Recommendations for Conducting Longitudinal Experience Sampling Studies

Niels van Berkel and Vassilis Kostakos

Abstract The Experience Sampling Method is used to collect participant self-reports over extended observation periods. These self-reports offer a rich insight into the individual lives of study participants by intermittently asking participants a set of questions. However, the longitudinal and repetitive nature of this sampling approach introduces a variety of concerns regarding the data contributed by participants. A decrease in participant interest and motivation may negatively affect study adherence, as well as potentially affecting the reliability of participant data. In this chapter, we reflect on a number of studies that aim to understand better participant performance with Experience Sampling. We discuss the main issues relating to participant data for longitudinal studies and provide hands-on recommendations for researchers to remedy these concerns in their own studies.

Keywords Experience sampling method · Ecological momentary assessment · ESM · EMA · Self-report · Data quality · Reliability

1 Introduction

Responding to an increased interest in studying human life more systematically than traditional surveys—and in a more realistic and longitudinal setting than possible through observations—Larson and Csikszentmihalyi introduced the Experience Sampling Method in 1983 [1]. Researchers using the Experience Sampling Method (ESM) ask their participants to intermittently complete a short questionnaire assessing their current state, context, or experience over an extended period of time (typically a couple of weeks). Questionnaires are typically designed to ensure that participants focus on their current experience rather than to reflect over a longer

N. van Berkel (✉)
Aalborg University, Aalborg, Denmark
e-mail: nielsvanberkel@cs.aau.dk

V. Kostakos
The University of Melbourne, Melbourne, Australia
e-mail: vassilis.kostakos@unimelb.edu.au

E. Karapanos et al. (eds.), *Advances in Longitudinal HCI Research*,
Human–Computer Interaction Series, https://doi.org/10.1007/978-3-030-67322-2_4

period of time, thus minimising the effects of participants' (in)ability to accurately recollect past events [2].

Early ESM studies focused on capturing the daily activities and corresponding experiences of study participants [3]. In those studies, participants were asked to answer what they were currently doing repeatedly. Collecting self-reports at random slots throughout the day, as opposed to a one-off survey or interview, ensured that responses are collected during the participant's "*interaction with the material and social environment*" [3]. In other words, the idea to collect self-report data in situ and thereby increase the ecological validity of a study was motivated by a desire to increase the reliability of participant responses.

A recent survey indicated an increased adoption of the Experience Sampling Method, with a focus on (personal) mobile devices [4]. The use of mobile devices as opposed to paper-based questionnaires provides a number of advances in terms of control over participant entries (e.g. prevent 'parking lot compliance' [5]), interactive design opportunities [6, 7], and contextual sensing possibilities [8–10]. We discuss how these opportunities provided by mobile devices can be utilised in the assessment, improvement, and analysis of the reliability of participant data in longitudinal experience sampling studies.

1.1 Longitudinal Experience Sampling

The timescale of ESM studies varies significantly, with a recent literature review (analysing 461 papers) reporting studies ranging between 1 and 365 days [4]. The median duration of an ESM study was found to be 14 days, while a majority of 70.9% of studies reported a duration of less than one month [4]. The one-day studies in the sample are mostly trials to investigate the (technological) feasibility of a given study configuration (e.g. Westerink et al. [11]). The longest study, totalling a year, investigated long term patterns in location sharing among a large sample of Foursquare users [12]. The typical range of ESM studies is in the duration of weeks rather than months as researchers aim to find a "*balance between study duration and intervention frequency*" [13].

Longitudinal experience sampling is relatively short-term when compared to cross-sectional repeated surveys (also called periodic surveys or simply a survey using a longitudinal design), typically covering months or years [14]. These survey-type designs are often used to investigate changes in attitudes or behaviours over extended periods of time [14], for example in consumer research [15] or within professional organisations [16]. In addition to their usual shorter duration, there are a number of other key differences between repeated surveys and longitudinal experience sampling: the frequency of the questionnaires, the reflective nature of surveys vs. the in-the-moment perspective of ESM questionnaires, and the fact that ESM questionnaires are collected 'in the wild' aiming to cover a variety of contexts. The ESM shares many of the same challenges encountered in other methodologies employing human sensing [10, 17], such as citizen science or situated crowdsourcing.

1.2 Challenges

The sustained effort required of participants over an extended period of time introduces a number of challenges. First, the motivation of participants is likely to decrease over time as initial interest drops. Techniques to maintain a base level of motivation, whether through intrinsic or extrinsic motivation, are therefore key in enabling successful longitudinal use of the ESM. Participant motivation, or lack thereof, plays a key role in relation to data quality and quantity, the two remaining challenges. Second, adherence to study protocol—typically quantified as the number of questionnaires that have been answered—has been shown to decline over time due to study fatigue [18]. Another concern is the variance in the number of responses between participants, which could skew the analysis of ESM results—a critical type of bias introduced by such variance is 'selective non-responses', in which the responses of specific groups of the study's sample are over- or under-represented [19]. An analysis of four recent ESM studies reveals significant differences across participants in terms of their response rate [20]. Third, ensuring a sufficient level of response reliability is key in collecting participant responses, and critical in generating sound study inferences. Novel sampling techniques and filtering mechanisms can support the increase in reliability of participant responses.

Here, we discuss these three challenges in detail and provide concrete recommendations for researchers to address these challenges in their own studies (Sects. 2, 3, and 4). Following this, we discuss analysis techniques specific to the analysis of longitudinal response data (Sect. 5) as well a number of concrete guidelines for the design and subsequent reporting of ESM studies through a 'checklist for researchers' (Sect. 6). Finally, we present a number of future trends in the area of longitudinal experience sampling studies (Sect. 7) and conclude this chapter (Sect. 8).

2 Participant Motivation

Larson and Csikszentmihalyi classify the "*dependence on respondents' self-reports*" as the major limitation of the ESM, while simultaneously highlighting examples that show how these self-reports are "*a very useful source of data*" [2]. Regardless of whether we consider the quantity or quality of participant responses, participant motivation is key in ensuring a successful study outcome. Given the longitudinal and oftentimes burdensome nature of ESM studies, a number of research streams have explored how to increase and maintain participant motivation over time and its subsequent effects on participant responses. Here, we distinguish between intrinsic and extrinsic means of motivation.

2.1 Intrinsic Motivation

Intrinsic motivation has simply been defined as "*doing something for its own sake*" [21] rather than expecting a direct or indirect compensation. It is, however, incorrect to state that researchers can therefore not (positively) influence a participant's intrinsic motivation. As already stated by Larson and Csikszentmihalyi in their original publication on the Experience Sampling Method: "*Most participants find that the procedure is rewarding in some way, and most are willing to share their experience. However, cooperation depends on their trust and on their belief that the research is worthwhile*" [1]. Here, Larson and Csikszentmihalyi refer to what they later classify as 'establishing a research alliance'. This research alliance aims to establish a vested interest of the participant in the study and the research outcome.

However, identifying *how* to give concrete form to such a research alliance remains under-explored in the current ESM literature. Related methodologies such as citizen science face similar challenges and have investigated how to build and sustain engagement among participants. These results show that interest and curiosity, perceived self-competence, and enjoyment in the task all contribute to an individual's intrinsic motivation [22, 23]. Furthermore, Measham and Barnett found that fulfilling a participant's initial motivation for participation increases the duration of a participant's engagement [24]. Although direct empirical evaluations of these factors are scarce for the ESM, given the methodological overlap we can hypothesise that these factors have a similar positive effect on participation motivation in ESM studies. We note that the potential side effects of increasing participants' motivation have not yet been sufficiently explored, and could potentially influence study results.

Recommendation 1 Provide rich feedback regarding the study goals and the participants' contribution to those goals. Provide information throughout the study period.

Recommendation 2 Target participant recruitment to communities with a vested interest in the study outcomes.

2.2 Extrinsic Motivation

Extrinsic motivation, which Reiss defines as "*the pursuit of an instrumental goal*" [21], consists of various methods of motivation, including (financial) rewards or a competition between participants. Although earlier work in Psychology stated that extrinsic motivators would undermine an individual's intrinsic motivation (cf. the self-determination theory [25]), recent work largely refutes this claim [21].

A (financial) compensation of participants is common for ESM studies, with a fixed compensation at the end of the study period being the most widely used (45.7%) [4]. The effect of different financial compensation structures on participant motivation has not been extensively explored, in part due to incomplete reporting of study details [4]. These initial reports do highlight, however, that the use of micro-compensations (a small payment for each completed response) motivates participants

in responding to ESM questionnaires. Although already applied by Consolvo and Walker in 2003 [26], this compensation structure has not been widely adopted in the HCI literature [4]. Mushtag et al. compare three different micro-compensation structures but do not contrast their results with, e.g. a fixed compensation [27]. Although the use of micro-compensation warrants further investigation, we note that this compensation structure may not be applicable to all studies due to potential negative effects on the study's ecological validity. As highlighted by Mushtag et al., participant reactivity to micro-compensation may confound self-reports in studies focusing on participant affect. Stone et al. warn of using excessive financial incentives, which could attract participants solely interested in the monetary reward rather than participating in the study [28].

Recommendation 3 Avoid excessive financial compensation and consider the use of micro-compensation when applicable.

The literature on the ESM has also explored a number of extrinsic motivation techniques besides financial compensation, with promising results. Hsieh et al. show that providing participants with visual information on their provided self-reports increased participant adherence by 23% over a 25 day period (study with desktop users) [6]. The visual feedback provided by Hsieh et al. allowed participants to explore their prior answers on questions related to interruption or mood. The authors state that such visualisations "*makes the information personally relevant and increases the value of the study to participants*" [6]. Van Berkel et al. studied the effect of gamification (e.g. points, leaderboard) on participant responses in a between-subject study. Their results show that participants in the gamified condition significantly increased both their response quality (quantified through crowd-evaluation) and their number of provided responses as compared to the participants in the non-gamified condition [7].

Recommendation 4 Include interactive feedback mechanisms in the study protocol to keep participants engaged and motivated.

3 Study Adherence

Participant adherence to protocol, i.e. the degree to which the questionnaire notifications are opened and answered, is critical in ensuring an informative study outcome. In Experience Sampling, study adherence is typically quantified as 'response rate' or 'compliance rate', defined as the "*number of fully completed questionnaires divided by the number of presented questionnaires*" [4]. Unsurprisingly, studies typically report a decrease in study adherence over time, see for example [29–31]. As researchers can expect a decrease in participant adherence over time, it is key to consider the trade-offs when designing a longitudinal study. Balancing the number of daily questionnaires, number of questionnaire items, questionnaire scheduling,

and duration of the study, as well as other factors such as participant compensation and availability, in accordance with the research question is key. A number of studies have aimed to systematically study the effect of these variables, see, e.g., a recent study by Eisele et al. on the effect of notification frequency and questionnaire length on participant responses [32], or Van Berkel et al.'s investigation on the effect of notification schedules [31]. We argue that any researcher should consider these study parameters in relation to their research question and population sample. As such, there is not one study configuration that would be applicable to every study. Below, we outline some of the decisions that can motivate the balancing of these variables.

3.1 Questionnaire Scheduling

The literature describes three global techniques for questionnaire scheduling: signal contingent, interval contingent, and event contingent [33]. In a signal contingent schedule configuration, notification arrival is randomised over the course of a given timespan. In an interval contingent configuration, notification schedules follow a predefined interval, for example every other hour between 08:00 and 17:00. For event contingent configurations a predefined event is determined which triggers the notification (typically as recognised by the questionnaire system, but can also refer to a 'detection' by the participant) [4, 33–35]. The use of an event-based notification system enables more advanced study designs, and allows researchers to optimise the moment of data collection to contexts which are most relevant.

In a direct comparison between the three aforementioned scheduling techniques, results indicate that an interval-informed event contingent schedule, in which questionnaire notifications are presented upon smartphone unlock with a maximum number per given timespan, result in fewer total notifications sent but a higher overall number of completed responses as compared to a signal or interval contingent schedule [31]. Kapoor & Horvitz use contextual information to predict participant availability and find that using such a predictive model outperforms randomised scheduling in terms of identifying the availability of participants [36]. Church et al. recommend researchers to adjust the questionnaire schedule to match the participant's schedule [26]. Rather than imposing an identical start and end time on all participants, this approach would allow for custom start and end times, e.g. in the case of nightshift workers. Other work has explored more active-based scheduling techniques, where the presentation questionnaires are determined based on the participant's current contextual information. For example, Rosenthal et al. calculate individualised participant interruptibility costs [37], Mehrotra et al. expand on this through the notion of interruptibility prediction models [38], and Van Berkel et al. show that contextual information such as phone usage can be used to schedule questionnaires at opportune moments [39].

Regardless of the chosen scheduling approach, the timing of questionnaires can have a significant impact on participants' ability to respond to a questionnaire and

therefore the respective data being collected. The aforementioned scheduling techniques all have their own strengths and weaknesses. Signal contingent scheduling (i.e. randomised) can be used to capture participants spontaneous (psychological) states but can be skewed towards commonly occurring events. An interval contingent configuration is useful to capture events which are expected to occur regularly and provides a consistent sampling strategy which allows for the modelling of time as a factor in relation to the answers provided by the participant. Due to the regular schedule with which notifications are presented, it increases the risk of (over)sampling the same event (e.g. start of a lecture). Finally, event contingent configurations are useful for capturing isolated or infrequently occurring events that can be detected either through sensor data or manually by the participant. Event-based schedules can result in an incomplete view of the participant's life if the event of interest only occurs in a limit variety of contexts [40].

Recommendation 5 Carefully consider the effect of the chosen questionnaire scheduling approach on the selection of participant responses.

3.2 Study Duration

The literature on ESM study design has recommended roughly similar maximum durations for ESM studies, e.g. a minimum duration of one week [1], two weeks [39], and two–four weeks [28]. Determining an appropriate study duration is a careful consideration that involves a variety of factors such as the frequency with which the phenomenon of interest occurs, the required effort to complete the questionnaire, and expected levels of motivation among the participant sample.

Researchers interested in longitudinal studies of extensive duration, e.g. months or years, will find that participants are likely unable or unwilling to repeatedly answer a set of questionnaires for the duration of the study. Given the extensive participant burden in ESM studies, we advise against the collection of self-reports across the entire duration of studies of this duration. Instead, researchers should consider the collection of manual responses for a (number of) period(s) within the duration of the entire longitudinal study—embedding the ESM within a larger study design. As such, researchers can combine the insights obtained through frequent ESM questionnaires with the information gained from repeated data collection over an extensive period of time. This approach, which has been called as 'wave-based' experience sampling, has been successfully employed in emotion research in a decade-long study consisting of three one-week sampling periods investigating the effect of age on emotion [41]. Similarly, already in 1983 Savin-Williams & Demo ran a one-week ESM study with a cohort of participants enrolled in a six-year longitudinal study [42].

The use of modern mobile devices allows researchers to passively collect an extensive amount of sensor data from study participants [9]. This data is collected unobtrusively and without additional burden to the participant, and can provide additional insights to the researcher. The unobtrusive nature of this data collection stands in stark

contrast to the continuous effort required from participants in human contributions and can provide a continuous long-term data stream simply not feasible with manual data collection. As such, we recommend that researchers interested in extensive longitudinal studies combine both continuous passive sensing with intermittent periods of extensive questionnaire collection. Recent development work shows the possibility of changing ESM questionnaire schedules throughout the study period [43], enabling the possibility of intermittent periods of questionnaires.

From a participant perspective, being enrolled in a longitudinal study makes it easy to forget that sensor data is being collected. We stress that, given the potential sensitive nature of the unobtrusively (naturally following participant's informed consent) collected sensor data, researchers should aim to remind participants of any ongoing data collection. A practical approach for this in the context of smartphone-based studies is the continuous display of an icon in the smartphone's notification bar, reminding participants of their enrolment in the study and the active data collection [18]. Researchers have also allowed participants to temporarily halt data collection, see, e.g., Lathia et al. in which participants can (indefinitely) press a button to pause data collection for 30 min [40].

Recommendation 6 Combine longitudinal passive sensing with focused periods of ESM questionnaires to obtain both long-term and in-depth insights.

4 Response Reliability

A core idea behind the introduction of the ESM was to increase the reliability of self-report data by reducing the time between an event of interest and the moment when a participant provides data on this event, thus reducing reliance on a participant's ability to recall past events [1]. Although this approach has been widely embraced in a number of disciplines, recent work points out that the quality of participant data in ESM studies cannot be expected to be consistently of high reliability [18]. This is an important concern for longitudinal studies, as response quality reliability typically degrades over time. As such, recent work in the HCI community has explored techniques to infer and improve the reliability of participant responses. Here, we discuss the use of the crowd, quality-informed scheduling techniques, and the application of additional validation questions to infer response quality.

4.1 Use of the Crowd

Although the ESM traditionally collects data on observations or experiences as captured by participants individually, recent work has drawn out creative ways of combining the contributions of multiple individuals to increase the reliability of the collected data.

One strain of work has explored the use of 'peers' to obtain multiple datapoints on one individual. Using this approach, which has been labelled as 'Peer-MA' [44], a selected number of the participant's peers report what they believe to be the participant's current state with regard to the concept of interest. As described by Berrocal & Wac, this approach "*has the potential to enrich the self-assessment datasets with peers as pervasive data providers, whose observations could help researchers identify and manage data accuracy issues in human studies*" [44]. Chang et al. show how the use of peer-based data collection can also increase the quantity of the data collected [45]. By recruiting a sufficiently large (and motivated) network of participant peers, researchers may be able to distribute the burden of questionnaire notifications and thereby sustain data input for a more extensive period of time—increasing the prospective of longitudinal ESM studies. A critical open question with regard to this novel approach is the assessment and interpretation of the contributions of peers and the potential biases introduced through, e.g., different peer-relationships and the (absence of) peer physical presence.

In contrast to the aforementioned perspective in which the crowd contributions are focused on individuals, others have applied the crowd to increase the reliability of observations. For example, the aforementioned work by Van Berkel et al. not only asked participants to contribute a label regarding a given place, but also asked participants to judge the relevance of the contributions of others [7]. Based on these relevance labels, the quality of participant contributions can be quantified. Another example is the work by Solymosi et al., in which participants generated a map indicating a crowd's 'fear of crime' through repeated and localised experience sampling data collection [46]. A main advantage of this approach, in which the quality assessment is done by participants, is that the quality of contributions can be assessed without the need for a priori ground truth on the presented data. From a longitudinal study perspective, integrating crowd assessment into the study design may enable the study population to rotate, i.e. for participants to drop out and new participants to join, as study fatigue emerges.

Recommendation 7 Consider whether participant data can be validated or augment through the use of the crowd.

4.2 Quality-Informed Scheduling

Literature on questionnaire scheduling has primarily focused on participant availability following from a motivation to increase participant compliance. However, as pointed out by Mehrotra et al., an ill-timed questionnaire might lead participants to respond to a questionnaire without paying much attention, reducing the overall reliability of respondents' data [38]. In addition to increasing the quantity of responses, researchers have therefore also explored how the scheduling of questionnaires can affect the quality of participant responses. In the study by Van Berkel et al., participants completed a range of questions (working memory, recall, and arithmetic)

while contextual data was being passively collected [39]. Their results show that participants were more accurate when they were not using their phone the moment a questionnaire arrived. Optimising the quality of responses by not collecting data when participants are actively using their phone may, however, negatively effect the quantity of answered questionnaires. Previous work shows participants are more likely to *respond* to questionnaires (i.e. focused on response quantity) when questionnaires are presented upon phone unlock (as compared to randomised or interval-based schedules) [31].

Recommendation 8 Introduce intelligent scheduling techniques to avoid interrupting participants when they do not have time to respond.

4.3 Validation Questions

Here, we discern two types of validation questions: explicitly verifiable questions (also known as ground truth questions) and reflective questions.

In order to assess the reliability and effort of online study participants, work on crowdsourcing has recommended the use of 'explicitly verifiable questions', also known as 'golden questions' [47]. These explicitly verifiable questions are often—but not always—quantitative in nature, relatively easy to answer, and the responses can be automatically assessed to be correct or incorrect. For example, Oleson et al. asked crowdworkers to verify whether a given URL matched with a given local business listing [48]. Kittur et al. describe two main benefits of using these questions. First, explicitly verifiable questions allow researchers to easily identify and subsequently exclude from data analysis those participants who do not provide serious input. Second, by including these questions participants are aware of the fact that their answers will be scrutinised, which Kittur et al. hypothesise may "*play a role in both reducing invalid responses and increasing time-on-task*" [47].

Although widely used in crowdsourcing, the uptake of explicitly verifiable questions in ESM studies is thus far limited. A challenging aspect for the uptake of explicitly verifiable questions in longitudinal ESM studies is the need to provide participants with varying question content. This would require the creation of a question database, use of an existing and labelled dataset, or automated generation of verifiable questions (see, e.g., Oleson et al. [48]). An earlier ESM study with 25 participants included a simple, and randomly generated, arithmetic task as means of verification [39]. In this task, participants were asked to add two numbers together, both numbers were randomly generated between 10 and 99 for each self-report questionnaire. Results showed a remarkably high accuracy of 96.6%, which could be indicative of differences in motivation and effort between online crowdsourcing markets and the participant population often encountered in ESM studies. However, whether the motivation of the respective study population indeed differs between online crowdsourcing and ESM studies requires further investigation across multiple studies as well as evaluation across a wider variety of explicitly verifiable questions.

Another approach which has seen recent uptake is the creation of verifiable questions based on participant sensor data [39]. This includes, for example, passive data collection on the participants' smartphone usage and subsequently asking participants to answer questions on, e.g., the duration of their phone use. The answer to this question is verifiable, is variable (changes throughout the day), and often challenging to answer correctly. Assessing the correctness of participant answers does, however, also raise questions. In particular, answer correctness should not be quantified as a binary state as it is unlikely that answers are completely correct.

Recent work has also explored the use of 'reflective questions' in increasing the reliability of participant contributions. In this approach, participants reflect on earlier events while supported by earlier data points—either collected actively by the participant or passively through, e.g., smartphone sensors. Rabbi et al. introduce 'ReVibe', introducing assisted recall by showing participants an overview of their location, activity, and ambience during the past day [49]. Their results show a 5.6% increase in the participants' recall accuracy. Intille et al. propose an image-based approach, in which participants take a photo or short video and use this material to reflect on past experiences [50]. This concept was further explored by Yue et al., who note that the images taken by participants can also provide additional information and insights to researchers [51].

Recommendation 9 Consider including additional questions (verifiable, ground truth, or reflective) to increase the reliability of participant answers.

5 Analysing Longitudinal ESM Data

Longitudinal research faces a unique set of challenges in the analysis of participant data not typically encountered in short-term or lab-based studies. The longitudinal nature of a study can alter a participant's perception or understanding of the variables of interest, and may result in an increasing inequality of the number of responses between participants and different contexts. Here, we discuss these three challenges—respectively known as response shift, compliance bias, and contextual bias—as faced in the analysis of longitudinal ESM studies.

5.1 Response Shift

Response shift can either refer to an individual's change in meaning of a given construct due to re-calibration (a change in internal standards), re-prioritisation (change in values or priorities), or re-conceptualisation (change in the definition) [52, 53]. As studies often focus on the same construct(s) for the entire study period, participants may experience a shift in their assessment of this construct. As an example by Ring et al. illustrates: "*a patient rates her pre-treatment level of pain as 7 on a*

10-point pain scale. She subsequently rates her post-treatment level of pain as 3. This is taken to indicate that the treatment has caused an improvement of 4 points. However, if she retrospectively rates her pre-treatment pain as having been a 5, the actual treatment effect is 2. Likewise, if she retrospectively rates her pre-treatment pain as having been 10, the actual treatment effect is 7." [54]. Similar to a change in a participant's internal standards of a given construct, a participant may also evaluate various constructs as carrying higher or lower importance as compared to the onset of the study. By asking participants to rate the relative importance of individual constructs prior and following the study, the degree of re-prioritisation can be assessed. Finally, re-conceptualisation can occur when participants re-evaluate the meaning of a concept in relation to their personal circumstances. For example, a patient may re-conceptualise their quality of life, either following their recovery or by adjusting their perspective when confronted with a chronic disease.

A commonly used technique to identify the occurrence of response shift among participants is the 'thentest', also known as the 'retrospective pretest-posttest design'. At the end of the study, participants complete a posttest questionnaire immediately followed-up with a retrospective questionnaire asking participants to think back to their perception of a construct at the start of the study. By collecting these data points at almost the same time, participants share the same internal standards during questionnaire completion. Therefore, the mean change between these two questionnaires gives insight into the effect of time or treatment. For more details on the thentest, we refer to Schwartz & Sprangers's guidelines [55].

Recommendation 10 Include a thentest in the design of your study when participant perception of a given construct may change over the duration of the study.

5.2 Compliance Bias

Inevitable differences between participants' availability and motivation will result in a difference in the number of collected responses between participants. As such, the experience of response participants can skew the overall study results, a phenomenon known as compliance bias [20]. Participants with a higher than average response rate may have a more vested interest in responding to notifications, for example as they are personally affected by the phenomenon being investigated. Similarly, participants with a high or low response rate may have different psychological characteristics or simply different smartphone usage behaviours. It is not unlikely that these factors are a confounding factor in relation to the phenomenon being studied—capturing responses primarily from a subset of the study population may therefore decrease the reliability of the results. Although not widely reported, recent work that re-analysed four independent ESM studies finds substantial differences between study participants in the number of responses collected [20]. Researchers can reduce compliance bias by balancing data quantity between participants during the study through intelligent scheduling techniques—i.e. increasing the likelihood

that questionnaires will be answered by targeting notifications to arrive at a time and context suitable to the participant. Although this requires considerable infrastructure implementation and researcher ought to be careful not to introduce other biases, reducing compliance bias can increase the usefulness and reliability of a collected dataset.

Recommendation 11 Use intelligence scheduling techniques to improve response rates among low-respondents to balance response quantity between participants.

Recommendation 12 Analyse and report the differences between the number of participant responses post-data collection.

5.3 Contextual Bias

The schedule through which questionnaires are presented to participants, i.e. the chosen sampling technique, can significantly bias the responses of participants towards a limited number of contexts over time. As stated by Lathia et al., "*[...] time-based triggers will skew data collection towards those contexts that occur more frequently, while sensor-based triggers [...] generate a different view of behaviour than more a complete sampling would provide*" [40]. These concerns are amplified for longitudinal studies, in which researchers typically aim to cover a wide variety of contexts and identify longitudinal trends. If participants, however, only provide self-reports at contexts most convenient to them (e.g. by dismissing questionnaires arriving in the early morning or while at work), resulting data can be heavily skewed towards a limited number of contexts and therefore diminish the value of longitudinal data collection. The risk of contextual bias can be reduced by taking into account the context of completed self-reports in the scheduling of questionnaires. By considering to context in which individual participants have already answered questionnaires, researchers can diversity the context of collected responses.

Recommendation 13 Diversify the context of collected responses by scheduling questionnaires in contexts underrepresented in the existing responses of a participant.

6 Researcher Checklist

In order to increase a study's replicability and allow for a correct interpretation of presented results, it is critical that researchers report both the methodological choices and the outcomes of a presented study in detail. Current practice does not align with these standards, with prior work indicating that the majority of studies do not report on, e.g., the compensation of participants [4]. As compensation can affect participant motivation and compliance [28], it is important to report such metrics.

Building on previous work [4, 26, 56], we present a list of study design and result decisions which should be considered by researchers. We hope that this 'checklist' proves a useful starting point for researchers designing their ESM studies, as well as an overview of the variables we consider key in the reporting of the results of ESM studies.

Study design

1. Consider the target participant population and their potential interest in participation.
2. Determine the duration of the study, taking into account the study fatigue of prospective participants. Extensive longitudinal studies can combine longitudinal passive sensing with focused periods of self-report data collection.
3. Determine the most suitable questionnaire schedule in light of the respective trade-offs and benefits of scheduling techniques [31, 40].
4. Determine the length and frequency of questionnaire items, aiming for a short completion time of the questionnaire [18, 26].
5. Determine the timeout time for individual questionnaires, especially when sampling participant responses following a predetermined event as to reduce participant recall time.
6. Consider whether it is valuable to assess response shift in participant responses and consider including a thentest in the study design.
7. Consider the use of verifiable, ground truth, or reflective questionnaires to assess the quality of participant responses.
8. Consider whether it is important to achieve a balanced number of responses between participants. If desired, implement intelligent scheduling techniques to increase response rates among low-respondents.
9. Assess how participants can be best motivated to enrol and maintain compliance throughout the study period.
10. Assess the possibility of using the crowd to either assess or compare the contributions of participants.

Study results

1. Report both the number of participants who completed and dropped out of the study.
2. Report the (average) duration of participant enrolment.
3. Report the number of completed, dismissed, and timed-out responses.
4. Report the overall response rate.
5. Analyse and report the difference in response rate between participants [20].
6. Analyse and report any significant differences in the context of completed responses (e.g. time or location of completion) [40].
7. If relevant, analyse and report on the (differences in the) accuracy of participants on ground truth questions.
8. If relevant, analyse and report on any changes in the participants' perception of the study's construct, e.g. with the help of the thentest [55].

6.1 Overview of Recommendations

Finally, we present an overview of the recommendations introduced in this chapter in Table 1. The included references offer additional information on the motivation, methods, and guidelines with regard to the respective recommendation.

Table 1 Overview of recommendations with references for further reading

No.	Recommendation	References
1	Provide rich feedback regarding the study goals and the participants' contribution to those goals. Provide information throughout the study period	[1, 24]
2	Target participant recruitment to communities with a vested interest in the study outcomes	[21, 23]
3	Avoid excessive financial compensation and consider the use of micro-compensation when applicable	[27, 28]
4	Include interactive feedback mechanisms in the study protocol to keep participants engaged and motivated	[6]
5	Carefully consider the effect of the chosen questionnaire scheduling approach on the selection of participant responses	[31, 39, 40]
6	Combine longitudinal passive sensing with focused periods of ESM questionnaires to obtain both long-term and in-depth insights	[41, 42]
7	Consider whether participant data can be validated or augment through the use of the crowd	[7, 44, 45]
8	Introduce intelligent scheduling techniques to avoid interrupting participants when they do not have time to respond	[36–39]
9	Consider including additional questions (verifiable, ground truth, or reflective) to increase the reliability of participant answers	[49–51]
10	Include a thentest in the design of your study when participant perception of a given construct may change over the duration of the study	[55]
11	Use intelligence scheduling techniques to improve response rates among low-respondents to balance response quantity between participants	[20]
12	Analyse and report the differences between the number of participant responses post-data collection	[20]
13	Diversify the context of collected responses by scheduling questionnaires in contexts underrepresented in the existing responses of a participant	[40]

7 Future Trends

Since the introduction of the Experience Sampling Method in the late 1970s [1], its main use has been in the application of intensive but relatively short-term data collection (i.e. weeks rather than months). In this foundational work, Larson & Csikszentmihalyi describe a typical ESM study to have a duration of one week. Technological and methodological developments have had, and continue to have, a significant impact on how the ESM is used by researchers throughout their projects. For example, the introduction and widespread usage of smartphones has enabled researchers to collect rich contextual information [8, 9]. Similarly, researchers have come up with novel scheduling techniques to increase the sampling possibilities offered through the ESM. Following the impact of these developments on how the ESM is applied, we expect future innovations to increase the ability for researchers to apply the ESM in a longitudinal setting.

From a technological perspective, recent work has pointed to the further integration of self-report devices in the participants' daily life. This includes (stationary) devices physically located in a participant's home or work location [57], integration of questionnaires in mobile applications already frequently used by participants (e.g. messaging applications [58]), or through the use of (tangible) wearables [59, 60]. Although the effect of these alternative questionnaire delivery techniques on (sustained) response rate and input accuracy still needs to be explored in more detail, these alternative input methods can reduce participant strain as compared to a smartphone-based approach (retrieving phone, unlocking, opening a specific application, locking away the phone). Future studies can also consider the collection of questionnaires across multiple platforms, such as the use of a stationary device at home and work, combined with a mobile device or application for on-the-go.

Methodologically, a number of under-explored avenues may prove useful in enabling longitudinal ESM studies. In Sect. 3.2, we refer to 'wave-based' experience sampling, in which participants actively contribute only for a number of (discontinuous) periods within a larger duration consisting of passive sensing. Although already explored in the early days of the ESM [42], this approach has thus far not been extensively applied. Furthermore, although prior work shows the positive effect of including extrinsic motivators [6, 7], the studies were limited to weeks. Further works is required to study the impact of these incentives in longitudinal settings. Finally, we note that an extensive amount of work has explored ways to infer participant availability and willingness to answer a questionnaire, both within the scope of ESM research [38, 61] as well as the broader research on attention and availability [36, 62–64]. Translating these findings into practical and shareable implementations which can be readily used by other researchers remains a formidable challenge. Addressing this, e.g., by releasing the source code of these implementations, allows for experimentation with advanced scheduling techniques while simultaneously enabling research groups to validate, compare, and extend these scheduling algorithms.

Numerous open questions regarding the use of the ESM beyond a couple of weeks (e.g. covering months of active data collection) remain. In this chapter, we outlined both practical suggestions which are applicable to researchers *today* when designing their studies, as well as offer a number of potential areas for future work in the domain of longitudinal self-report studies.

8 Conclusion

The Experience Sampling Method has enabled researchers to collect frequent and rich responses from study participants. Enabled by the wide uptake of mobile devices, researchers can deliver a highly interactive and increasingly intelligent research tool straight into the hands of participants. Our overview shows that the introduction of smaller and more connected mobile hardware alone is not sufficient in enabling a push towards truly longitudinal studies. In order to extend the viable duration of ESM studies, further development of methodological practices is required. Investigating the effect of novel hardware solutions and study design configurations, both in the lab and in situ, will require a focused effort from the research community.

References

1. Larson R, Csikszentmihalyi M (1983) The experience sampling method. New Directions Methodol Soc Behav Sci
2. Csikszentmihalyi M, Larson R (1987) Validity and reliability of the Experience-Sampling method. J Nerv Ment Dis 175:526–536
3. Csikszentmihalyi M, Larson R, Prescott S (1977) The ecology of adolescent activity and experience. J Youth Adolescence 6:281–294
4. van Berkel N, Ferreira D, Kostakos V (2017) The experience sampling method on mobile devices. ACM Comput Surv 50(6):93:1–93:40
5. Smyth JM, Stone AA (2003) Ecological momentary assessment research in behavioral medicine. J Happiness Stud 4:35–52
6. Hsieh G, Li I, Dey A, Forlizzi J, Hudson SE (2008) Using visualizations to increase compliance in experience sampling. In: Proceedings of the 10th international conference on ubiquitous computing, UbiComp '08, (New York, NY, USA). ACM, pp 164–167
7. van Berkel N, Goncalves J, Hosio S, Kostakos V (2017) Gamification of mobile experience sampling improves data quality and quantity. In: Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies (IMWUT), vol 1, no 3, pp 107:1–107:21
8. Raento M, Oulasvirta A, Eagle N (2009) Smartphones: an emerging tool for social scientists. Sociol Methods Res 37(3):426–454
9. Ferreira D, Kostakos V, Schweizer I (2017) Human sensors on the move. Springer International Publishing, pp 9–19
10. van Berkel N, Goncalves J, Wac K, Hosio S, Cox AL (2020) Human accuracy in mobile data collection. Int J Hum-Comput Stud, p 102396
11. Westerink J, Ouwerkerk M, de Vries G, de Waele S, van den Eerenbeemd J, van Boven M (2009) Emotion measurement platform for daily life situations. In: 3rd international conference on affective computing and intelligent interaction and workshops, pp 1–6

12. Guha S, Wicker SB (2015) Spatial subterfuge: an experience sampling study to predict deceptive location disclosures. In: Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing, UbiComp '15, (New York, NY, USA). Association for Computing Machinery, pp 1131–1135
13. Heron KE, Smyth JM (2010) Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Brit J Health Psychol 15(1):1–39
14. Shaughnessy JJ, Zechmeister EB, Zechmeister JS (2011) Research methods in psychology. McGraw-Hill, New York
15. Armantier O, Topa G, Van der Klaauw W, Zafar B (2017) An overview of the survey of consumer expectations. Econ Policy Rev 23–2:51–72
16. Stein RE, Horwitz SM, Storfer-Isser A, Heneghan A, Olson L, Hoagwood KE (2008) Do pediatricians think they are responsible for identification and management of child mental health problems? Results of the AAP periodic survey. Ambulatory Pediatr 8(1):11–17
17. van Berkel N, Budde M, Wijenayake S, Goncalves J (2018) Improving accuracy in mobile human contributions: an overview. In: Adjunct proceedings of the ACM international joint conference on pervasive and ubiquitous computing, pp 594–599
18. van Berkel N (2019) Data quality and quantity in mobile experience sampling. Phd thesis, The University of Melbourne
19. Hektner JM, Schmidt JA, Csikszentmihalyi M (2007) Experience sampling method: measuring the quality of everyday life. Sage
20. van Berkel N, Goncalves J, Hosio S, Sarsenbayeva Z, Velloso E, Kostakos V (2020) Overcoming compliance bias in self-report studies: a cross-study analysis. Int J Hum-Comput Stud 134:1–12
21. Reiss S (2012) Intrinsic and extrinsic motivation. Teach Psychol 39(2):152–156
22. Eveleigh A, Jennett C, Blandford A, Brohan P, Cox AL (2014) Designing for dabblers and deterring drop-outs in citizen science. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '14, (New York, NY, USA). ACM, pp 2985–2994
23. Rotman D, Preece J, Hammock J, Procita K, Hansen D, Parr C, Lewis D, Jacobs D (2012) Dynamic changes in motivation in collaborative citizen-science projects. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, CSCW '12, (New York, NY, USA). ACM, pp 217–226
24. Measham TG, Barnett GB (2008) Environmental volunteering: motivations, modes and outcomes. Australian Geographer 39(4):537–552
25. Deci E, Ryan RM (1985) Intrinsic motivation and self-determination in human behavior. Springer, Berlin
26. Consolvo S, Walker M (2003) Using the experience sampling method to evaluate ubicomp applications. IEEE Pervas Comput 2:24–31
27. Musthag M, Raij A, Ganesan D, Kumar S, Shiffman S (2011) Exploring micro-incentive strategies for participant compensation in high-burden studies. In: Proceedings of the 13th international conference on ubiquitous computing, UbiComp '11, (New York, NY, USA). ACM, pp 435–444
28. Stone AA, Kessler RC, Haythomthwatte JA (1991) Measuring daily events and experiences: decisions for the researcher. J Personal 59(3):575–607
29. Shih F, Liccardi I, Weitzner D (2015) Privacy tipping points in smartphones privacy preferences. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, CHI '15, (New York, NY, USA). Association for Computing Machinery, pp 807–816
30. Tollmar K, Huang C (2015) Boosting mobile experience sampling with social media. In: Proceedings of the 17th international conference on human-computer interaction with mobile devices and services, MobileHCI '15, (New York, NY, USA). Association for Computing Machinery, pp 525–530
31. van Berkel N, Goncalves J, Lovén L, Ferreira D, Hosio S, Kostakos V (2019) Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. Int J Hum-Comput Stud 125:118–128
32. 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

33. Wheeler L, Reis HT (1991) Self-recording of everyday life events: origins, types, and uses. J Personal 59(3):339–354
34. Barrett LF, Barrett DJ (2001) An introduction to computerized experience sampling in psychology. Soc Sci Comput Rev 19(2):175–185
35. Bolger N, Davis A, Rafaeli E (2003) Diary methods: capturing life as it is lived. Ann Rev Psychol 54(1):579–616
36. Kapoor A, Horvitz E (2008) Experience sampling for building predictive user models: a comparative study. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '08, (New York, NY, USA). Association for Computing Machinery, pp 657–666
37. Rosenthal S, Dey AK, Veloso M (2011) Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In: Lyons K, Hightower J, Huang EM (eds) Pervasive computing. Springer, Berlin, Heidelberg, pp 170–187
38. Mehrotra A, Vermeulen J, Pejovic V, Musolesi V (2015) Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In: Adjunct proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2015 ACM international symposium on wearable computers, UbiComp/ISWC'15 Adjunct, (New York, NY, USA). Association for Computing Machinery, pp 723–732
39. 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 ACM SIGCHI conference on human factors in computing systems, pp 51:1–51:12
40. Lathia N, Rachuri KK, Mascolo C, Rentfrow PJ (2013) Contextual dissonance: design bias in sensor-based experience sampling methods. In: Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing, UbiComp '13, (New York, NY, USA). Association for Computing Machinery, pp 183–192
41. Carstensen LL, Turan B, Scheibe S, Ram N, Ersner-Hershfield H, Samanez-Larkin GR, Brooks KP, Nesselroade JR (2011) Emotional experience improves with age: evidence based on over 10 years of experience sampling. Psychol Aging 26(1):21–33
42. Savin-Williams RC, Demo DH (1983) Situational and transituational determinants of adolescent self-feelings. J Personal Soc Psychol 44(4):824
43. Bailon C, Damas M, Pomares H, Sanabria D, Perakakis P, Goicoechea C, Banos O (2019) Smartphone-based platform for affect monitoring through flexibly managed experience sampling methods. Sensors 19(15):3430
44. Berrocal A, Wac K (2018) Peer-vasive computing: leveraging peers to enhance the accuracy of self-reports in mobile human studies. In: Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers. ACM, pp 600–605
45. Chang Y-L, Chang Y-J, Shen C-Y (2019) She is in a bad mood now: leveraging peers to increase data quantity via a chatbot-based ESM. In: Proceedings of the 21st international conference on human-computer interaction with mobile devices and services, MobileHCI '19 (New York, NY, USA). Association for Computing Machinery
46. Solymosi R, Bowers K, Fujiyama T (2015) Mapping fear of crime as a context-dependent everyday experience that varies in space and time. Legal Criminol Psychol 20(2):193–211
47. Kittur A, Chi EH, Suh B (2008) Crowdsourcing user studies with mechanical turk. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '08, (New York, NY, USA). Association for Computing Machinery, pp 453–456
48. Oleson D, Sorokin A, Laughlin G, Hester V, Le J, Biewald L (2011) Programmatic gold: targeted and scalable quality assurance in crowdsourcing. In: Workshops at the Twenty-Fifth AAAI conference on artificial intelligence
49. Rabbi M, Li K, Yan HY, Hall K, Klasnja P, Murphy S (2019) Revibe: a context-assisted evening recall approach to improve self-report adherence. In: Proceedings of the ACM Interaction Mobile Wearable Ubiquitous Technology, vol 3
50. Intille S, Kukla C, Ma X (2002) Eliciting user preferences using image-based experience sampling and reflection. In: CHI '02 extended abstracts on human factors in computing systems, CHI EA '02, (New York, NY, USA). Association for Computing Machinery, pp 738–739

51. Yue Z, Litt E, Cai CJ, Stern J, Baxter KK, Guan Z, Sharma N, Zhang GG (2014) Photographing information needs: the role of photos in experience sampling method-style research. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '14, (New York, NY, USA). Association for Computing Machinery, pp 1545–1554
52. Sprangers MA, Schwartz CE (1999) Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med 48(11):1507–1515
53. Schwartz CE, Sprangers MA, Carey A, Reed G (2004) Exploring response shift in longitudinal data. Psychol Health 19(1):51–69
54. Ring L, Höfer S, Heuston F, Harris D, O'Boyle CA (2005) Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patients. Health Qual Life Outcomes 3(1):55
55. Schwartz CE, Sprangers MA (2010) Guidelines for improving the stringency of response shift research using the thentest. Qual Life Res 19(4):455–464
56. Christensen TC, Barrett LF, Bliss-Moreau E, Lebo K, Kaschub C (2003) A practical guide to experience-sampling procedures. J Happiness Stud 4(1):53–78
57. Paruthi G, Raj S, Gupta A, Huang C-C, Chang Y-J, Newman MW (2017) Heed: situated and distributed interactive devices for self-reporting. In: Proceedings of the 2017 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2017 ACM international symposium on wearable computers, UbiComp '17, (New York, NY, USA). Association for Computing Machinery, pp 181–184
58. Gong Q, He X, Xie Q, Lin S, She G, Fang R, Han R, Chen Y, Xiao Y, Fu X et al (2018) LBSLAB: a user data collection system in mobile environments. In: Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers, UbiComp '18, (New York, NY, USA). Association for Computing Machinery, pp 624–629
59. Adams AT, Murnane EL, Adams P, Elfenbein M, Chang PF, Sannon S, Gay G, Choudhury T (2018) Keppi: a tangible user interface for self-reporting pain. In: Proceedings of the 2018 CHI conference on human factors in computing systems, CHI '18 (New York, NY, USA). Association for Computing Machinery
60. Hernandez J, McDuff D, Infante C, Maes P, Quigley K, Picard R (2016) Wearable ESM: differences in the experience sampling method across wearable devices. In: Proceedings of the 18th international conference on human-computer interaction with mobile devices and services, MobileHCI '16, (New York, NY, USA). Association for Computing Machinery, pp 195–205
61. Liono J, Salim FD, van Berkel N, Kostakos V, Qin AK (2019) Improving experience sampling with multi-view user-driven annotation prediction. In: IEEE international conference on pervasive computing and communications (PerCom, pp 1–11
62. Pielot M, Vradi A, Park S (2018) Dismissed! a detailed exploration of how mobile phone users handle push notifications. In: Proceedings of the 20th international conference on human-computer interaction with mobile devices and services, MobileHCI '18 (New York, NY, USA). Association for Computing Machinery
63. Visuri A, van Berkel N, Okoshi T, Goncalves J, Kostakos V (2019) Understanding smartphone notifications' user interactions and content importance. Int J Hum-Comput Stud 128:72–85
64. Weber D, Voit A, Auda J, Schneegass S, Henze N (2018) Snooze! investigating the user-defined deferral of mobile notifications. In: Proceedings of the 20th international conference on human-computer interaction with mobile devices and services, MobileHCI '18 (New York, NY, USA). Association for Computing Machinery