was I try not to spend a lot of time with my mobile phone .... And I had the feeling that I did really have to pay attention to it [the study phone]. That it gives the same feeling as a normal mobile phone, it ... it gives a bit of pressure, I think. That you have to spend time on things that are actually not that important. I mean with the normal phone, it's expected that you are reachable and that you pay attention to it. But I found it was a bit the same. I didn't expect that. I underestimated it a bit. That you do really have to pay attention to it.

Temporal changes in the level of disturbance were also reported. Nine participants (18%) reported an increase in their negative experience over time, while three participants (6%) reported the opposite trend, namely that responding to the beeps became less disturbing or effortful over time.

Changes in Awareness. More than half of the interviewed participants ($N = 32$; 64%) reported becoming more aware of their emotions due to the participation in the study. For example, Participant 54 described the following experience:

> Yes I think it made me more aware of my own emotions. And it showed me that I do have certain patterns in my emotions. And that for example I wouldn't feel down if I am alone in my room, but I would maybe if I am in public .... If I have to take public transport. So that I did maybe feel more down there. And also that, that was something I hadn't realized, that I frequently feel irritated and I hadn't realized. And when filling in the questions I noticed that I do feel more irritated than would be good for me.

Similarly, Participant 50 noted:

> Euhm I did think more about how I feel, yes, because you are getting these questions all the time, like 'Are you happy?', 'Do you feel.' euhm. Definitely in the moment itself, I always had something like a bit a self-reflection, like 'Do I really feel happy? Am I ehm.' and I did find it very interesting for myself just ... because I always saw myself as a pessimist, a negative person, but if I now think back about how I filled in the questionnaire, I almost always scored more than 5 on happiness. So well, yes, I have something like 'Ah yes, okay, then I do not really feel down that often.'

Three participants (6%) reported thinking about what to fill in in-between beeps, and one participant reported increased awareness of what they were doing during the day.

Changes in Emotion Regulation. Seven participants (14%) also reported changes in emotion regulation as a result of the assessments. Participant 64 made the following observation:

> You do think more about how you are feeling and then, yes, how you are feeling but also how you are dealing with it. Sometimes I realize that I am very happy. But if you then ask 'Did you express this emotion', then the answer is like 'Well, almost not at all' .... I tried to deal with it more consciously and to express more clearly like ... or to show more clearly, I feel good now by laughing and also to be more honest. That if people are asking 'how are you?'. And it's not going well, then I want to, well, with friends and not with random people, but I dare to tell friends more honestly, like, I feel a bit questionable.

Changes in Behavior. A large part of participants ($N = 40$; 80%) did report not changing their behavior or routine because of the participation in the study. Only one participant (105; 2%) explicitly mentioned avoiding certain activities not to miss assessments:

> **Did you adapt your daily routine to respond to the beeps?** Sometimes, like when for example someone asked: 'Oh do you want to go swimming?' Then I thought: 'Next week I can go swimming' or to do something, but it doesn't matter. **Okay.** Because for example during all the other things that I did I could always take it, but then when going running, swimming, watching a movie. Well no that would have been possible. But yea, if they asked something like that, then I thought, I will join next week, I just won't now.

Six participants (12%) also reported interrupting their sleep to respond to the assessment at least once. More subtle changes in

behavior included going back to get the phone when it was forgotten somewhere ($N = 2$; 4%) and waiting for beeps ($N = 3$; 6%), which did also lead to active changes in the routine of Participant 113: "I thought, it's almost half past 10, I will wait a bit with sleeping because there were only ten more minutes." Finally, three participants (6%) reported being tempted to change their routine to avoid missing assessments, but did not actually change their behavior.

Discussion

Our aim was to systematically investigate measurement reactivity to ESM by looking at objective changes in the data and analyzing subjective reports of participants given during interviews. Reactivity was broadly defined as any changes in participants that were caused by the ESM assessments. We identified several potential reactive effects in ESM data. While increases in emotional awareness were frequently reported in interviews, the ESM measure of momentary emotional awareness was found to decrease over time in the study. In addition, positive affect was also found to become lower over time. Quantitative analyses revealed decreases in completion time and in the within-person variance of variables over the duration of the study. Qualitative data offer support for an interpretation of these observed changes as a habituation effect. The effects of the sampling protocol and individual characteristics were inconsistent and did not support our hypotheses on stronger reactivity in higher sampling frequency groups and for individuals scoring higher on depression or neuroticism. In addition, individual participants reported various other reactive effects during the interviews.

Changes in Underlying Construct

Some of the observed changes suggest that ESM led to changes in some of the underlying constructs that were assessed. Most pronounced were increases in emotional awareness that were reported by most of the participants during follow-up interviews. Such increases in (emotional) awareness are in line with previous findings from qualitative studies in healthy participants and patients (Bos et al., 2020; Kauer et al., 2012; Moitra et al., 2017; Smelror et al., 2019; Turner et al., 2019; Van Dam et al., 2019; Widdershoven et al., 2019). To our surprise, the reported increase in emotional awareness was not apparent in the ESM data, in which momentary emotional clarity was not found to change and momentary emotional awareness was even found to decrease significantly over the first 6 days of the study (estimated drop from 3.76 to 3.49 on a 7-point Likert-type scale over 14 days, which represents a 7% decrease). This was also the case in additional exploratory analyses in which we investigated the changes in ESM measures of emotional awareness and clarity only in individuals who had reported increases in emotional awareness during interviews. These observations are not consistent with previous findings that detected increases in retrospectively assessed emotional awareness (Kauer et al., 2012; Ludwigs et al., 2018) and emotion differentiation (Hoemann et al., 2021; Widdershoven et al., 2019) due to participation in an ESM study. However, the effect of ESM on momentary emotional awareness and clarity as assessed in the present study has, to our knowledge, not been investigated previously. There are different processes that may explain the observed decrease in emotional awareness over time. Considering the qualitative findings, it is, for instance, possible that emotional awareness was artificially increased during the first assessment moments as a reactive effect and returned to baseline after participants became increasingly habituated to the assessments (in line with changes in response behavior that are discussed later). Following this interpretation, the qualitative and quantitative results may not contradict each other. Alternatively, a retrospective bias may be present in interview data, which may have been tainted by specific key experiences during the study rather than by a consistent increase in emotional awareness over time.

Aside from the changes in emotional awareness, we also detected a small but statistically significant decrease in positive affect over the first 6 days of the study (estimated drop from 4.74 to 4.63 on a 7-point Likert-type scale over 14 days, which represents a decrease of 2% of the initial score). Similar decreases in positive affect have been previously detected for individuals high in neuroticism or depression with similar sampling frequencies, while increases in happiness have been observed for individuals scoring lower on neuroticism and depression (Conner & Reid, 2012). The current results do not allow us to identify the mechanism underlying these changes with certainty. However, an increase in burden is one possible explanation and would be in line with qualitative reports of burden in general, as well as of increases in burden over time in particular. It is also possible that both changes in emotional awareness and affect were methodological artifacts driven by changes in response behavior, which will be discussed in the next paragraph. However, it is not clear why not all variables would have been affected by such changes. In sum, more work is needed to explore these changes and their underlying causes in more detail. However, these findings underline the need to use ESM control groups when evaluating the effect of an intervention with ESM, as simple changes in ESM data over time are apparent also in the absence of an intervention.

Changes in Response Behavior

Several changes in response behavior were detected that point toward a habituation effect. Specifically, we detected consistent decreases in the within-person variance of responses, in line with previous studies (Csikszentmihalyi & Larson, 1987; Fuller-Tyszkiewicz et al., 2013; Vachon et al., 2016). The within-person variances of variables decreased with ratios varying from 0.90 for clarity to 0.74 for positive affect over the 2-week period, which corresponds to 11% and 26% of the initial within-person variance. In addition, participants were becoming faster at responding to questions over time, which is also in line with previous results (Arslan et al., 2021; Husky et al., 2010; Johnson et al., 2009). The predicted decrease of 23% from 4.4 s per item on Day 1 to 3.4 s per item on Day 14 is comparable to what has been reported in a previous diary study, where response times evolved from 5 s per item on Day 1 to 2–2.5 s on Day 30 of data collection. Alongside these changes in the data, many participants reported becoming more habituated to the measures over time in interviews, as has been documented in one previous study (Paterson et al., 2020).

In itself, habituation to the ESM measures can be seen as a beneficial change over time because the ESM assessments take less time for participants to complete and therefore interfere less with their routines. However, whether the observed changes in response behavior are also associated with a decrease in data quality, in line with a fatigue effect, is more difficult to judge based on the current results. We observed decreases in the strength of the association

between event valence and positive affect over time, which may point toward a decrease in data quality over time. Yet, the association between event valence and negative affect was not found to change over time. It is possible that the size and change of this association was distorted by the overall low levels of negative affect in the current student sample. Participants did also not mention becoming less accurate over time in interviews; however, these reports are likely influenced by social desirability. Recent findings that combine objective measures of social media use with self-report data do indicate a decrease in accuracy of ESM ratings over time (Verbeij et al., 2021). Also, previously reported decreases in compliance over time in the same data set (Eisele et al., 2020) and other ESM data (Forkmann et al., 2018; Ono et al., 2019; Rintala et al., 2018; Silvia et al., 2013) are consistent with a fatigue effect. Qualitative reports in the present study also confirm that at least some participants experience an increase in assessment burden over time. However, previous analyses of the current data did not support increases in ESM measures of perceived burden or careless responding over time, which would also be expected in case of such a fatigue effect (Eisele et al., 2020).

Initial Elevation Bias

Detected changes either spanned the first 6 days or progressed continuously over the whole study period. Additional exploratory analyses indicated that the group-level changes over time were not specific for the first day or beep (except for the completion time, which was significantly higher at the first filled-in compared to subsequent beeps). The changes were therefore not consistent with an initial elevation bias that specifically affects the first assessment day or beep. However, "initial" can be defined in other ways. For example, the first ESM measure was preceded by several self-report questionnaires during the baseline session in the present study. It is possible that an initial elevation bias would have been limited to these cross-sectional questionnaires. Nevertheless, temporal changes relatively early in the study were detected, and qualitative reports describing changes in response behavior in the beginning could be in line with a change during the first days of assessments and therefore consistent with other definitions of an initial elevation bias.

The Role of the Sampling Protocol and Individual Characteristics

Our initial hypothesis of increased reactive effects based on sampling frequency and baseline depression and neuroticism level was not confirmed. These findings contrast with previous reports of differential reactive effects based on the sampling frequency (Conner & Reid, 2012; McCarthy et al., 2015), but are in line with results reported by Stone et al. (2003). Some changes in response behavior (i.e., within-person variance and completion time) varied based on questionnaire length and sampling frequency. This suggests that there may be differences in changes in response behavior over time between these groups. However, the differences between different sampling frequency groups were not consistent. Additionally, the decrease in completion time was found to flatten out in the six beeps and long questionnaire groups, but changes in within-person variance were found to be stronger in these groups. This combination of changes is difficult to explain in terms of changes in response behavior, as both habituation and fatigue effects were expected to be associated with a simultaneous decrease in completion time and within-person variance of ratings.

When it comes to the influence of individual characteristics on reactivity, some effects on response behavior were found to be less pronounced for individuals scoring higher on neuroticism and depression. However, we did not find the effect described by Conner and Reid (2012) or effects consistent with some individuals being more vulnerable to changes in constructs with sampling frequencies as high as ours. Differences in protocols between Conner and Reid's and our study may explain this discrepancy, as literature suggests that more focused questionnaires (i.e., assessing fewer constructs) may induce more reactivity (Korotitsch & Nelson-Gray, 1999). Our study used a questionnaire assessing multiple constructs, while participants in Conner and Reid's study did solely rate different aspects of their happiness. Alternatively, the sample size of the present study may have been too small to detect an interaction effect between individual characteristics and time. Exploratory analyses with other personality variables highlight that the changes in response behavior over time do also depend on the level of extraversion, agreeableness, openness, and, to a lesser extent, conscientiousness of a participant. In addition, there was considerable variation in participants' individual trajectories over time, as highlighted by significant random effects of time in the models. These individual trajectories could be the result of the sampling and hence be signs of heterogeneity in reactive effects. However, it is important to note that random changes in variables over time can also appear in the absence of measurement reactivity (e.g., the mood of a participant could change in response to events in their daily life).

Lessons From Qualitative Feedback

The experiences during the study that participants described in interviews showed a lot of variation. However, a number of themes were applicable to a large number of participants and may point toward issues that could be tackled in the future to optimize the following of instructions, the experience of participants during the study, and to reduce reactive changes in participants. It became, for instance, apparent that the timing of the beeps (9 a.m. to 10:30 p.m.), which is commonly used in ESM studies, conflicted with the sleeping schedules of many participants. This problem may be addressed by individualizing the beep schedules to fit the daily lives of participants, as has already been done in some ESM studies (e.g., Bastiaansen et al., 2020). Additionally, we noticed that a large part of participants experienced discomfort when responding to questionnaires during social interactions, highlighting that this is an important topic to address during briefing sessions to avoid missing data. Although not the focus of the current article, the qualitative data also gave some insights as to when beeps are missed. Specifically, participants frequently found themselves unable to respond at work or while attending classes or studying. A large part of missed beeps was also due to participants forgetting to take the study phone with them, which may be reduced by relying on the participants own phone in the future, a suggestion that was also specifically made by some participants.

Constraints on Generality, Limitations, and Directions for Future Work

The present study was conducted in a young student sample. Even though the current results do not support differences in reactivity

based on neuroticism or depression, it is unclear to what extent the current findings can be generalized to other populations, which may be more or less affected by responding to ESM assessments. The effects of personality variables that emerged in exploratory analyses point toward the idiographic nature of reactive changes in ESM research. The significance of the random effects of time in the mean models also points in this direction, as we observed large variation in the size and direction of reactive changes. It was not possible to investigate the variations in changes in the within-person variance over time between individuals, as such modeling approaches are not currently available. It would be interesting to explore the variability of changes in variance in the future. While the current results highlight changes in the data that appear early on during ESM monitoring, it is possible that other changes take place later on and that therefore could not be detected in the present study. As personalized approaches to psychiatry with large numbers of data points per individual and monitoring of individuals over extended periods of time become more popular, reactive changes that take more time to appear become increasingly relevant to investigate. Further, while our sample offered the unique opportunity to directly compare the effects of different sampling protocols on reactivity, the sample size was also limited. The resulting limited power to detect effects, especially interaction effects, represents a serious limitation of the current research. To establish robust effects, individual data meta-analyses may be a useful approach in the future.

The present study aimed to investigate the presence of reactive effects. Previous work has gathered a number of possible mechanisms that could explain reactive changes over time (Fuller-Tyszkiewicz et al., 2013; Patrick & Gilbert, 1998; Shrout et al., 2018). To further our understanding of the underlying mechanisms once robust effects are established, carefully designed experiments (along the lines of De Vuyst et al., 2019; Johar & Sackett, 2018; Shrout et al., 2018) or the analysis of data sets that combine both objective and self-report measures (Verbeij et al., 2021) seem promising approaches. Relatedly, more work is needed to judge the practical significance of the observed affects, for example, by investigating how reactive changes affect results of analyses on a practical level (see, e.g., Weermeijer et al., 2022). Furthermore, it is difficult to compare the size of observed effects to other effects in the ESM literature due to the large variability in ESM study designs and the absence of an agreement on standardized effect sizes for multilevel model (Rights & Sterba, 2019).

Finally, it is important to note that awareness, clarity, and rumination were all measured with individual items. It is possible that reactive effects are different depending on the number of items that make up a measure, as was pointed out by a reviewer. It is, for instance, imaginable that individual items are more affected by a reactive increase in random noise, while composite measures may be more likely to trigger changes in the underlying construct (e.g., the inclusion of four items about rumination may make it more likely that participants recognize and act upon unhealthy amounts of rumination).

Conclusion

Measurement reactivity could threaten the validity of findings and should therefore be a concern for every researcher relying on self-report data. Our results indicate the presence of some reactive effects in the level and within-person variance of ESM measures, as well as the speed with which responses are given. Combined with qualitative data, these changes seem to be the result of a habituation effect. The shape of the observed changes did not support an initial elevation bias limited to the first assessment moment or first day of data. Qualitative data also gave insight into the wide variety of experiences of participants during ESM studies based on which we formulate recommendations for future studies. The current results further suggest that some research questions may be more affected by reactive effects than others. Based on the current findings, researchers should be especially aware of the possibility of reactive changes when investigating variability in momentary experiences, completion times, positive affect and its association with event valence, or emotional awareness. The current results do not allow us to conclude whether the observed reactive changes are detrimental to the validity of the collected data. As the influence of reactive changes remains poorly understood, researchers should routinely test for reactive changes in their ESM data and discuss openly how reactivity may have influenced their results.

References

Aaron, L. A., Turner, J. A., Mancl, L., Brister, H., & Sawchuk, C. N. (2005). Electronic diary assessment of pain-related variables: Is reactivity a problem? *Journal of Pain*, *6*(2), 107–115. https://doi.org/10.1016/j.jpain.2004.11.003

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

Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 108–123). The Guilford Press.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S. M., de Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., … Bringmann, L. F. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, Article 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Beal, D. J. (2015). ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 383–407. https://doi.org/10.1146/annurev-orgpsych-032414-111335

Bos, F. M., Snippe, E., Bruggeman, R., Doornbos, B., Wichers, M., & van der Krieke, L. (2020). Recommendations for the use of long-term experience sampling in bipolar disorder care: A qualitative study of patient and clinician experiences. *International Journal of Bipolar Disorders*, *8*(1), Article 38. https://doi.org/10.1186/s40345-020-00201-5

Broderick, J. E., & Vikingstad, G. (2008). Frequent assessment of negative symptoms does not induce depressed mood. *Journal of Clinical Psychology in Medical Settings*, *15*(4), 296–300. https://doi.org/10.1007/s10880-008-9127-6

Conner, T. S., & Reid, K. A. (2012). Effects of intensive mobile happiness reporting in daily life. *Social Psychological & Personality Science*, *3*(3), 315–323. https://doi.org/10.1177/1948550611419677

Cruise, C. E., Broderick, J., Porter, L., Kaell, A., & Stone, A. A. (1996). Reactive effects of diary self-assessment in chronic pain patients. *Pain*, *67*(2), 253–258. https://doi.org/10.1016/0304-3959(96)03125-9

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, *175*(9), 526–536. https://doi.org/10.1097/00005053-198709000-00004

Denissen, J. J. A., Geenen, R., van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008). Development and validation of a dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment*, *90*(2), 152–157. https://doi.org/10.1080/00223890701845229

De Vuyst, H. J., Dejonckheere, E., Van der Gucht, K., & Kuppens, P. (2019). Does repeatedly reporting positive or negative emotions in daily life have an impact on the level of emotional experiences and depressive symptoms over time? *PLOS ONE*, *14*(6), Article e0219121. https://doi.org/10.1371/journal.pone.0219121

Ebner-Priemer, U. W., Kuo, J., Kleindienst, N., Welch, S. S., Reisch, T., Reinhard, I., Lieb, K., Linehan, M. M., & Bohus, M. (2007). State affective instability in borderline personality disorder assessed by ambulatory monitoring. *Psychological Medicine*, *37*(7), 961–970. https://doi.org/10.1017/S0033291706009706

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-germeys, I., & Viechtbauer, W. (2020). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

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

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

French, D. P., & Sutton, S. (2010). Reactivity of measurement in health psychology: How much of a problem is it? What can be done about it? *British Journal of Health Psychology*, *15*(3), 453–468. https://doi.org/10.1348/135910710X492341

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

Harrell, F., Jr. (2021). *rms: Regression modeling strategies* (R package Version 6.2-0). https://CRAN.R-project.org/package=rms

Harrell, F. E. (2015). *Regression modeling strategies. Springer series in statistics* (Vol. 64). Springer. https://doi.org/10.1007/978-1-4757-3462-1

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328–3336. https://doi.org/10.1002/sim.5338

Hoemann, K., Barrett, L. F., & Quigley, K. S. (2021). Emotional granularity increases with intensive ambulatory assessment: Methodological and individual factors influence how much. *Frontiers in Psychology*, *12*, Article 704125. https://doi.org/10.3389/fpsyg.2021.704125

Husky, M. M., Gindre, C., Mazure, C. M., Brebant, C., Nolen-Hoeksema, S., Sanacora, G., & Swendsen, J. (2010). Computerized ambulatory monitoring in mood disorders: Feasibility, compliance, and reactivity. *Psychiatry Research*, *178*(2), 440–442. https://doi.org/10.1016/j.psychres.2010.04.045

Johar, O., & Sackett, A. M. (2018). The self-contaminating nature of repeated reports of negative emotions. *Basic and Applied Social Psychology*, *40*(5), 293–307. https://doi.org/10.1080/01973533.2018.1496336

Johnson, E. I., Grondin, O., Barrault, M., Faytout, M., Helbig, S., Husky, M., Granholm, E. L., Loh, C., Nadeau, L., Wittchen, H. U., & Swendsen, J. (2009). Computerized ambulatory monitoring in psychiatry: A multi-site collaborative study of acceptability, compliance, and reactivity. *International Journal of Methods in Psychiatric Research*, *18*(1), 48–57. https://doi.org/10.1002/mpr.276

Kauer, S. D., Reid, S. C., Crooke, A. H., Khor, A., Hearps, S. J., Jorm, A. F., Sanci, L., & Patton, G. (2012). Self-monitoring using mobile phones in the early stages of adolescent depression: Randomized controlled trial. *Journal of Medical Internet Research*, *14*(3), Article e67. https://doi.org/10.2196/jmir.1858

Knowles, E. S., Coker, M. C., Scott, R. A., Cook, D. A., & Neville, J. W. (1996). Measurement-induced improvement in anxiety: Mean shifts with repeated assessment. *Journal of Personality and Social Psychology*, *71*(2), 352–363. https://doi.org/10.1037/0022-3514.71.2.352

Korotitsch, W. J., & Nelson-Gray, R. O. (1999). An overview of self-monitoring research in assessment and treatment. *Psychological Assessment*, *11*(4), 415–425. https://doi.org/10.1037/1040-3590.11.4.415

Kramer, I., Simons, C. J. P., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., Schruers, K., van Bemmel, A. L., Myin-Germeys, I., Delespaul, P., van Os, J., & Wichers, M. (2014). A therapeutic application of the experience sampling method in the treatment of depression: A randomized controlled trial. *World Psychiatry*, *13*(1), 68–77. https://doi.org/10.1002/wps.20090

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. H. T. Reis (Ed.), *Naturalistic approaches to studying social interaction (new directions for methodology of social and behavioral sciences)* (Vol. 15, pp. 41–56). Jossey-Bass.

Law, M. K., Furr, R. M., Arnold, E. M., Mneimne, M., Jaquett, C., & Fleeson, W. (2015). Does assessing suicidality frequently and repeatedly cause harm? A randomized control study. *Psychological Assessment*, *27*(4), 1171–1181. https://doi.org/10.1037/pas0000118

Litt, M. D., Cooney, N. L., & Morse, P. (1998). Ecological momentary assessment (EMA) with treated alcoholics: Methodological problems and potential solutions. *Health Psychology*, *17*(1), 48–52. https://doi.org/10.1037/0278-6133.17.1.48

Ludwigs, K., Lucas, R., Burger, M., Veenhoven, R., & Arends, L. (2018). How does more attention to subjective well-being affect subjective well-being? *Applied Research in Quality of Life*, *13*(4), 1055–1080. https://doi.org/10.1007/s11482-017-9575-y

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

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*(4), 1510–1515. https://doi.org/10.3758/s13428-019-01330-1

Moitra, E., Gaudiano, B. A., Davis, C. H., & Ben-Zeev, D. (2017). Feasibility and acceptability of post-hospitalization ecological momentary assessment in patients with psychotic-spectrum disorders. *Comprehensive Psychiatry*, *74*, 204–213. https://doi.org/10.1016/j.comppsych.2017.01.018

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

Palmier-Claus, J. E., Ainsworth, J., Machin, M., Barrowclough, C., Dunn, G., Barkus, E., Rogers, A., Wykes, T., Kapur, S., Buchan, I., Salter, E., & Lewis, S. W. (2012). The feasibility and validity of ambulatory self-report of psychotic symptoms using a smartphone software application. *BMC Psychiatry*, *12*(1), Article 172. https://doi.org/10.1186/1471-244X-12-172

Paterson, C., Primeau, C., & Lauder, W. (2020). What are the experiences of men affected by prostate cancer participating in an ecological momentary assessment study? *Cancer Nursing*, *43*(4), 300–310. https://doi.org/10.1097/NCC.0000000000000699

Patrick, J., & Gilbert, D. G. (1998). Effects of repeated administratin of the beck inventory and other measures of negative mood states. *Personality and Individual Differences*, *24*(4), 457–463. https://doi.org/10.1016/S0191-8869(97)00193-1

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2021). *_nlme: Linear and nonlinear mixed effects models_* (R package Version 3.1-152). https://CRAN.R-project.org/package=nlme

QSR International Pty Ltd. (2020) *NVivo* (release 1.3). https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Radloff, L. (2012). Center for epidemiologic studies depression scale. In M. Sajatovic & L. F. Ramirez (Eds.), *Rating scales in mental health* (pp. 109–111). Johns Hopkins University Press.

Ram, N., Brinberg, M., Pincus, A. L., & Conroy, D. E. (2017). The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. *Research in Human Development*, *14*(3), 253–270. https://doi.org/10.1080/15427609.2017.1340052

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338. https://doi.org/10.1037/met0000184

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

Rowan, P. J., Cofta-Woerpel, L., Mazas, C. A., Vidrine, J. I., Reitzel, L. R., Cinciripini, P. M., & Wetter, D. W. (2007). Evaluating reactivity to ecological momentary assessment during smoking cessation. *Experimental and Clinical Psychopharmacology*, *15*(4), 382–389. https://doi.org/10.1037/1064-1297.15.4.382

Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, *4*(1925), 5–34. https://doi.org/10.1023/A:1023605205115

Shiffman, S., Engberg, J. B., Paty, J. A., Perz, W. G., Gnys, M., Kassel, J. D., & Hickcox, M. (1997). A day at a time: Predicting smoking lapse from daily urge. *Journal of Abnormal Psychology*, *106*(1), 104–116. https://doi.org/10.1037/0021-843X.106.1.104

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(1), E15–E23. https://doi.org/10.1073/pnas.1712277115

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

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Smelror, R. E., Bless, J. J., Hugdahl, K., & Agartz, I. (2019). Feasibility and acceptability of using a mobile phone app for characterizing auditory verbal hallucinations in adolescents with early-onset psychosis: Exploratory study. *JMIR Formative Research*, *3*(2), Article e13882. https://doi.org/10.2196/13882

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

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

Turner, C. M., Arayasirikul, S., Trujillo, D., Lê, V., & Wilson, E. C. (2019). Social inequity and structural barriers to completion of ecological momentary assessments for young men who have sex with men and trans women living with HIV in San Francisco. *JMIR mHealth and uHealth*, *7*(5), Article e13241. https://doi.org/10.2196/13241

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

Van Dam, L., Rietstra, S., Van der Drift, E., Stams, G. J. J. M., Van der Mei, R., Mahfoud, M., Popma, A., Schlossberg, E., Pentland, A., & Reid, T. G. (2019). Can an emoji a day keep the doctor away? An explorative mixed-methods feasibility study to develop a self-help app for youth with mental health problems. *Frontiers in Psychiatry*, *10*, Article 593. https://doi.org/10.3389/fpsyt.2019.00593

van de Velde, S., Levecque, K., & Bracke, P. (2011). Vlaanderen versus Nederland: Verschillen in depressieve klachten bij mannen en vrouwen gemeten met de CES-D8. *Tijdschrift voor Psychiatrie*, *53*(2), 73–82.

Verbeij, T., Pouwels, J. L., & Valkenburg, P. M. (2021). The accuracy and validity of self-reported social media use measures among adolescents. *Computers in Human Behavior Reports*, *3*, Article 100090. https://doi.org/10.1016/j.chbr.2021.100090

von Baeyer, C. L. (1994). Reactive effects of measurement of pain. *The Clinical Journal of Pain*, *10*(1), 18–21. https://doi.org/10.1097/00002508-199403000-00004

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

Widdershoven, R. L. A., Wichers, M., Kuppens, P., Hartmann, J. A., Menne-Lothmann, C., Simons, C. J. P., & Bastiaansen, J. A. (2019). Effect of self-monitoring through experience sampling on emotion differentiation in depression. *Journal of Affective Disorders*, *244*, 71–77. https://doi.org/10.1016/j.jad.2018.10.092

Zawadzki, M. J., Scott, S. B., Almeida, D. M., Lanza, S. T., Conroy, D. E., Sliwinski, M. J., Kim, J., Marcusson-Clavertz, D., Stawski, R. S., Green, P. M., Sciamanna, C. N., Johnson, J. A., & Smyth, J. M. (2019). Understanding stress reports in daily life: A coordinated analysis of factors associated with the frequency of reporting stress. *Journal of Behavioral Medicine*, *42*(3), 545–560. https://doi.org/10.1007/s10865-018-00008-x

Received August 19, 2021
Revision received August 2, 2022
Accepted August 9, 2022 ■