The Experience Sampling Method on Mobile Devices

NIELS VAN BERKEL, The University of Melbourne
DENZIL FERREIRA, University of Oulu
VASSILIS KOSTAKOS, The University of Melbourne

The Experience Sampling Method (ESM) is used by scientists from various disciplines to gather insights into the intra-psychic elements of human life. Researchers have used the ESM in a wide variety of studies, with the method seeing increased popularity. Mobile technologies have enabled new possibilities for the use of the ESM, while simultaneously leading to new conceptual, methodological, and technological challenges. In this survey, we provide an overview of the history of the ESM, usage of this methodology in the computer science discipline, as well as its evolution over time. Next, we identify and discuss important considerations for ESM studies on mobile devices, and analyse the particular methodological parameters scientists should consider in their study design. We reflect on the existing tools that support the ESM methodology and discuss the future development of such tools. Finally, we discuss the effect of future technological developments on the use of the ESM and identify areas requiring further investigation.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; **User studies**; **Field studies**; *Ubiquitous and mobile computing*; *Ubiquitous and mobile computing design and evaluation methods*; *Smartphones*; *Mobile phones*; Mobile devices; • **General and reference** → **Surveys and overviews**;

Additional Key Words and Phrases: Experience sampling method, mobile devices, smartphone, data collection, methodology, qualitative data, sensor, in situ, ecological momentary assessment, ambulatory assessment, ESM, EMA

ACM Reference format:
Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *ACM Comput. Surv.* 50, 6, Article 93 (December 2017), 40 pages.
https://doi.org/10.1145/3123988

1 INTRODUCTION

Literature from as early as the 1900s already shows a scientific interest in the systematic collection of information about daily life (Bevans 1913). The advent of personal technologies in the late 1970s gave rise to the Experience Sampling Method (ESM) (Larson and Csikszentmihalyi 1983), aimed at measuring the behaviour, thoughts, and feelings of participants throughout their

This work is partially funded by the Academy of Finland (Grants 276786-AWARE, 286386-CPDSS, 285459-iSCIENCE, 304925-CARE), the European Commission (Grant, 6AIKA-A71143-AKAI), and Marie Skłodowska-Curie Actions (645706-GRAGE).
Authors' addresses: N. van Berkel, The University of Melbourne, Department of Computing and Information Systems, Parkville 3010, Melbourne, Australia; email: n.vanberkel@student.unimelb.edu.au; D. Ferreira, University of Oulu, Center for Ubiquitous Computing, Pentti Kaiteran katu 1, PO Box 4500, FI-90014 Oulu, Finland; email: denzil.ferreira@oulu.fi; V. Kostakos, The University of Melbourne, Department of Computing and Information Systems, Parkville 3010, Melbourne, Australia; email: v.kostakos@unimelb.edu.au.

https://doi.org/10.1145/3123988

day-to-day activities. With experience sampling, data is collected through self-reports provided by participants, much like in traditional diary studies. But unlike diary studies, participants are proactively triggered at various points throughout the day (Larson and Csikszentmihalyi 1983). The method employs *in situ* self-reports for systematic in-context data collection, as well as close to the onset or completion of the investigated phenomena. This reduces reliance on the participants' long-term memory to reconstruct past events or experiences, and data collection can be primed to those particular events that are of interest to the researcher. As described by Myin-Germeys et al. (2009), "[t]he Experience Sampling Method (ESM) allows us to capture the film rather than a snapshot of daily life reality of patients."

Today, the availability of personal mobile devices enables widespread deployment of mobile phones as a research tool (Raento et al. 2009). Combining these technologically powerful and widely available devices with the ESM research method allows for increasingly more powerful and insightful research probes. Most prominently, utilising the sensors on mobile devices allows for the inference of many elements of the participants' context. As a result, it becomes possible to adjust both the presentation and content of the presented self-reports following the requirements of the researcher(s) and the participants' availability and context. In addition, the collection of information using mobile devices over traditional (analogue) techniques allows researchers to keep a closer eye on the data provided by their participants as it is being collected – for example, by assessing their adherence to the study protocol.

1.1 Goals and Method

The use of mobile devices over the traditional pen-and-paper approach provides researchers with new experimental opportunities, but also gives rise to new challenges on conceptual, methodological, and technological levels. In this survey, we discuss these challenges in light of recently published ESM studies and through a synthesis of related work. Despite the large number of studies employing the ESM on mobile devices, no comprehensive literature overview exists on the use of ESM on such devices. Our literature review aims to cover this gap and provide researchers with practical advice on running ESM studies on mobile devices, and in particular smartphones. A recent survey on the use of experience sampling for behaviour research identifies recent developments in mobile technologies for ESM (Pejovic et al. 2016). Our review extends this work by providing a structured overview of current ESM practices and developments in a larger time-frame. In addition, we provide advice on methodological decisions and identify both future trends and requirements in the scope of mobile experience sampling. Ultimately, our work develops a systematic framework for conducting and reporting ESM studies in the era of widespread smartphone usage.

We conduct a systematic literature search on the use of the ESM and the methodologically related Ecological Momentary Assessment (EMA), covering three established digital libraries from the computer science domain. We then select and analyse those papers involving mobile devices. Following this collection of the literature, we analyse several study parameters that vary across studies. Using these study parameters, we aim to establish an overview of potential study configurations and recommended practices for other researchers. In addition, we evaluate currently available software tools and compare their functionality to the study parameters following from the literature review. Finally, we draw on the findings of our literature review to identify future developments in the area of experience sampling.

We wish to immediately acknowledge that, despite our best efforts, our survey may not be all inclusive. More specifically, there may be literature employing ESM/EMA which may not be indexed on the selected databases. On the other hand, we believe that the selected databases cover most research venues in the computer science domain. Similarly, the list of software tools may not be all inclusive and we encourage the reader to contact the authors if we omitted any important tool.

1.2 Organisation of the Survey

Section 2 provides a background of the ESM, offering an historical overview as well as acknowledging the potential challenges and alternatives to the usage of the ESM. Section 3 presents the literature review and discusses the identified study parameters. In Section 4, we discuss the different approaches with regards to data collection and give practical advice for researchers on the implementation of these parameters. Section 5 focuses on the reporting of ESM study results. In Section 6, we provide an overview of available ESM tools for researchers, and identify their strengths and weaknesses in relation to our previous findings. Finally, we draw our expectations of future developments in the domain of experience sampling in Section 7.

2 BACKGROUND

The ESM is used by researchers from a wide variety of academic disciplines to increase their understanding of human behaviour. We present a basic overview to the methodology and its applications (Section 2.1), providing a starting point for those unfamiliar with the ESM. This is followed by a historical perspective on the use of the ESM (Section 2.2) considering the close methodological ties to the diary study. The historical overview is followed by an impression of the technological developments that have come to shape the ESM as it is used today (Section 2.3). Finally, we discuss the challenges encountered by researchers when using the ESM in studies (Section 2.4), as well as some of the most well-known alternatives to this methodology (Section 2.5).

2.1 Experience Sampling

The ESM is employed by researchers with an interest in studying human behaviour. Participants of ESM studies are requested to provide self-reports on their activities, emotions, or other elements of their daily life multiple times per day. These self-reports are provided by answering a short, usually identical, questionnaire upon receiving a notification (e.g., smartphone notification, text message). Following the collection of self-reports across multiple days and among various participants provides researchers with a profound insight into the studied daily life experience(s). As participants record their answers in their natural environment, as opposed to in a laboratory, these *in situ* self-reports provide a more accurate representation of the participants' natural behaviour.

One of the initial studies that applied this methodology investigated adolescent activity and experience (Csikszentmihalyi et al. 1977). Participants were asked to complete self-reports at random points throughout the day for the duration of a week. These self-reports consisted of questions such as "What was the main thing you were doing?" and "Were you in control of your actions?", combined with various Likert scale questions regarding the quality of the participants' interaction. With a combined total of 753 self-reports, the researchers could construct the activities on which adolescents spend their time and how they experienced these activities.

To date, the methodology has been applied in a wide variety of fields with review papers on the use of this method in areas such as *mood disorder and dysregulation* (Ebner-Priemer and Trull 2009), *substance usage* (Shiffman 2009), and *binge eating* (Haedt-Matt and Keel 2011). Furthermore, the ESM is also actively used in computer science—most notably in the discipline of human-computer interaction (HCI). Consolvo and Walker (2003) discuss the possibilities of the ESM for ubiquitous computing research, specifically touting its ability to evaluate a technological artefact. Given their nature of blending with the users' environment and being available 'on the go', ubiquitous applications are notoriously challenging to evaluate. Laboratory studies are therefore unable to capture the user experience of such applications during actual usage, something much more feasible using the ESM.

We provide two examples to highlight the possibilities of ESM for computer scientists. In our study on smartphone usage (van Berkel et al. 2016b), participants answered a single binary question each time they started using their phone. In addition, sensor logging was used to infer usage of the device. This combination of human and sensor data allowed us to quantify smartphone interaction sessions. In a different study, Church et al. (2014) employ the ESM to gain insights into daily information needs and how these needs are addressed in both mobile and non-mobile settings. For both examples, researchers rely on the participants' self-reports to construct an understanding of their motives and needs.

2.2 Historical Perspective

The methodological constructs for the later development of the ESM were provided by diary studies. The diary study is a popular research method that captures a wide variety of aspects on daily human life. In a diary study, participants are asked to fill out (daily) self-reports regarding their experiences, activities, and feelings—often focusing on a selection of particular event(s) or feeling(s). Bevans (1913) already applied a variation of the diary study technique as early as 1913 to investigate the daily patterns of working men.

Diary studies allow for longitudinal data collection outside the laboratory, in the participants' natural environment, and outside of public life. Consequently, a diary study does not rely on direct observational methods that may skew the collected data (Bolger and Laurenceau 2013). As such, diary studies are highly suitable to study commonly occurring events in daily life. These events determine our lives to a great extent, since, as stated by Wheeler and Reis (1991): "little experiences of everyday life fill most of our waking time and occupy the vast majority of our conscious attention."

Two types of diary studies can be distinguished: *feedback* studies and *elicitation* studies (Carter and Mankoff 2005). In a feedback study, participants answer a set of questions at a predetermined timeslot or event. In an elicitation study, participants capture media (e.g., photographs) when an event occurs and discuss the collected media at a later point in time with the researcher(s). Both types of diary studies have their shortcomings, most prominently the limited reach of data collection (feedback studies) and humans' inability to reliably reconstruct past events (elicitation studies) (Iida et al. 2012). A common drawback in the diary study lies in the fact that study participants must remember and be sufficiently motivated to complete these diary entries. In addition, participants can fabricate (analogue) diary entries at a later point in order to counterfeit study compliance (Broderick et al. 2003).

The ESM relies heavily on the concept of the diary study, but aims to mitigate some of the aforementioned drawbacks. This follows from the need to measure the behavioural and intrapsychic aspects of (day-to-day) human life in a more *reliable* and *consistent* approach than possible with the diary study. The ESM fulfils this requirement by reducing reliance on the participants' ability to accurately reproduce earlier experiences, minimising cognitive bias. Cognitive bias has been shown to reduce the validity of collected data (Iida et al. 2012). The ESM aims to preserve the *ecological validity* during a study, which is defined as "the occurrence and distribution of stimulus variables in the natural or customary habitat of an individual" (Hormuth 1986).

In a diary study, the time between onset of, and a participant's reflection on a studied phenomenon is shorter when compared to traditional debriefing interviews occurring at the end of a study, reducing retrospection bias. An example of such a bias is "*Rosy retrospection*," in which study participants evaluate past events (e.g., a vacation trip) more positively retrospectively than they did during the actual event (Mitchell et al. 1997). The ESM significantly differs from diary studies in the way questions are delivered to study participants. In diary studies, participants receive question sets passively and at their own initiation, while the ESM actively prompts participants to answer

a question set. This enables a further reduction of the time gap in-between onset of and reflection on the studied phenomena.

As previously mentioned, Csikszentmihalyi et al. (1977) conducted one of the first ESM studies while analysing adolescent activity and experience in the late 1970s. The authors instructed participants to complete a paper self-report form upon each incoming pager signal. The self-report contained questions regarding the participants' context and subjective state at that particular moment. Participants were required to carry with them both an electronic pager and paper questionnaires.

Following the introduction of the PDA (Personal Digital Assistant), many studies have adopted such digital devices. For example, Taylor et al. (1990) and Totterdell et al. (1992) summarise early explorations into applying the ESM on PDAs in the early 90's. Back then, the main concerns of these researchers regarded the PDAs limited battery life, unstable data storage, and high device costs—in addition to the risk of introducing novelty bias with the new technology. These concerns ensured the continued usage of pagers in combination with pen-and-paper data collection. The introduction and widespread usage of smartphones and other mobile devices has eventually replaced the use of PDAs.

2.3 Digital Evolution of the ESM

The ESM, originally used on electronic pagers (Csikszentmihalyi et al. 1977), has traditionally been quick to adapt to new device types as they became available. This uptake has allowed the ESM to evolve alongside these new developments—further exploiting its advantages over alternate methodologies. Technological advances have given rise to new possibilities for the ESM (Barrett and Barrett 2001). Advantages of smartphone-based ESM studies include:

- *Improved Data Quality through Validation.* Unusually short response completion time could indicate a participant 'skipping' through the questions. Questionnaires can also be configured to expire after a set time period after which they are withdrawn. Using a paper-based approach, participants may reconstruct or fabricate data after the intended time period has expired (Broderick et al. 2003).
- *Context Reconstruction.* Utilising the sensors of the mobile device employed during the study, a wide variety of sensors and data modalities become available. This enables researchers to collect not only the explicit answers of the participants, but also the *context* in which these answers are provided.
- *Real-Time Study Status.* Researchers can receive and analyse study data in real-time. This allows researchers to identify any possible errors while the study is still running, such as identification of participants not adhering to the study protocol, technical problems, or participants dropping out of the study.
- *Advanced Question Logic.* Presented questionnaires can embed logical constructs. This way, questions can depend on previous input from the participant or the participant's current context. This allows for potentially richer data collection and a reduction of unnecessary questions (thus reducing participant strain).
- *Rich Media Collection.* Participants can collect visual and auditory material explicitly. This augments the collected data, allowing for a richer qualitative insight into the daily experiences of the study participants.

Technological developments have allowed ESM studies to increase in both the attainable survey complexity and richness of the captured context. This leads to new possibilities in data collection, as shown in the quadrants of Figure 1. These opportunities of data collection, as represented in Figure 1, are further discussed in Section 7.2.

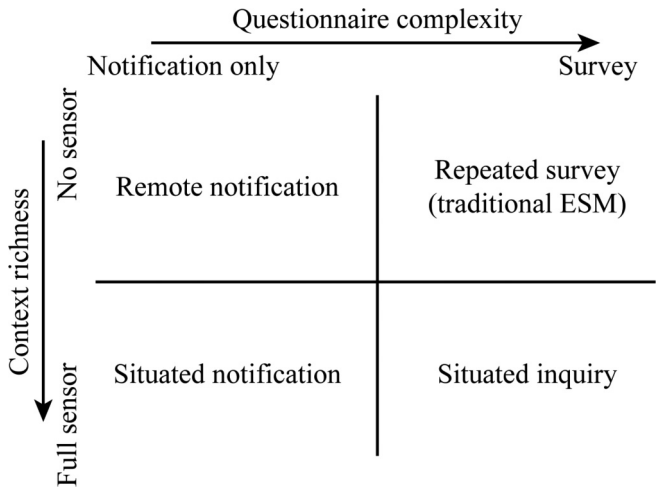

Fig. 1. Matrix overview of ESM opportunities following an increase in question complexity and context richness.

2.4 Potential Challenges for the ESM

Studies employing the ESM face a number of potential difficulties. In the following text, we outline several concerns based on findings from the literature (Barrett and Barrett 2001, Scollon et al. 2003), and provide alternative or updated point of views.

— *Participant Burden.* Answering ESM notifications throughout the day is a considerable burden for participants. Study factors such as the number of questions, daily alerts, and question types can be configured to reduce this burden (Klasnja et al. 2008). By assessing the participants' context in the background, it is possible to reduce interruption to those moments relevant to the researcher. The literature also describes the requirement for participants to visit the laboratory throughout the study to transfer their data (Barrett and Barrett 2001). While the widespread availability of Internet connectivity removes this requirement, eliminating regular visits to the researcher's laboratory may also reduce participant motivation as there is less opportunity for the creation of a mutual research alliance between researcher and study participant.

— *Participant Retention.* Because of the frequent action required by participants, study dropout rates are generally high. Literature suggest to adjust compensation rates accordingly (Barrett and Barrett 2001) or for researchers to make the collection of data intrinsically rewarding to the participants (Hektner et al. 2007). This requires additional software implementation, and is highly dependent on the nature of the collected data. We give the example of participants' own self-reports being used in discussion with their professional caregiver, providing a direct benefit to data collection (van Berkel et al. 2016a). Another example is found in the visualisation of self-reports—allowing participants to measure progress and answer their personal (quantified-self) questions (van Berkel et al. 2015). Lastly, the attitude of research assistants in communication with participants has also been discussed as an influencing factor on participant retention (Conner Christensen et al. 2003). Section 4.5 contains concrete advice on this matter.

— *Programming*. A mobile application or service is often required to run an ESM study. In the current smartphone era, freely available software packages (e.g., AWARE (Ferreira et al. 2015b), PACO (PACO 2016), and Purple Robot (Karr 2015)) have decreased the costs and required level of technical knowledge to run an ESM study. Yet, the availability of accessible tools for researchers without programming skills is still lacking (Raento et al. 2009, Rough and Quigley 2015)—acting as a deterrent for researchers to carry out mobile experience sampling studies.
— *Study Equipment*. Cost, software compatibility, and device characteristics have all been considered as important factors when buying study equipment (Conner Christensen et al. 2003). In addition, literature highlights the difficulties in managing study devices and ensuring that the devices are not misused (Barrett and Barrett 2001). Market penetration of mobile devices is currently so high (GSMA 2016), that researchers should consider utilising the participants' own devices during the study. This also removes the need for participants to carry an extra device, contributing to a more natural experience for the participant and reducing novelty and learning effects introduced by the study equipment.
— *Platform Heterogeneity*. The large number of different devices produced by hardware manufacturers has led to a diversified product landscape, with variations in both hardware and software components (e.g., screen size, CPU, memory, operating system). This introduces unique limitations for each configuration, and requires flexible software to support a wide range of devices.
— *Data Quality*. Relying on participant provided data introduces several concerns with regards to the quality of collected data. This includes a lack of participant answers, deliberately wrong or careless answers, response shift (changes in the participants' internal standard when rating events or emotions), and participant reactivity. Participant reactivity is the occurrence of participant behaviour adjustments as the result of knowing that they are being observed—a problem that is strongly present in observational studies, but has been found to be limited in ESM studies (Barrett and Barrett 2001). To increase data quality, researchers can clean their data to filter out any suspicious information. We further discuss the aforementioned concerns regarding data quality, including how to identify and resolve these potential problems, in Section 4.6.

2.5 Critique and Alternatives to the ESM

Several papers have published critiques on the ESM as a method for measuring the intra-psychic aspects of human life. One category of criticism addresses the fact that the ESM measures data 'in the moment', therefore being unable to capture participant reflection on the measured phenomenon (Isaacs et al. 2013). In case the researcher is interested in investigating reflection, other research methods are indeed preferable (e.g., interview studies (Kvale 2007)). However, the aim of the ESM is to allow for *in situ* data collection of current events, as this provides a more reliable way of collecting data on everyday life without relying on the reconstruction of past events.

Gouveia and Karapanos (2013) present two major points of criticism regarding the ESM: "While the experience sampling and diary methods are considered the gold standard of *insitu* data collection [...], they also entail important drawbacks as they are disruptive to participants' daily activities and suffer from a lack of realism as the remote researcher does not have rich data about the situations on which participants report [...]." By including sensor-based data collection in a study, researchers are able to reconstruct a rich situational picture for each individual participant at each moment in time (Dey 2001). Evidently, the context as inferred from the mobile device's sensors is likely unable to provide a detailed or complete understanding as obtainable in a laboratory

observation. On the other hand, the ESM does allow for deployment in the participants' naturalistic and familiar environment. Critique regarding the level of participant disruption of *in situ* data collection (Gouveia and Karapanos 2013) is valid, i.e., ESM notifications may come at inopportune moments and interfere with the participants' current activity.

Researchers have developed a few alternatives to the ESM. The Ecological Momentary Assessment (EMA) was developed to perform *in situ* data collection in the discipline of behavioural medicine (Shiffman et al. 2008). Some scholars state that "EMA is a more broadly defined construct than ESM" (Stone and Shiffman 2002), arguing that the focus of ESM is traditionally limited to randomized sampling (i.e., the presentation of notifications at random times) as opposed to other sampling techniques, and is focused on private and subject phenomenon as opposed to behavioural and physiological measures. However, the terms ESM and EMA are nowadays often used interchangeably (Scollon et al. 2003; Yue et al. 2014), as the methods for data collection become analogous. For this reason, when we use the term ESM in this survey, we also refer to studies employing EMA.

Other related methodologies found in the literature include the Day Reconstruction Method (DRM), shadowing studies, and video-based observational studies. We shortly highlight these methods and include references for the interested reader.

Kahneman et al. (2004) developed the DRM to be used for well-being research. Like the ESM, it aims to characterise the experiences of daily life through self-reports. Instead of several short questionnaires throughout the day, participants reconstruct their experiences either at the end of the day or at the following day. This reduces participant burden, but consequently the DRM is unable to capture *in situ* data. Participants are usually asked to first reconstruct the events of the day into a sequence of (chronological) episodes, after which they answer a set of questions by drawing on the reconstructed sequences.

The shadowing methodology, first observed in management studies from the 1950s, describes an observation technique in which researchers follow, 'shadow', study participants, noting down observations which they consider relevant. A study on foremen working in an assembly plant is often seen as a key example of this method (Walker 1956). As the researchers are present during observation, they can record the events of interest (as opposed to relying on the participants' judgement). On the other hand, the presence of a researcher may skew the behaviour of participants and introduce ethical issues—as for example observed in the measurement of hospital hand hygiene compliance (Haas and Larson 2007).

Lastly, video-based observation consists of the (audio)visual recording of the environment of interest (Asan and Montague 2014). Researchers are not directly present during the observation period, thus reducing the intrusiveness of the study. In addition, researchers can observe the scenario multiple times and from multiple angles. The downsides of this method include the limited observation area covered by static cameras, change in behaviour of participants as the result of being observed, and the amount of manual work involved with reviewing and annotating video data. As observation is limited to the instrumented area(s), video-based observation is most suited for studies assessing a phenomenon restricted to a single location. An alternative is found in wearable cameras, which can often be clipped on to a participant's clothing and are small enough to be relatively unobtrusive. While this allows for *in situ* observation, it remains tedious to analyse the recordings and introduces new ethical considerations for both the study participants and their surroundings.

3 LITERATURE REVIEW

We conduct an extensive and systematic literature review on the use of the ESM. For this, we utilise three established digital libraries focusing on the computer science domain (*IEEE Xplore*

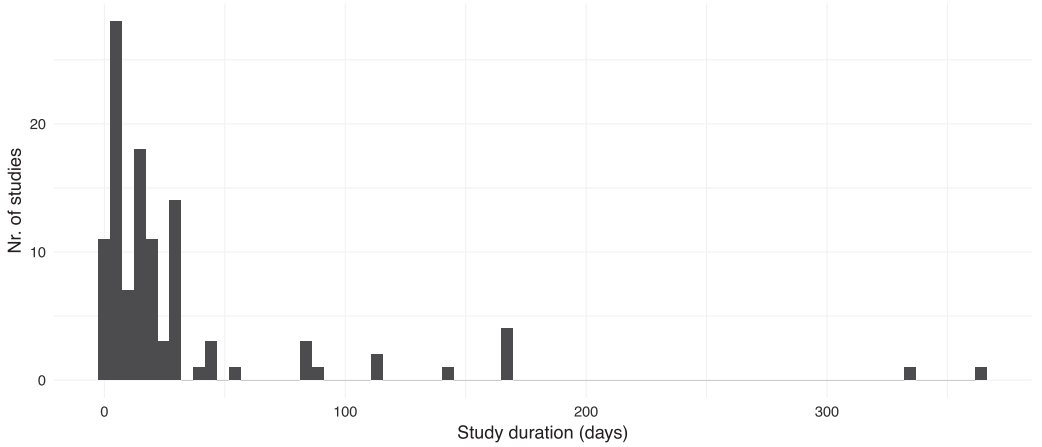

Fig. 2. Study duration in analysed studies.

Digital Library, ACM Guide to Computing Literature, and *USENIX*). Together, these libraries provide a distinctive overview of peer-reviewed publications in our community. We applied an 'inclusive or' search query using the terms *Experience Sampling Method, Ecological Momentary Assessment,* and *Experience Sampling*. Abbreviations such as ESM and EMA were not included as a search term as they result in a high number of false positives by referring to different scientific concepts (e.g., 'Effective Mass Approximation'). Our search strategy includes both full-text and metadata (e.g., title, abstract) in the search results. Furthermore, we apply a search filter to only consider publications since 2005 to ensure a focus on modern mobile devices.

After merging these search results and removing duplicate papers, a total number of 461 papers remained. These 461 papers were each individually analysed by one of the authors of this literature review. In case of ambiguity or confusion, we consulted among ourselves to find consensus in our classifications. Many papers were included in the search results, but excluded from further analysis. Typically, these papers either asked participants to answer questions on a device while being supervised by a researcher (either in a lab setting or in the field), deployed one-off surveys on mobile devices or online, described a system or software without evaluating it, were a literature review, or were a brief work in progress. Following this curation, a total of 110 relevant papers remained.

3.1 Study Parameters

Based on the analysed literature, we identify common study parameters in ESM studies. We would like to acknowledge once more that our analysis may not be all inclusive and is based solely on the information available in the publications' content.

3.1.1 Study Duration. The average reported study duration is 32 days, with a high standard deviation of 57.1 due to several high outliers (see Figure 2). As such, the median study duration of 14 days gives a more accurate representation of study duration in the analysed studies. A reported study duration of 1 month is assumed to consist of (4 * 7 =) 28 days. For studies reporting a different study duration between participants, we assigned the mean duration value. Most studies (70.9%) report a study duration of less than 1 month. We believe this is a natural consequence of the ESM methodology and self-reports in general, as a safeguard to reduce participant burden.

In their review of Ecological Momentary Interventions, Heron and Smyth (2010) found a similar large difference in study duration and urge researchers to find a "balance between study duration and intervention frequency." The currently common practice, in which studies last for around 2 weeks (Table 1) and request multiple short questionnaires per day, usually results in a good response rate. This duration is in line with previous recommendations from diary studies (Stone et al. 1991) and ESM studies employing PDAs (Reis et al. 2000).

3.1.2 Number of Participants. The number of participants differ drastically in the analysed papers, ranging from 1 to 1013 participants (SD = 124.6). Due to several outliers, the mean number of participants is relatively high with 53 participants. Our results indicate a median number of 19 participants, providing a more representative insight into ESM practices. This number of participants is in line with the local standards in the HCI community (Caine 2016) (sample size has a mode of 12 and a median of 18).

Researchers use various methods to determine a study's sample size, including prospective power analysis, data saturation, cost analysis, or by using guidelines (e.g., local standards) (Caine 2016). While a power analysis provides an objective approach to determining sample size for a quantitative analysis, it also comes with various limitations. As a power analysis is based on previous data on the topic at hand (e.g., effect size), it is less useful to study novel technology, as is often the case in computer science. A moderate sample size makes it feasible for researchers to carry out exit interviews or other qualitative methods. For these reasons, we believe the current practice of sample size in experience sampling, which is in line with local standards of the HCI community, is appropriate. However, we do recommend researchers to use Bayesian statistics, as this allows for statistical inference on small sample sizes (Kay et al. 2016).

3.1.3 Response Rate. Participant response rate, often referred to as *compliance rate* in medical literature, describes the ratio of answered ESM notifications across the study population. We define response rate as: number of fully completed questionnaires divided by the number of questionnaires presented. If participant responses were removed from the results (e.g., nonsense answers were provided), we subtracted these responses from the number of *completed* questionnaires only. A high response rate indicates a more complete picture of the studied phenomena. Furthermore, it is a sign of data being collected over an array of timeslots and more likely to be contextually diverse. The opposite also holds: a low response rate indicates a low level of *in situ* measurements. A lower response rate (missing measurements) may be the result of multiple factors, including most prominently; the study's ESM configuration options (e.g., notification expiry time), low participant motivation, or instrumentation flaws.

A total of 65 papers (59.1%) did not report response rates. In an additional three papers, the response rate was considered 'not applicable', as participants did not receive any notifications but were free to submit data at any time. Average response rate for the remaining papers is 69.6% (SD = 22.8%).

Response rate is affected by many factors, including study subject, methodological configuration, and questionnaire length. *Equity theory* (Adams 1963) suggests that participants are more willing to provide input if the costs to participation (e.g., time, energy, resources) are lower than the value of the expected outcome. This might explain some of the differences observed in Table 1. The study with the lowest response rate, (Fischer and Benford 2009), asked participants to report their engagement while playing in a long-term game. Participants were therefore occasionally interrupted during game-play, without any direct personal benefits following from the study. Studies with higher response rates typically provide a concrete goal that is of direct influence to the participant (e.g., providing context information prior to answering a phone call (Grandhi and Jones 2015)). Larson and Csikszentmihalyi (1983) originally advised researchers to establish a "*viable*

Table 1. Overview from Systematic Literature Review

Reference	Duration (in days)	N	Trigger: Signal	Trigger: Interval	Trigger: Event	Trigger: Free input	Response rate	Compensation: Fixed	Compensation: Raffle	Compensation: # responses	Compensation: None	Personal device	Device sensor: Location	Device sensor: Phone calls	Device sensor: Network related	Device sensor: Bio-signal(s)	Device sensor: Phone usage	Device sensor: Accelerometer	Device sensor: App. usage	Device sensor: Other(s)
(Consolvo et al. 2005)	14	16	•				90.4%	•		•		no								
(Abowd et al. 2005)	21	12				•						no								
(Khalil and Connelly 2006)	10	20	•				80.0%					no								
(Tsai et al. 2006)	28	15		•			80–100%					yes								
(Froehlich et al. 2006)	28	16			•		80.5%			•		no	•							
(Sala et al. 2007)	3	19	•				35.0%	•				no								
(Hareva et al. 2007)	14	5		•								yes				•				
(Mihalic and Tscheligi 2007)	7	8		•				•				no								
(Markopoulos et al. 2015)	7	36		•								no								
(Anthony et al. 2007)	7	25	•				92.0%	•				no								
(ter Hofte 2007)	7	10	•				74.0%	•				yes			•					
(Muukkonen et al. 2008)	14	55		•								no								
(Khan et al. 2008)	7	11		•	•							no								
(Klasnja et al. 2008)	8	20	•	•				•				no				•				
(Khan et al. 2009)	7	20		•	•			•				no	•		•					
(Westerink et al. 2009)	1	32		•	•							no				•				
(Mancini et al. 2009)	21	6			•						•	yes								•
(Fischer and Benford 2009)	13	16		•	•		17.9%					yes					•			
(Ara et al. 2009)	14–28	60	•									no						•		•
(Motahari et al. 2009a)	28	129			•			•				no	•							
(Pärkkä et al. 2009)	90	17				•	Not applicable					no				•				
(Motahari et al. 2009b)	21	165		•	•		20.0%		•			no	•							
(Lee et al. 2010)	84	10			•							no								•
(Tejani et al. 2010)	336	304	•									no								
(Fischer et al. 2010)	10	11		•			65.0%					yes								
(Cowan et al. 2010)	14	16	•		•		31.3%					no								
(Ketykó et al. 2010)	1	18			•							no			•					
(Grandhi and Jones 2010)	6	38			•				•			no		•						
(Ellingson and Oken 2011)	1	1		•								no				•				
(Maxhuni et al. 2011)	5	3		•		•						no								•
(Kim et al. 2011)	21	11			•							no		•	•					•
(Fischer et al. 2011)	14	20	•		•		68.9%	•				no		•						•
(Pessemier et al. 2011)	7	29			•		100%	•				no								
(Moturu et al. 2011a)	168	54		•								no		•	•					•
(Fisher and Simmons 2011)	7	5		•								yes	•	•				•		•
(Rosenthal et al. 2011)	28	19	•		•			•				yes								•
(Moturu et al. 2011b)	168	54		•								no		•	•					•
(Fletcher et al. 2011)	14	25	•		•							no			•					
(Rieger et al. 2012)	1	128	•		•		87.7%					no				•				
(Wang et al. 2012)	28	28		•				•				no								

(Continued)

Table 1. Continued

Reference	Duration (in days)	N	Trigger				Response rate	Compensation				Personal device	Device sensor						
			Signal	Interval	Event	Free input		Fixed	Raffle	# responses	None		Location	Phone calls	Network related	Bio-signal(s)	Phone usage	Accelerometer	App. usage	Other(s)
(Korunka et al. 2012)	28	6		•			75.0%													
(Ickin et al. 2012)	28	30	•	•								yes	•	•	•					•
(Moreno et al. 2012)	7	189	•				93.2%	•		•		yes								
(Wilson et al. 2012)	2	15					92.2%	•				no								
(López et al. 2012)	5	12	•				46.6%					no								
(Lepri et al. 2012)	42	51		•			83.9%					yes								•
(Gomes et al. 2012)	168	12		•	•		53.5%					yes								
(Gaggioli et al. 2013)	7	6	•				90.0%				•	no				•		•		
(Costa et al. 2013)	14	7				•					•	yes	•							•
(Hasan et al. 2013)	21	5		•		•							•							•
(Liang et al. 2013)	14–28	11		•								yes			•					
(Lathia et al. 2013)	18	22	•		•		83.5%				•	yes	•	•				•		•
(Teso et al. 2013)	144	54		•								no								•
(Burgin et al. 2013)	7	749	•				65.2%		•	•		no								
(Ickin et al. 2013)	2–5	5				•	Not applicable					yes							•	•
(Weppner et al. 2013)	84	9	•	•								no								
(Nguyen et al. 2014)	15	13																		•
(Pergler et al. 2014)	1	12			•							yes							•	
(Church et al. 2014)	84	108		•			43.0%	•	•			yes								
(Ferreira et al. 2014)	14	15			•		76.0%					yes					•			
(Sabatelli et al. 2014)	3	7										no			•					
(Hasan et al. 2014)	42	19	•																	
(Faiola and Srinivas 2014)	3	24		•								yes								
(Ayzenberg and Picard 2014)	10	10				•	Not applicable					yes	•	•						•
(Harbach et al. 2014)	27	52	•		•			•		•		yes					•			
(Smith et al. 2014)	112	11			•							yes		•						
(Van den Broucke et al. 2014)	14	13			•		64.5%					yes	•		•					•
(Yue et al. 2014)	5	1013	•							•		yes								•
(Liu et al. 2014b)	7	20			•			•				yes	•							
(Patil et al. 2014)	15	35	•					•		•		yes	•							
(Gonzales 2014)	6	76	•					•				no								
(Linnap and Rice 2014)	1	24		•	•								•							
(Seto et al. 2014)	14	12				•							•					•		
(Adams et al. 2014)	10	7		•			40.0%				•	mixed								•
(Alcañiz et al. 2014)	7	21										no								
(Tollmar and Huang 2015)	7	40		•			65.3%					yes								
(Miu et al. 2015)	1	10			•													•		
(Kim et al. 2015b)	37	57	•	•								no								
(Niforatos and Karapanos 2014)	7	13			•		50.0%					yes		•			•		•	
(Sabra et al. 2015)	1	10			•							no	•							

(Continued)

Table 1. Continued

Reference	Duration (in days)	N	Trigger: Signal	Trigger: Interval	Trigger: Event	Trigger: Free input	Response rate	Compensation: Fixed	Compensation: Raffle	Compensation: # responses	Compensation: None	Personal device	Device sensor: Location	Device sensor: Phone calls	Device sensor: Network related	Device sensor: Bio-signal(s)	Device sensor: Phone usage	Device sensor: Accelerometer	Device sensor: App. usage	Device sensor: Other(s)
(Ghosh et al. 2015)	16	2		•	•							yes								•
(Hasan et al. 2015)	42	34	•			•														
(Patil et al. 2015)	15	35	•					•		•		yes	•							
(Grandhi and Jones 2015)	7–14	30			•		93.3%					no		•						
(Reyal et al. 2015)	28	12		•				•				yes								•
(Shih et al. 2015)	28	34		•				•		•		yes	•						•	
(Kim et al. 2015a)	2	22	•	•			92.0%					no				•				
(Ferreira et al. 2015a)	168	218			•						•	yes							•	
(Guha and Wicker 2015)	365	204			•						•	yes								•
(Saeb et al. 2015)	14	18		•				•				mixed	•							
(Xu et al. 2015)	7	97		•								yes					•			
(Intille et al. 2016)	28	33	•				40-90%				•	mixed								
(Gustarini et al. 2016)	27	42	•		•		52.1%					yes								•
(Zhang et al. 2016)	18	24			•			•				yes					•			
(Johansen and Kanstrup 2016)	30	5	•									yes								
(Yang et al. 2016)	14	91		•			91.7%	•	•			yes								
(Maekawa et al. 2016)	15–25	15	•				78.8%					no	•		•			•		•
(Vhaduri and Poellabauer 2016)	28	17		•		•	29.3%				•									
(Weber et al. 2016)	7	16	•	•						•		no					•			•
(Ciman and Wac 2016)	28	25		•			58.2%		•			yes				•				•
(Smyth and Heron 2016)	11	90					89.0%	•				no								
(Mehrotra et al. 2016)	56	20	•		•							yes								•
(Spanakis et al. 2016)	14	100	•		•															
(Nguyen et al. 2016)	21	13				•														•
(Ashour et al. 2016)	112	10	•				97.8%					yes				•				
(Lee et al. 2016)	25	16		•			88.0%					yes								
(Walsh and Brinker 2016)	2	179	•					•				mixed								
(Buschek et al. 2016)	30	18			•		49.1%	•				yes					•			
(Mayer et al. 2016)	4	85	•					•		•		yes								
(Hernandez et al. 2016)	5	15	•				75–90%	•				no								

research alliance," in which participants understand the importance of their contribution to the study to ensure high response rates. While we acknowledge the importance of this notion, this alone might not be sufficient. Results from our literature review indicate that a direct connection between the participants' effort and an intrinsic reward are beneficial. We therefore advise researchers to make the collection of self-reports intrinsically rewarding for the participant when possible (see Section 4.5).

The method used to designate responses as valid or invalid also influences the study's overall response rate. For example, Shih et al. (2015) present a questionnaire every hour, yet only dismiss this questionnaire after 3 hours. This essentially allows participants to answer multiple

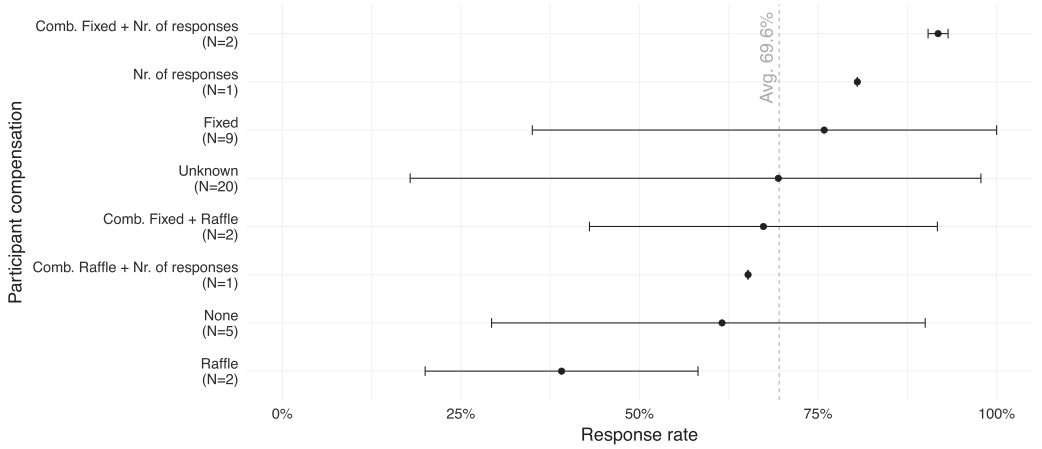

Fig. 3. Response rates for different compensation techniques (min, mean, max).

questionnaires in one session. In a different study, notifications were automatically dismissed after three minutes (Ferreira et al. 2014). These decisions highly effect the level of *in situ* measurement and the resulting response rate. Because of the negative effect on the level of *in situ* measurement, we generally advise against the stacking of questionnaires. These examples highlight the importance of reporting methodological study parameters (see also Section 5).

3.1.4 Compensation. A variety of participant compensation structures are present in the analysed literature. Various studies also combine different structures in their reward schema. Following a classification of the literature, we define four compensation structures:

- —Fixed reward paid to participants in which a certain good is offered as an incentive. Common examples are monetary rewards, vouchers, or study credits.
- —A raffle, in which participants have a certain chance to win one of the advertised goods. This compensation structure is often combined with a fixed reward.
- —Compensation based on the (number of) participant responses. Either by offering participants a fixed reward after a certain response rate is achieved, or by offering a small reward for each answer provided.
- —No reward provided to the participant.

Unfortunately, a total of 64 studies (58.2%) do not mention whether or how participants are compensated. This limits our ability to compare study results. These and other such issues regarding reproducibility and comparability are further discussed in Section 5. The most popular form of reward is a fixed reward (21 studies), in which participants receive an agreed-upon reward. Three studies compensated their participants solely on a *per answer* basis, and another three studies compensated through a raffle—resulting in a reward for only one or several participants. A total of 9 studies specifically note that participants received no reward. In addition, we note that various studies combine two compensation structures. This includes the combination of a fixed reward and compensation based on the number of responses (7 studies), a fixed reward and a raffle (2 studies), and a raffle in combination with the number of responses (1 study).

Figure 3 provides an overview of the effects of compensation on participant response rate in our sample. Please note that the majority of studies in our sample did not report response rates, making

it challenging to identify the relationship between compensation and response rate. Additionally, participants do not solely consider economic motives when participating in a study (Lynn 2001) (see Section 4.5). Although our sample is limited, the data suggests that the use of micro-incentive compensation leads to higher response rates and thus warrants further investigation. Raffle-based compensation has a relatively low response rate in our sample.

3.1.5 ESM Trigger. Notifying participants is a fundamental element of the ESM (Consolvo and Walker 2003; Lathia et al. 2013). Participant burden is reduced by reminding participants to provide data, as opposed to participants providing data at their own accord (Chang et al. 2015). Three different types of notification triggers are described in the literature (Barrett and Barrett 2001; Wheeler and Reis 1991):

— *Signal contingent* entails the presentation of alerts randomised over the course of a given timespan (usually a single day). This timespan typically contains a certain schedule to avoid night-time alerts. Additionally, the number of alerts can be restricted to a set daily maximum.
— *Interval contingent* entails the presentation of alerts according to a predefined interval or schedule—for example, to present a questionnaire every hour. Given the standardised time gaps between alerts, interval contingent is well suited for time series analysis (Conner Christensen et al. 2003). Again, it is common to define a schedule to avoid night-time alerts.
— *Event contingent* entails the presentation of alerts according to the start or completion of a predefined event. This event can result from changes in hardware sensor readings (e.g., GPS (Sabra et al. 2015)), detected events on the device (e.g., an incoming phone call (Grandhi and Jones 2010)), or even an event external to the device (*e.g.*, food intake (Seto et al. 2014)).

Each of these different notification modes can introduce potential biases; "[...] 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" (Lathia et al. 2013). In addition to the three notification triggers, we observe several studies encouraging participant-initiated data submissions (e.g., Costa et al. (2013) and Seto et al. (2014)). Allowing participants to submit data is especially useful for events that occur seldom or irregularly, and are in addition very difficult or impossible to measure reliably with mobile sensor information. However, if this is not the case, it is advisable to steer away from active participant input in an ESM study. Its methodology is considerably different, and the data collected from ESM responses and active participant input cannot simply be analysed as one collection.

In Table 1, we observe that interval contingent triggers are most common (26 studies), followed by signal contingent triggers (23 studies), and lastly event contingent triggers (21 studies). From our sample, 7 studies relied solely on data collection initiated by participants rather than notifications. Surprisingly, 5 studies do not report the trigger used in their study. The remaining studies (a total of 28) apply a range of combinations in their contingency configurations. Of these, the most common combination is a random and event contingent trigger (10 studies). This combination ensures both measurements throughout the day and at occurrence of a specific event.

3.1.6 Sensor Usage. Of the identified studies, a total of 70 studies (63.6%) passively or actively collect sensor data from the participants' study device. Our analysis includes both the use of hardware and software sensors. The location sensor proves to be the most popular sensor, being used in 20 studies, followed by the logging of phone call activity in 12 studies. Other common information types include network-related events (including Bluetooth, 12 studies), and physiological measurements (including activity, 10 studies). Sensor data is used to either construct the participants' context, or to trigger notifications based on the occurrence of a specified event. For the

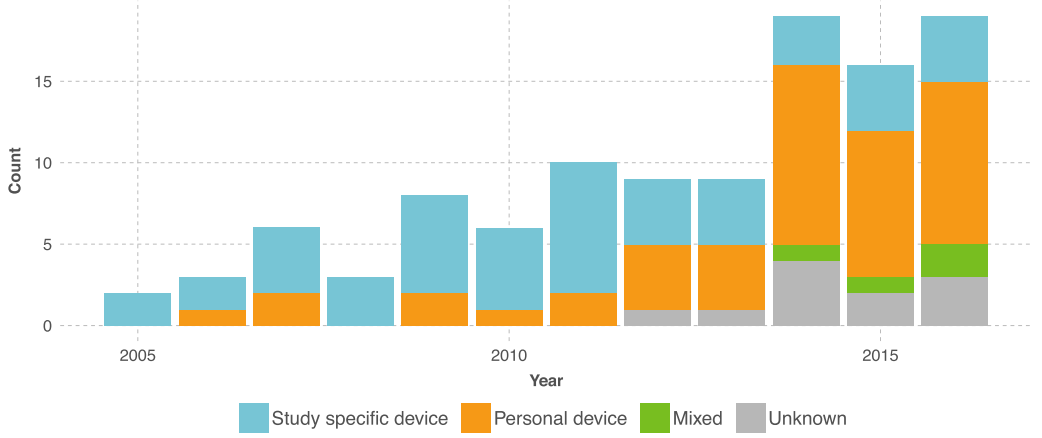

Fig. 4. Device ownership in analysed studies.

most commonly used sensor, this for example includes inferring interruptibility based on location (Fisher and Simmons 2011), sending a questionnaire at a specified location (Sabra et al. 2015), or distributing participant efforts over a geographic area (Linnap and Rice 2014).

The aforementioned examples highlight the possibilities of experience sampling on mobile devices. In their article on smartphones as a tool for social scientists, Raento et al. (2009) discuss the possibilities of a combined human and sensor data collection approach: "[...] since experience sampling can be easily applied in smartphones to complement background logging, smartphones themselves can provide a partial solution for the need for triangulation. Indeed, smartphones provide three modes of data collection: (1) automatic data logging in the background, (2) experience sampling as a way to collect subjective data, and (3) integration of the two." (Raento et al. 2009). While thus deemed useful for the social sciences, collection of sensor data remains rare, citing high development costs and lack of specialised skills (Raento et al. 2009). Our review shows that the use of sensor readings is relatively common in computer science, with 63.6% studies reporting the use of sensor data.

3.1.7 Device Ownership. In a total of 49 studies (44.5%) summarised in Table 1, researchers provided their participants with a mobile device for the duration of the study. In 46 studies (41.8%), participants used their personal device. In 4 studies (3.6%), participants used either their own device or a device provided by the researcher. For a total of 11 papers (10.0%), it was not indicated whether participants used their personal device. Figure 4 provides an overview of the changes in device ownership over time.

As seen in Figure 4, the use of personal devices is increasing. This is a positive development, as the use of a study-specific device might have unintentional side-effects. Carrying around a separate research device will affect the participants' day-to-day activities. "In principle, the less aware the subject is of the presence of the observing device, the less its presence should affect the study." (Raento et al. 2009). In addition, participants will not be completely comfortable in operating an unknown device, introducing a novelty effect. Unfortunately, we also note an increase in studies in which device ownership is not specified. While this may be indicative of a trend in which researchers assume personal devices used as study device, this cannot be presumed as a considerable number of studies still employ a study-specific device.

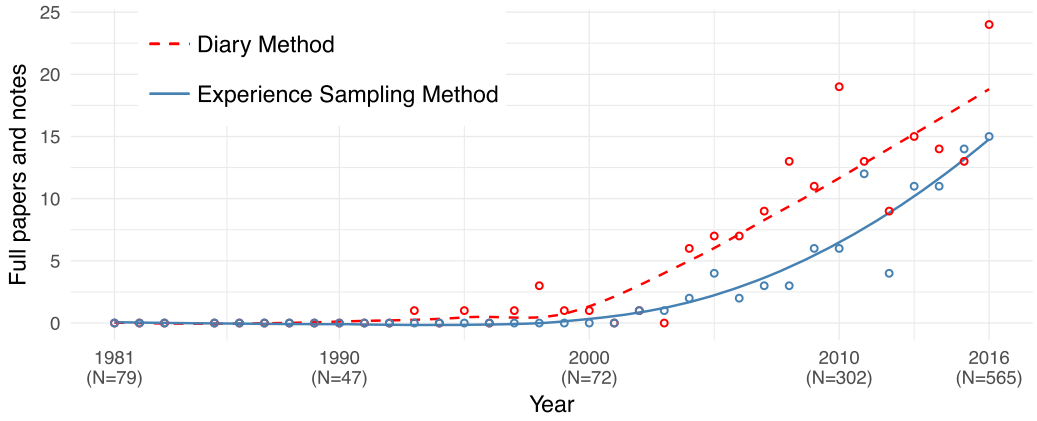

Fig. 5. Mentions of the Diary Method and the ESM over time in CHI proceedings.

3.2 Popularity of ESM Over Time

We analyse the popularity of experience sampling over time in the Conference on Human Factors in Computing Systems (CHI). CHI is generally considered as the flagship conference in HCI, and therefore is often used to provide an indication of changes in the HCI landscape (e.g., Caine (2016) and Liu et al. (2014a)). We use the ACM Digital Library to search full-text within each CHI proceeding for usage of one (or more) of the following terms *Experience Sampling Method, Ecological Momentary Assessment*, and *Experience Sampling*. We restrict our search to full papers and notes. In addition, we search for the methodologically related diary method (*Diary Method*, and *Diary Study*) to allow for a compare and contrast.

The results of this analysis are shown in Figure 5. The first mentions of diary studies are visible in the 1990s, with a continuous uptake in usage following the new millennium. The ESM follows a similar pattern, with first studies carried out in the 2000s. An increase in usage can be seen from 2010 till date, with the trend line indicating increased usage in the coming years. While it is evident that the number of publications at the CHI conference has increased considerably (see x-axis of Figure 5), both methodologies have established themselves in the HCI community and show continued uptake. Although the first publications of the ESM in Psychology already appeared in the late 1970s (e.g., Csikszentmihalyi et al. (1977)), the first publication in CHI which made use of this method appeared in 2002.

To verify whether this aforementioned trend holds true for the general scientific community, we carry out a similar search using Google Scholar. Google Scholar provides a general overview on the output of the scientific community, including a wide range of sources and disciplines. We use the same search terms as for our analysis of the CHI proceedings, and exclude patents and citations. We collect the number of search results per year, ranging from 1981 (as per Figure 5) up to the year 2016.

The results of this analysis are shown in Figure 6. Results leading up to the year 2000 show a similar pattern as seen in the analysis of the CHI proceedings, with mentions of the diary method outranking those of the ESM. The number of search results of the ESM overtake the number of results for the Diary Method in 2009. The trend lines (Loess Curves) indicate that the number of mentions for both methods is expected to increase over the coming years. We must note that the number of *total* yearly publications within the scientific community have increased considerably over the past decades. In addition, search results returned by Google Scholar come from a wide

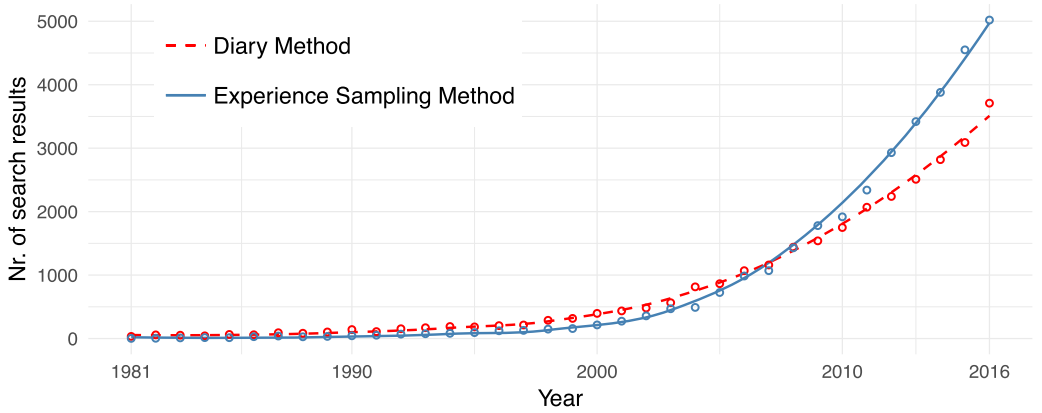

Fig. 6. Number of search results of Diary Method and ESM in Google Scholar.

range of sources, not all of which are peer-reviewed academic venues. However, this analysis clearly shows the increased popularity of the ESM over time.

4 DATA COLLECTION IN ESM STUDIES

Central to the ESM is the collection of *in situ* data. With increasing capabilities of mobile devices, researchers can collect a wide array of different data sources. These new possibilities raise the question as to exactly *which* data researchers should collect. While determining relevant datatypes is evidently dependent on the research goals, it is prudent to consider all data collection possibilities. Table 2 provides an overview of commonly used question types. We briefly explain each input type, when it is used, and which design challenges one might encounter in a mobile study.

The collection of multiple data types can be used to extract additional value out of participant responses. For example, Yue et al. (2014) investigated the role of photos in ESM studies. Photos not only provide a way to elicit recall in a post-study interview, but can serve as added value during data analysis. Researchers may utilise collected photos to clarify participant answers (e.g., "bought a new keyboard" could refer to both a musical instrument and a computer input device).

It is possible to acquire participants' context through the sensors of the participants' mobile device. Context is defined as "any information that can be used to characterise the situation of an entity" (Dey 2001). A key benefit is that acquired context can enrich the collected data without explicit input by the participant. For example, upon completion of a questionnaire, the current GPS location can be stored and linked. This would allow a researcher to determine the meteorological conditions at the time and place of each individual participant response.

4.1 Methodological Decisions

Next to the study parameters already compared and contrasted in Section 3, the literature review highlights multiple methodological decisions that are pertinent to the design of an ESM study. We discuss our findings regarding these critical methodological decisions faced by researcher running an ESM study, categorised into five key categories: notification schedule, notification expiry, inter-notification time, inquiry limit, and study duration.

4.1.1 Notification Schedule. Traditionally, it has been considered good practice to design questionnaires that minimise participants' burden (Consolvo and Walker 2003). Yet, in ESM studies the goal of the researcher is to collect rich data on a participants' experiences—resulting in two

Table 2. Common ESM Question Types

Input type	Description	Question type usage	Mobile design challenge
Text field	Text entry using keyboard. Can be restricted to only accept text conforming to a set format.	Input unable to be captured in a small set of choices; e.g., the participant's current thoughts.	Form factor constraints make it tedious for participants to input large amount of text.
Radio button	Select only *one* option out of the presented possibilities. Can take a variety of visual appearances.	Make a distinctive choice between a set of predefined options.	Often impossible to include all possibilities, consider the addition of an 'other' option.
Checkbox	Select *multiple* options out of a presented list. Can take a variety of visual appearances.	Select 'all that apply' from a presented list. Both no and only one selection are valid input.	Often impossible to include all possibilities, consider the addition of an 'other' option.
Likert scale	Select one option on a scale. Special type of radio button input type.	Indicate level or degree of (dis)agreement with an accompanying statement.	Limited possibility to display a (large) scale. A 'default' selected state allows participants to quickly skip the question(s).
Slider	Select any value between a set minimum and maximum.	Indicate level or degree of (dis)agreement with an accompanying statement.	Limited level of precision using a slider input.
Photo/video	Take a picture or shoot a video clip. Often also allows for the selection of a previously captured media file.	Provide media to augment the current context of the participant.	Storage and transfer of files can be problematic on a mobile data plan. Privacy concerns for participant and surrounding.
Audio	Access to the phone's microphone(s) to start an audio recording.	Provide media to augment the current context or allow participant to record own thoughts and ideas.	Privacy concerns for both the participant and the participant's surrounding. Time consuming to analyse and transcribe.
Affect grid	Describe mood as compared to the total possibility space. Grid of 81 radio buttons.	Indicate mood at a given time, especially valuable when tracked over time.	Limited screen place to explain the workings of the affect grid to participants.

clearly opposing requirements. While questionnaires cannot be completed without interrupting participants, the interruption can be kept as brief as possible by reducing the number of questions (e.g., by considering context), providing a list of pre-populated options through radio buttons, and increasing text legibility.

Studies should adapt their notification schedule to minimise participant burden. For instance, when utilising a *signal-* or *interval*-contingent alert type, it may be desirable to avoid waking the participants during the night. Several studies therefore enforce a notification schedule during which no notifications will be sent to the participants at this time of day. For example, Pejovic and Musolesi (2014) apply a 08:00–22:00 schedule, while others allow participants to determine the notification schedule (e.g., Church et al. (2014)).

In addition, researchers can utilise the context of participants to determine a timing that is both suitable for participants and in line with the research question(s). Park (2005) investigated mobile phone addiction and found that 70% of participants use their phone directly after waking up. Hence, by inferring sleep and wake-up time, researchers can adjust notification schedule parameters for each participant. In fact, the use of context information over strict timeslots in determining user availability can prove to be much more efficient. For example, Randall and Rickard (2013) present a questionnaire at the event of participants' listening to music on their device and report 27% of their data being recorded outside of a regular 'collection window' from 08:00–22:00. We therefore recommend researchers make use of these sensor readings, as they can be successful indicators to users' attentiveness.

Lastly, researchers may consider scheduling notifications according to literature findings on mobile phone usage. According to data from Dingler and Pielot (2015), mobile phone users are attentive to messages for over 12 hours per day, with higher levels of attentiveness during evenings and weekdays. Researchers can utilise this knowledge in multiple ways. For example, longer questionnaires can be sent during the evening hours, while short questionnaires are presented during daytime. However, researchers must also be wary of the consequences of this behaviour. If questionnaires are answered primarily in the evening hours, this skews data collection and reduces the validity of the study. Therefore, it is important to visually inspect the distribution of answers over time, both over the total duration of the study, as well as the spread of answers over a cumulated day. To combat this effect, researchers can combine an event-contingent study configuration, in which a questionnaire is triggered by a specified event, with a random or pre-determined time schedule. Using this approach, questionnaires are triggered following an event, but only once per timeslot. This prevents an imbalance in questionnaire answer times.

4.1.2 Notification Expiry. Participants' response time to ESM notifications is an influencing factor on the measured level of *in situ*. Previous work has argued that ESM notifications should be withdrawn if they are ignored for a given amount of time (Abdesslem et al. 2010; Froehlich et al. 2007), ensuring that the question was answered in the context in which it was presented. Researchers can configure a *notification expiry* to specify the time after which unanswered questionnaires are withdrawn. This concept has also been called *notification time-out* (Consolvo and Walker 2003) and *notification lifetime* in the literature.

Our literature review indicates different notification expiry times (e.g., 3 minutes (Ferreira et al. 2014), 5 minutes (Khan et al. 2008)). Sahami Shirazi et al. (2014) found that if a user is to interact with a notification, there is a 50% chance this happens within the first 30 seconds after notification onset. Furthermore, there is a 83% probability that users will interact with a notification within the first 5 minutes after notification onset (Sahami Shirazi et al. 2014). Therefore, introducing notification expiry in a study can reduce participant burden while increasing data quality by maintaining a high level of *in situ*. Determining a suitable notification expiry time is dependent on

the content of the ESM questionnaire and the chosen ESM scheduling parameters. For example, an event-contingent study with a frequently occurring event is likely to benefit from a short expiry time. Ensuing notification expiry, the issued notification should be removed from the notification drawer to prevent further user-interaction. We strongly advise researchers to expire an ongoing notification when a new notification is send to the participant, preventing the 'stacking' of multiple notifications. Multiple of the same questionnaires answered at the same point in time provides no new information to the researcher, and negatively affects reliability of the data analysis.

4.1.3 Inter-Notification Time. Presentation of frequent notifications can overload the participant and reduce willingness to participate over a longer period of time. Intuitively, users are more willing to answer a question frequently when the question is short and does not require serious mental effort. The *inter-notification time* allows the researcher to configure the *minimum* time in-between two notifications. The literature shows considerable differences between ESM studies regarding the timeout between notifications. Consolvo and Walker (2003) restrict to one notification per 72-minute interval, with a maximum of 10 questionnaires per day. Ferreira et al. (2014) determined an inter-notification of 15 minutes for their study after analysing how long users are likely to leave their phone screen turned off. As a suitable inter-notification time is dependent on the measured parameters, experimentation is often advisable in order to determine a suitable timing (Iida et al. 2012).

The acceptability of a certain number of notifications is highly dependent on the time and effort required to complete each questionnaire (Consolvo and Walker 2003). In case the goal of the study is to observe long-term behaviour while also requiring fine-grained details on the observation, Collins and Graham (2002) suggest the combination of two separate data collection methods: "One attractive alternative is to use a longer measurement interval for the bulk of the data collection, but also to collect a subset of data using a temporal design with a shorter measurement interval." Using this combination of methods, long-term strain on participants is reduced, preventing higher drop-out rates and a near-certain drop in response quality.

4.1.4 Inquiry Limit. The *inquiry limit* determines the maximum number of questionnaires sent during the determined notification schedule. A sensible inquiry limit can be used to reduce the burden on study participants (Consolvo and Walker 2003). When participants are asked to answer open-ended questions (typically take longer to answer), the inquiry limit should be reduced, while short binary questions allow for a higher inquiry limit. One study indicates that "a sampling frequency of five to eight times per day may yield an optimal balance of recall and annoyance" (Klasnja et al. 2008). The questionnaire prompted to participants in this study contained eight questions, requiring participants to enter numbers on the occurrence and length of their walking and sitting behaviour up till the notification. Different research settings (both on the level of required recall and time required to complete questionnaires) will require a different inquiry limit. In order to get a feeling for the experienced participant burden, Hektner et al. (2007) advise researchers to run their own study as test-participant prior to inviting real participants. This is a good idea, as it will allow the researcher to not only experience the participant burden, but also to validate the technical implementation of the system.

4.1.5 Study Duration. The duration of a study is typically dependent on various factors, most importantly the frequency of occurrence of the studied phenomenon, number of questions per day, and questionnaire length. A study observing a rare event will require a longer duration to capture sufficient data points. Sending many questionnaires in a day will quickly reduce participant motivation, making it unfeasible to run for longer durations of time. Consequently, researchers aiming for more questionnaires per day should reduce either study duration or questionnaire length.

Despite the study duration being dependent on various factors, some common guidelines can be established.

To ensure a variety of days in the participants' life are captured, a minimum duration of 1 week is advisable: "Seven days are likely to yield a fairly representative sample of the various activities individuals engage in and to elicit multiple responses from many of these activities." (Hektner et al. 2007). For the maximum duration, no clear guidelines exist in the literature. Early diary studies report on the potential effect of self-reflection among participants in studies with a long duration and find the quality of collected data to deteriorate after a period of 2 to 4 weeks (Stone et al. 1991). Similarly, PDA-based ESM studies usually have a duration of 2 weeks (Reis and Gable 2000). We recommend similar durations for experience sampling on modern mobile devices, as this aspect has not yet been extensively studied. Therefore, a study investigating a frequently occurring phenomenon (i.e., multiple times a day) is typically well represented using a duration of 1 to 3 weeks. It is the responsibility of the researcher to analyse the collected data for any significant differences between the onset of the study and the last days of data collection. Any observed differences might be the result of reduced participant commitment, and/or an effect of study measurements on the participants' behaviour.

4.2 Privacy

With the use of mobile devices as a research instrument, an enormous amount of personal data can be collected. This is even more true when the participants' *personal device* is used for the data collection process. As described by Raento et al. (2009), participants may not realise the potential impact of the data collected—even if the researcher explicitly states what data is being collected. In fact, even the researcher may not be fully aware of what other revealing information exists within the dataset (Raento et al. 2009).

To grant participants more control over the data being collected, Lathia et al. (2013) allow sensing to be 'paused.' Participants have to actively enable this feature, which will stop the data collection for 30 minutes—and which can be repeated indefinitely. Raento et al. (2009) state that informed consent entails reminding participants that they are being observed, giving the example of playing a "*beep*" sound prior to their study participants engaging in a (recorded) phone call. The researchers state that this has in some cases led to altered conversations, missing important/relevant data, or simply biasing the participant. Therefore, there is a clear consideration to be made between reminding the participant about being observed and collecting more complete and truthful information.

4.3 Interruptibility

Given the personal nature of mobile devices, interruption of the user during day-to-day usage can quickly become an annoyance (Church and de Oliviera 2013), leading to reduced response rates. There is a growing literature on interruptibility on mobile devices, with notifications playing a key role in obtaining a user's attention (Pielot et al. 2014). These mobile notifications have been described as using visual, auditory, and/or tactile signals to inform the device owner (Sahami Shirazi et al. 2014).

Previous work has shown that smartphone notifications often interrupt the ongoing task of the user (Cutrell et al. 2001), and frequent notifications can start to annoy users (Church and de Oliviera 2013). These and similar findings have prompted the development of machine learning classifiers exploring opportune moments of notification presentation (Mehrotra et al. 2015). Knowing when and how to interrupt study participants is, therefore, of value to a researcher running an ESM study. For instance, Pielot et al. (2015) successfully show an increase in response rates when issuing

notifications during periods of "boredom." A ground truth on the level of experienced boredom was achieved using an ESM questionnaire.

Pejovic and Musolesi (2014) considered context in analysing interruptibility and accompanying sentiment. Using a classifier and manual participant labelling, the authors identify opportune moments of interruption, while also showing that sentiment decreases as the number of notifications increases during the day. Interruption load is therefore an important factor that researchers should consider when determining the timing of ESM questionnaires.

It is also important to highlight that incoming notifications by other applications running on the participants' phone are challenging to anticipate. Still, this information could be valuable to determine an opportune moment for interruption. Based on the aforementioned work (Pejovic and Musolesi 2014), it can be argued that when a participant has not received notifications for a while, willingness to reply increases. On the other hand, it is also reasonable to assume that when the participant is already responding to an incoming notification, it is less of a strain to respond to a study questionnaire. Surprisingly, no study has investigated the effect of participants' own application notifications on ESM response rate.

4.4 Presentation of ESM on Mobile Devices

Consolvo and Walker (2003) discuss several parameters that should be considered when designing a questionnaire, including question readability (e.g., font-size, contrast) and modality (e.g., text-based, audio-based). Some researchers propose that these parameters should be configurable by the participant to ensure accessibility for all (Froehlich et al. 2007). However, this results in a significant increase in development costs, which is simply not feasible for every study. We therefore advise researchers to focus their questionnaire design on a single modality (most likely text-based), and to make one design as accessible as possible.

Validated paper-based questionnaires can be considered as an effective basis for ESM questionnaires, for example the use of the Patient Health Questionnaire-9 (Jelenchick et al. 2013). However, questionnaires developed for traditional media (e.g., paper, computer monitor) cannot always be directly transformed for presentation on mobile devices. Long sentences, (complex) text input, and large selection lists may become unusable on small screens (Väätäjä and Roto 2010). Thus, the presentation becomes a methodological concern that researchers may have to review in light of the possibilities and restrictions of mobile devices.

As an example, Meschtscherjakov et al. (2009) explored the use of emoticons as a way of allowing for quick user input on a mobile device. Using a horizontal alignment of five emoticons, they propose a non-verbal question format—changing both the visual presentation and methodological construct. Their results indicate no bias caused by the emoticon-based presentation technique.

Hektner et al. (2007) propose various ESM question formats for capturing the *external coordinates* of experience; *date, time of day, physical location, activities*, and *companions*. Note that for several of these questions, the mobile device used by the participant can (partially) infer this information in the background. According to Hektner et al. (2007): "[...] these elements paint the backdrop against which one's daily experience is lived out." The thoughts and feelings of the participant are considered the *internal coordinates* of experience, and should be inferred through other questions. To measure these internal coordinates, researchers may consider questionnaires used in prior studies in order to increase inter-study comparability.

4.4.1 Logic and ESM Questions. The use of logical-based constructs (e.g., if *condition x* is met → ask *question y*) allows for the presentation of questions based on predetermined conditions. Froehlich et al. (2007) distinguish between *prescripts* and *postscripts*. Prescripts are executed before a question is presented, allowing the question to take into account the current context. Postscripts

allow researchers to influence the question flow based on a (set of) predetermined condition(s), such as current sensor readings or answers provided by the participant. This is often called *question branching*, *question skipping*, or *answer piping*. Using these logical constructs, data collection can focus on the exact moment(s) required by the researchers. As an example, analysing the user's context can allow the software to send a questionnaire every time the user travelled for at least a certain specified distance or following the conversation with a colleague.

4.5 Response Rate

Given the possibility for researchers to collect data in real time, participant input can be continuously analysed to assess retention rates as the study progresses. Researchers should contact participants when questionnaires have remained unanswered for multiple days. This allows researchers to answer any question the participant might have, resolve any technical issues, potentially convince the participant to increase participation efforts, or otherwise discard the participant and recruit a new participant. In addition, providing feedback on participant progress can increase retention rates (Hsieh et al. 2008; Stone et al. 1991). The consequences of participants with a low response rate are dependent on the studied phenomena and the context in which they are studied. For a study investigating a long-term condition (e.g., dietary patterns (Seto et al. 2014)), participant results with a low response rate would be of little value—whereas a study regarding relative short-context situations (e.g., speed dial (Lee et al. 2010)) could still benefit from a participant with only a small set of collected data points.

Various theories exist regarding the impact of incentives or compensation on participants' willingness to complete a questionnaire. Literature suggests that participants often combine a variation of *economic, altruistic, assistive*, and *reciprocative* (i.e., offering something in return) motives (Lynn 2001). The literature also contains practical suggestions to entice a high response rate in ESM, including; visualising participant data as a feedback mechanism (Hsieh et al. 2008), complex remuneration structures (Conner Christensen et al. 2003), the use of gamification mechanisms (van Berkel et al. 2017), and enticing participant engagement to feel part of the study (Larson and Csikszentmihalyi 1983). We encourage researchers to explore the possibilities to integrate such an incentive in their studies. Engaging with study participants (e.g., during study intake) to entice participants to feel part of the study is a recommended practice for all studies. This requires the researcher to emphasise the importance of the participants' contribution, describe the global study goal, and establish a joint rapport. In our own work, we establish joint rapport through individual intake sessions of 10 to 15 minutes. During these sessions, we take a friendly and professional approach in explaining the study goal and the participants' task(s). This involves taking the time to answer any questions from participants, and showing a genuine interest in any concerns the participant may have. We are upfront about any (personal) data collection, and provide details about implemented privacy measures. In addition, we make sure the participant understands the task(s), and that any software is successfully installed. Lastly, we thank participants for their participation and tell them to reach out to us (email, visit our office) if they have any problems or questions.

A common incentive for participants is a monetary compensation. Interestingly, a high compensation may not lead to desirable results. Stone et al. (1991) offered a compensation of $250 for participation but encountered poor data quality in their collected data. One potential reason for this may be the fact that participants were attracted for the wrong incentive (receiving money over the desire for altruistic participation). As shown in Figure 3, the use of micro-incentives (i.e., small (monetary) incentive per completed questionnaire) shows promise in achieving high response rate. While the sample is relatively small due to missing reports on response rate, we encourage researchers to utilise this incentive structure more frequently in the future. We advise against raffle-based compensation, as it leads to the lowest response rate in our sample.

4.6 Data Quality

Following the collection of participant answers, study data should be carefully prepared and analysed. We discuss three factors related to data quality which researchers should consider following data collection.

Data cleaning is used to filter out both nonsensical information provided by participants and erroneous data as the result of technical problems. Examples as identified in the literature review include the removal of participants with atypical smartphone usage behaviour (Buschek et al. 2016), and discarding incomplete or incorrectly compiled responses (Gaggioli et al. 2013, Harbach et al. 2014). In addition, the literature describes the removal of responses with suspiciously fast completion times (McCabe et al. 2011). Determining a suitable threshold for this depends on the questions presented and the answer options available (binary choice buttons allow for faster answers than text fields). McCabe et al. (2011) advise a (somewhat arbitrary) threshold of 0.5 seconds. Lastly, participants with a low response rate should be removed, as their data does not provide a complete picture on their daily experiences. Although the removal of participants with a low response rate is common practice, examples from the literature indicate the use of different thresholds (e.g., participants with no responses (Moreno et al. 2012), participants who completed less than half of the assigned tasks (Yang et al. 2016)). As the current literature does not provide well-supported cut-off rates for either questionnaire completion time or response rate, we suggest researchers to compare the results of individual participants to that of their peers to determine outliers in their data (e.g., completion times shorter than twice the standard deviation).

Lastly, the response shift phenomenon is an effect in which participants change the meaning they assign to response scales over the course of a study. As stated by Barta et al. (2012), "a study participant who rated her depressed mood as 5 on a 5-point scale might on a subsequent report rate the same intensity of depressed mood as 3 or 4 because she has come to realise that her depressed mood could be considerably more severe than the mood to which she had previously assigned a rating of 5". Schwartz et al. (2004) apply a retrospective pretest–posttest design (i.e., 'thentest'), measuring the participants' recalibration through a pre-study assessment and a post-study assessment of the pre-study experience—thus calculating the response shift effect. We advise researchers to apply the thentest or similar methods when collecting participant data on reflective measures (e.g., mood, pain, quality of life).

5 INCONSISTENCIES IN REPORTING ESM STUDIES

We highlight the inconsistencies observed in ESM study reporting, followed by practical advice on reporting important methodological decisions. The conducted literature review (Table 1) highlights multiple methodological decisions currently underreported. From a total of 110 studies, 65 studies do not report response rates, 64 studies do not provide information on participant compensation, 11 studies do not specify whether participants were using their personal device or a study-specific device, and 5 studies do not report the configuration of notification triggers. This inconsistency in, or complete lack of, reporting results is a critical shortcoming in both research methodology and practice. For example, a high response rate provides the reader with the reassurance that the presented results can be projected to the chosen population sample (Barclay et al. 2002).

5.1 Towards Consistent Reporting of Study Metrics

The consistent reporting of all major study metrics enables scientific comparison across ESM studies and even across the various scientific disciplines in which this method is applied. In addition, it will allow for more accurate replication of study results. We identify a set of core ESM metrics,

common to all studies, which we consider critical to be reported by researchers. These core metrics are:

— *Number of Participants.* Both prior to and following the data cleaning process.
— *Study Duration.* In case of differences between participants report the average study duration (including SD).
— *Notification Schedule.* Trigger mechanism, frequency, and any excluded hours.
— *Notification Expiry.* Report even if notifications do not expire.
— *Inter-Notification Time.* Report even if no inter-notification time is established.
— *Inquiry Limit.* Report even if no maximum number of notifications is established.
— *Device Ownership.* Participants' usage of personal or study-specific device.
— *Response Rate.* Response rate after the data has been cleaned.

In addition, any abnormalities in the reported data should be reported. This is especially the case for studies with a lower response rate. Information on the distribution of responses, both across participants and across time, allows for the identification of any potential skewing in terms of participant behaviour. We are not the first to draw attention to the importance of consistency in reporting study metrics in self-report methodologies. We refer the interested reader to Stone and Shiffman et al. (2002) and Church et al. (2015) on reporting guidelines for additional information.

6 ESM SOFTWARE

A wide variety of ESM-related tools exists, ranging from mobile texting services to sophisticated data collection frameworks. With a constantly evolving software landscape, it is worthwhile to analyse the existing choices. Services disappear, become outdated, receive a large overhaul, and new services emerge. We base our selection of tools for evaluation on the lists of resources provided by Conner (2015), Rough and Quigley (2015), and additional applications discovered through our peers and online search. Following the topic of our work, this analysis dismisses SMS texting services, generic survey tools without a notification mechanism, (purpose-made) PDA's (e.g., PsyMate (2016)), discontinued/unmaintained services, and non-English tools. Researchers interested in these specific categories are directed to the extensive overview by Conner (2015). By discontinued/unmaintained, we mean tools that have not been updated in the last 3 years, that is, these tools can no longer serve as a reliable instrument for deployment in a user study as they do not work on the latest mobile software (i.e., Android, iOS).

Table 3 shows wide differences in terms of functionality, pricing, and configuration options between the various software tools. Of 12 tools, 8 offer support for both the Android and iOS mobile platform, 3 are Android only, and 1 is iOS only. A total of 6 tools are offered for free, while the other tools are part of a commercial service (often in various tiers, sometimes providing a 'free tier' with limited capabilities). An open source license is provided for five of the analysed tools.

While most tools support signal and interval contingent notifications, support for event contingent notifications is only supported by six tools. Out of these six tools with support for event-contingent notifications, only two support active-sensing of all five analysed events (GPS, accelerometer, participants' usage of messaging, screen state (on/off), and phone details (e.g., incoming-call)). This lack of support for event-contingent notifications can explain the difference in number of studies applying this notification technique as compared to signal and interval contingent triggers (as seen in Table 1). In addition, the ability to *collect* and *store* sensor data was limited in most of the studied tools. While location tracking was supported by 11 out of 12 tools, other commonly used sensor types (e.g., accelerometer 5/12, and messaging 4/12) are missing.

Table 3. Overview of Current ESM Software and Functionality

		Trigger				OS				
Name	Sensor logging	Signal	Interval	Event	Branching	Android	iOS	Free	OSS*	Configuration
AWARE		•	•	•	•	•	•	•	•	Code
ESmCapture		•	•		•		•			Online & App
Ilumivu		•	•		•	•	•			Online
iPromptU			•			•	•	•		App
iSURVEY					•	•	•			Online
Jeeves **		•	•	•	•	•		•	•	App
LifeData		•	•			•	•			Online
MetricWire		•	•	•	•	•	•			Online
movisensXS		•	•		•	•				Online
ohmage		•	•	•		•	•	•	•	Online & XML
Paco		•	•	•	•	•	•	•	•	Online
Purple Robot			•	•		•		•	•	Code

*Open Source Software **indicates tool is not (yet) publically released.

Location Accelerometer Messaging Screen Phone details

6.1 Software Usability

Raento et al. (2009) discuss technical obstacles regarding the use of smartphones as a tool for carrying out research: "Thus far, smartphones have been mainly used in applied interdisciplinary areas like HCI and computer-supported cooperative work, mainly because of high development costs and the specialised skills needed for their utilization, but the technology clearly has potential beyond these applied settings." A surprisingly high number of platforms allow for configuration through a form-based web interface, omitting the need for software development skills. Unfortunately, the overview suggests that this also often results in lack of support for more advanced functionality (e.g., sensor logging, notification triggers). Thus, the possibilities for a researcher are limited without performing some form of programming. One key example is the absence of sensor-integration in the majority of software tools that we surveyed. The few software tools that do include integration with sensor readings are both technically complex or limited to the *passive* recording of sensor values without considerable software development. A similar sentiment can be found in the literature: "a lack of programming knowledge often hinders researchers in creating ESM applications" (Rough and Quigley 2015). In order to combat this obstacle, Rough and Quigley (2015) introduce *Jeeves*, a visual programming environment that aims to make use of the ESM more accessible to researchers without programming experience. Jeeves allows researchers to construct logic conditional statements using an accessible 'drag-and-drop' design.

In order for the ESM to increase its methodological value, there is a clear opportunity in the *active* reading of sensor values. Sensor values of the participants' mobile phones are helpful to determine in which context the questionnaire appears, as well as selecting those questions which are most relevant to the identified context. This increases the consistency of ESM results through contextual awareness, and can further reduce cognitive bias (Consolvo and Walker 2003). However, the interpretation of such raw sensor values is a complex task. For example, to obtain useful information from raw accelerometer values requires not only periodic sensing, but in addition needs feature engineering, classifier training, classifier evaluation, and finally inference to recognise physical activities (e.g., running) from raw sensor values. These are highly specialised skills, and are out of scope for most ESM practitioners. To make smartphone sensors useful for researchers thus requires software tools to offer high-level inferences of sensor values, rather than just collecting raw data. Only when high-level inferences become available can researchers obtain an increased context richness of their participants, resulting in new possibilities for ESM studies (see Figure 1).

Furthermore, the majority of the analysed tools offer little flexibility concerning customisation of study parameters. Again, those tools with more extensive parameter configuration possibilities require considerable technical configuration. While customisability of study parameters is often valuable, the currently found implementations reduce accessibility for a majority of researchers. What is lacking is an approach that allows for both a guided configuration-based on best practices, and a more extensive customisation possibility. This way, a researcher could easily configure the essential study considerations, while allowing for further customisation of other parameters when required.

Lastly, most software tools support only a subsection of available questionnaire input types (most common input types summarised in Table 2). While it is clear that the need for a checkbox is more common than the need for an affect grid, the absence of a certain input type in the software can act as a barrier for researchers aiming to apply the ESM in their study. A noticeable absent input type among the majority of analysed software tools was the recording of multimedia data (image, video, audio). The capture of these types of multimedia data has been shown to be a useful addition during data collection (Yue et al. 2014).

7 FUTURE OF MOBILE EXPERIENCE SAMPLING

Technological developments have improved both the reliability and possibilities of the ESM. Going forward, *mobile context awareness*, in which both sensing of and reacting to a certain context is enabled by the mobile device itself (Lovett and O'Neill 2012), will take on a larger role in experience sampling. In addition, the participants' input device will no longer be the only device used to construct the participants' context, as data from external devices and networks will be integrated to construct a richer context of the participant. Simultaneously, new technological challenges and methodological decisions arise as these technologies are introduced. Technological challenges include inter-device communication across devices and software systems, strict battery optimisation techniques, and different input techniques between devices. Examples of methodological questions include ensuring consistent data quality across devices, presentation of questionnaires on different form factors, and the timing of notifications based on a rich contextual understanding.

The effect of mobile ESM on scientific areas outside of computer science will be profound. While our review indicates that mobile sensors are actively used in the computer science discipline, this is not yet the case for other fields despite proven interest and opportunities (e.g., Ben-Zeev et al. (2015) and Raento et al. (2009)). As described in the previous section, there is a pressing need for accessible ESM study design tools that allow researchers outside of the computer science domain to configure advanced mobile instrumentation studies. These new tools will result in a large increase

in the usage of passive and active sensor data as researchers from fields such as psychology and medicine embed these new possibilities in their study designs.

7.1 User Interfaces for ESM

Developments in mobile platforms enable new input methods for ESM questionnaires. A recent example is the introduction of interactive notifications. These notifications allow users to quickly respond to incoming requests without leaving the application they are currently using, thus reducing participant strain and potentially increasing response rate. These interactive notifications can prove useful for questionnaires, especially when answer options are limited.

An alternative to the use of questionnaire notifications is the collection of self-reports as a by-product of another task. An example is the use of 'unlock journaling' (Zhang et al. 2016), in which unlocking the phones' screen is used to answer a single-item question through a sliding movement. The authors report an increase in frequency of answers, and reduced levels of experienced intrusiveness. We expect the popularity of this method to increase as new input types are implemented. The vast amount of daily smartphone unlocks and the short duration of smartphone usage sessions (van Berkel et al. 2016b) ensures a high number of potential answers over the day. We urge researchers to explore other creative input methods on mobile devices to reduce participant burden and open the way for longer deployment periods.

The use of social networks as a data collection tool has not yet been explored in the context of ESM. Recently, several large-scale social networks introduced interactive chat-bots (e.g., Facebook Messenger). These chat-bots can be configured to ask questions at certain time intervals, and even make use of natural language interpretation to follow-up on participant answers. An advantage of these chat-bots is that no additional application needs to be installed, and multiple operating platforms can easily be supported. On the other hand, participants may easily be distracted by the context of the application while completing a questionnaire (e.g., simultaneously talking with friends), or even delay using the application altogether as to prevent showing their online peers that they are available to chat. A second example is the use of intelligent personal assistants (e.g., Siri, Alexa). As these assistants are increasingly aware of their user's context, and can interact with users through multiple input modalities, a more intelligent and participant-friendly ESM configuration is within reach. The effect of novel user interfaces on participant responses will require further investigation.

Researchers have begun to explore the use of wearable devices (e.g., smartwatches) for notification purposes and the completion of ESM questionnaires. Limited screen estate constraints the possibility of both user input and content presentation (Baudisch and Chu 2009; Xiao et al. 2014). Hernandez et al. (2016) compare three different form factors (smartphone, smartwatch, and Google Glass) and find a significant difference in time between an incoming notification and initial user interaction (shortest with Google Glass, longest with smartphone), though no difference in total response time. Questionnaire completion times were significantly longer for the Google Glass compared to both the smartwatch and the smartphone.

7.2 Context Sensing

The collection of context-related information during ESM studies will continue to increase. Mobile devices will allow for the collection of new information types such as physiological data (e.g., heart-rate, eye tracking) and proximity-based information (e.g., local social networks through Wi-Fi Aware), allowing researchers to obtain new information on their participants. For example, physiological data can be used to assess the participant's engagement when answering a questionnaire—possibly providing an indication of their response quality and reliability. In addition to data collected on the participants' mobile device, the emergence of a multi-device environment will allow

for a far more detailed construction of a participants' context. Intille et al. (2003) already explored the possibilities of such a multi-device environment in the context of experience sampling, instrumenting participants' homes with a high number of sensors. As the number of sensing devices in both public and private space increases, the need for extensive instrumentation is reduced. In addition to increased sensing capabilities, a ubiquitous network can offload content presentation to different devices based on context and questionnaire content (e.g., proximity to participant, screen size in relation to content).

Referring to the matrix shown in Figure 1, wearable device constraints limit the opportunity for a high level of survey complexity. However, a moderate level of context richness is attainable, thus placing wearable devices in the quadrant *situated notification*. The level of context richness can be further extended through integration with the aforementioned multi-device environment. Traditional pen-and-paper questionnaires allow for high survey complexity at a low level of context richness (*repeated survey*), whereas mobile devices often combine the opportunity for both a high level of survey complexity and context richness (*situated inquiry*). The continuous evolvement of mobile devices will allow for an ever increases level of context richness. However, the possibility to increase questionnaire complexity is largely dependent on use of a multi-device environment. Larger screen sizes or novel input techniques are required to display and answer questionnaires of higher complexity than possible today.

8 CONCLUSION

The wide variety of studies using the ESM shows that the methodology is well suited and extensively applied to study an extensive array of human related phenomena. The use of participant-owned smartphones to answer ESM questionnaires has become possible due to widespread smartphone usage and available software tooling. This has reduced the strain on participants, as they no longer have to worry about carrying around additional study related objects (e.g., pen-and-paper, study-administrated PDA). In effect, this increases the ecological validity of the study employing this method. Today's mobile devices have amplified the possibilities and reliability of data collection, while introducing new opportunities for context (re)construction.

In this survey, we discussed the historical background of the ESM and its evolution parallel to the rise of mobile devices and ubiquitous computing. We then discussed the methodology's challenges and alternatives. We conducted a systematic literature review in the computer science domain and from this identified the prominent study parameters in an ESM study. Following this, we compare the various configurations that researchers apply in peer-reviewed publications. These methodological decisions were further analysed and discussed. We observe that the reporting of ESM study parameters and results has been largely inconsistent. Our analysis provides guidelines on how to address this inconsistency to allow for better comparison across studies.

The combination of human input data and automated sensor data collection has proven to be valuable and is capable of providing previously unattainable insights. Our review shows that the functionality of today's mobile devices remains underused in ESM studies in the computer science domain, with the literature reporting severe difficulties in other disciplines when it comes to the use of advanced functionalities of mobile devices—partly due to the lack of available software tools. This provides opportunities for the HCI discipline to bring (further) improvements to the application of this methodology. Going forward, we expect the possibilities of the ESM to continue to expand as new technological possibilities arise. Applicability and availability of accessible software will be a key-element to make this methodology available to researchers from a variety of disciplines with varying degrees of technical knowledge. With a continuing improvement of mobile device capabilities and an increased availability in all layers of society, we expect usage of this methodology to increase even more in the future. Finally, in Table 4, we provide researchers a

Table 4. Checklist for Researchers

Methodological Decisions to Consider and Report	Outcomes to Report
—*Notification Schedule.* Trigger mechanism, frequency, and any excluded hours or contexts. —*Inter-Notification Time.* Report even if no inter-notification time is established. —*Notification Expiry.* Report even if notifications do not expire. —*Inquiry Limit.* Report even if no maximum number of notifications is established. —*Device Ownership.* Participants' usage of personal or study-specific device. —*Participant Compensation.* Describe which strategy was followed, and how much participants were rewarded (even if no reward is provided). —*ESM Question Type.* Report the input type and any relevant parameters (e.g., number of points on Likert scale). Provide screenshots of the questions. —*Rich Media Collection.* Participants can collect visual and auditory material explicitly. This enriches the collected data, allowing for a richer qualitative insight into the daily experiences of the study participants. —*Validated Questionnaire Adaptation.* Report any validated questionnaires that were used, and describe how they were adapted (e.g., rephrased) for a mobile form factor. —*Advanced Question Logic.* Presented questionnaires can embed logical constructs. This way, questions can depend on previous input from the participant or the participant's current context.	—*Number of Participants.* Both prior to and following the data cleaning process. —*Study Duration.* In case of differences between participants report the average study duration (including SD). —*Response Rate.* Response rate after the data has been cleaned. —*Context Logged.* Report which sensors were used to collect data, at what frequency, and how many records were collected. —*Abnormalities.* Report any differences in collected results over the duration of the study, over the duration of an accumulated day, and between participants.

checklist of methodological decisions and outcomes to report when conducting an ESM study on mobile devices.

REFERENCES

F. B. Abdesslem, I. Parris, and T. Henderson. 2010. Mobile experience sampling: Reaching the parts of facebook other methods cannot reach. In *Proceedings of the Privacy and Usability Methods Pow-wow.*

G. D. Abowd, G. R. Hayes, G. Iachello, J. A. Kientz, S. N. Patel, M. M. Stevens, and K. N. Truong. 2005. Prototypes and paratypes: Designing mobile and ubiquitous computing applications. *IEEE Pervas. Computing*, 4, 4, 67–73. DOI: https://doi.org/10.1109/MPRV.2005.83

J. S. Adams. 1963. Toward an understanding of inequity. *J. Abnorm. Social Psychol.* 67, 5, 422–436. DOI: https://doi.org/10.1037/h0040968

P. Adams, M. Rabbi, T. Rahman, M. Matthews, A. Voida, G. Gay, T. Choudhury, and S. Voida. 2014. Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild. In *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare, ICST*, 72–79. DOI: https://doi.org/10.4108/icst.pervasivehealth.2014.254959

M. Alcañiz, A. Rodríguez, B. Rey, and E. Parra. 2014. Using serious games to train adaptive emotional regulation strategies. In *Proceedings of the International Conference on Social Computing and Social Media*, Springer International Publishing, 541–549. DOI: https://doi.org/10.1007/978-3-319-07632-4_51

D. Anthony, T. Henderson, and D. Kotz. 2007. Privacy in location-aware computing environments. *IEEE Pervas. Computing*, 6, 4, 64–72. DOI: https://doi.org/10.1109/MPRV.2007.83

K. Ara, N. Sato, S. Tsuji, Y. Wakisaka, N. Ohkubo, Y. Horry, N. Moriwaki, K. Yano, and M. Hayakawa. 2009. Predicting flow state in daily work through continuous sensing of motion rhythm. In *Proceedings of the International Conference on Networked Sensing Systems*, 1–6. DOI: https://doi.org/10.1109/INSS.2009.5409930

O. Asan and E. Montague. 2014. Using video-based observation research methods in primary care health encounters to evaluate complex interactions. *Informat. Primary Care*, 21, 4, 161–170. DOI: https://doi.org/10.14236/jhi.v21i4.72

M. Ashour, K. Bekiroglu, C. H. Yang, C. Lagoa, D. Conroy, J. Smyth, and S. Lanza. 2016. On the mathematical modeling of the effect of treatment on human physical activity. In *Proceedings of the IEEE Conference on Control Applications (CCA)*, 1084–1091. DOI: https://doi.org/10.1109/CCA.2016.7587951

Y. Ayzenberg and R. W. Picard. 2014. FEEL: A system for frequent event and electrodermal activity labeling. *IEEE J. Biomed. Health Informat.* 18, 1, 266–277. DOI: https://doi.org/10.1109/JBHI.2013.2278213

S. Barclay, C. Todd, I. Finlay, G. Grande, and P. Wyatt. 2002. Not another questionnaire! Maximizing the response rate, predicting non-response and assessing non-response bias in postal questionnaire studies of GPs. *Family Pract.* 19 1, 105–111. DOI: https://doi.org/10.1093/fampra/19.1.105

L. F. Barrett and D. J. Barrett. 2001. An introduction to computerized experience sampling in psychology. *Soc. Sci. Comput. Rev.* 19, 2, 175–185. DOI: https://doi.org/10.1177/089443930101900204

W. Barta, H. Tennen, and M. Litt. 2012. Measurement reactivity in diary research. In *Handbook of Research Methods for Studying Daily Life*, M. R. Mehl, and T. S. Conner (Eds.). Guilford Press, New York.

P. Baudisch and G. Chu. 2009. Back-of-device interaction allows creating very small touch devices. In *Proceedings of the Conference on Human Factors in Computing Systems*, ACM, 1923–1932. DOI: https://doi.org/10.1145/1518701.1518995

D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehab. J.*, 38, 3, 218–226. DOI: https://doi.org/10.1037/prj0000130

G. E. Bevans. 1913. *How Workingmen Spend Their time*, Columbia University Press (1913).

N. Bolger and J.-P. P. Laurenceau. 2013. *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*, Guilford Press (2013).

J. E. Broderick, J. E. Schwartz, S. Shiffman, M. R. Hufford, and A. A. Stone. 2003. Signaling does not adequately improve diary compliance. *Ann. Behav. Med.* 26, 2, 139–148. DOI: https://doi.org/10.1207/S15324796ABM2602_06

C. J. Burgin, P. J. Silvia, K. M. Eddington, and T. R. Kwapil. 2013. Palm or cell? Comparing personal digital assistants and cell phones for experience sampling research. *Soc. Sci. Comput. Rev.*, 31, 2, 244–251. DOI: https://doi.org/doi:10.1177/0894439312441577

D. Buschek, F. Hartmann, E. V. Zezschwitz, A. D. Luca, and F. Alt. 2016. SnapApp: Reducing authentication overhead with a time-constrained fast unlock option. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, 3736–3747. DOI: https://doi.org/10.1145/2858036.2858164

K. Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, 981–992. DOI: https://doi.org/10.1145/2858036.2858498

S. Carter and J. Mankoff. 2005. When participants do the capturing: The role of media in diary studies. In *Proceedings of the Conference on Human Factors in Computing Systems*, ACM, 899–908. DOI: https://doi.org/10.1145/1054972.1055098

Y.-J. J. Chang, G. Paruthi, and M. W. Newman. 2015. A field study comparing approaches to collecting annotated activity data in real-world settings. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 671–682. DOI: https://doi.org/10.1145/2750858.2807524

K. Church, M. Cherubini, and N. Oliver. 2014. A large-scale study of daily information needs captured in situ. *ACM Trans. Comput.-Hum. Interact.* 21, 2, 1–46. DOI: https://doi.org/10.1145/2552193

K. Church and R. de Oliveira. 2013. What's up with whatsapp?: Comparing mobile instant messaging behaviors with traditional SMS. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, 352–361. DOI: https://doi.org/10.1145/2493190.2493225

K. Church, D. Ferreira, N. Banovic, and K. Lyons. 2015. Understanding the challenges of mobile phone usage data. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*, 505–514. DOI: https://doi.org/10.1145/2785830.2785891

M. Ciman and K. Wac. 2016. Individuals' stress assessment using human-smartphone interaction analysis. *IEEE Trans. Affect. Comput.* DOI: https://doi.org/10.1109/TAFFC.2016.2592504

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

T. Conner Christensen, L. Feldman Barrett, E. Bliss-Moreau, K. Lebo, and C. Kaschub. 2003. A practical guide to experience-sampling procedures. *J. Happiness Stud.* 4, 1, 53–78. DOI: https://doi.org/10.1023/A:1023609306024

T. Conner. 2015. Experience sampling and ecological momentary assessment with mobile phones. Retrieved 20 February 2017 from http://www.otago.ac.nz/psychology/otago047475.pdf.

S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. 2005. Location disclosure to social relations: Why, when, & what people want to share. In *Proceedings of the Conference on Human Factors in Computing Systems*, ACM, 81–90. DOI: https://doi.org/10.1145/1054972.1054985

S. Consolvo and M. Walker. 2003. Using the experience sampling method to evaluate Ubicomp applications. *IEEE Pervas. Comput.* 2, 2, 24–31. DOI: https://doi.org/10.1109/MPRV.2003.1203750

P. M. Costa, J. Pitt, T. Galvão, and J. F. e. Cunha. 2013. Assessing contextual mood in public transport: A pilot study. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, ACM, 498–503. DOI: https://doi.org/10.1145/2493190.2494429

L. Cowan, W. G. Griswold, L. Barkhuus, and J. D. Hollan. 2010. Engaging the periphery for visual communication on mobile phones. In *Proceedings of the Hawaii International Conference on System Sciences*, 1–10. DOI: https://doi.org/10.1109/HICSS.2010.184

M. Csikszentmihalyi, R. Larson, and S. Prescott. 1977. The ecology of adolescent activity and experience. *J. Youth Adoles.* 6, 3, 281–294. DOI: https://doi.org/10.1007/BF02138940

E. Cutrell, M. Czerwinski, and E. Horvitz. 2001. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Proceedings of the Human-Computer Interaction – INTERACT*, 263–269.

A. K. Dey. 2001. Understanding and using context. *Pers. Ubiq. Comput.* 5, 1, 4–7. DOI: https://doi.org/10.1007/s007790170019

T. Dingler and M. Pielot. 2015. I'll be there for you: Quantifying attentiveness towards mobile messaging. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 1–5. DOI: https://doi.org/10.1145/2785830.2785840

U. W. Ebner-Priemer and T. J. Trull. 2009. Ecological momentary assessment of mood disorders and mood dysregulation. *Psychol. Assess.* 21, 4, 463–475. DOI: https://doi.org/10.1037/a0017075

R. M. Ellingson and B. Oken. 2011. Ambulatory physiologic monitoring system supporting EMA with self-administered visual evoked potential recording at randomized intervals. In *Proceedings of the International Instrumentation and Measurement Technology Conference*. IEEE, 1–4. DOI: https://doi.org/10.1109/IMTC.2011.5944091

A. Faiola and P. Srinivas. 2014. Extreme mediation: Observing mental and physical health in everyday life. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 47–50. DOI: https://doi.org/10.1145/2638728.2638741

D. Ferreira, J. Goncalves, V. Kostakos, L. Barkhuus, and A. K. Dey. 2014. Contextual experience sampling of mobile application micro-usage. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 91–100. DOI: https://doi.org/10.1145/2628363.2628367

D. Ferreira, V. Kostakos, A. R. Beresford, J. Lindqvist, and A. K. Dey. 2015a. Securacy: An empirical investigation of android applications' network usage, privacy and security. In *Proceedings of the Conference on Security and Privacy in Wireless and Mobile Networks*. ACM, 1–11. DOI: https://doi.org/10.1145/2766498.2766506

D. Ferreira, V. Kostakos, and A. K. Dey. 2015b. AWARE: Mobile context instrumentation framework. *Frontiers in ICT*, 2, 6, 1–9. DOI: https://doi.org/10.3389/fict.2015.00006

J. E. Fischer and S. Benford. 2009. Inferring player engagement in a pervasive experience. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 1903–1906. DOI: https://doi.org/10.1145/1518701.1518993

J. E. Fischer, C. Greenhalgh, and S. Benford. 2011. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 181–190. DOI: https://doi.org/10.1145/2037373.2037402

J. E. Fischer, N. Yee, V. Bellotti, N. Good, S. Benford, and C. Greenhalgh. 2010. Effects of content and time of delivery on receptivity to mobile interruptions. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 103–112. DOI: https://doi.org/10.1145/1851600.1851620

R. Fisher and R. Simmons. 2011. Smartphone interruptibility using density-weighted uncertainty sampling with reinforcement learning. In *Proceedings of the International Conference on Machine Learning and Applications and Workshops*. 436–441. DOI: https://doi.org/10.1109/ICMLA.2011.128

R. R. Fletcher, S. Tam, O. Omojola, R. Redemske, and J. Kwan. 2011. Wearable sensor platform and mobile application for use in cognitive behavioral therapy for drug addiction and PTSD. In *Proceedings of the Engineering in Medicine and Biology Society*. IEEE, 1802–1805. DOI: https://doi.org/10.1109/IEMBS.2011.6090513

J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. 2007. MyExperience: A system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the International Conference on Mobile Systems, Applications and Services*. ACM, 57–70. DOI: https://doi.org/10.1145/1247660.1247670

J. Froehlich, M. Y. Chen, I. E. Smith, and F. Potter. 2006. Voting with your feet: An investigative study of the relationship between place visit behavior and preference, In *Proceedings of UbiComp 2006: Ubiquitous Computing*, P. Dourish, and A. Friday (Eds.). Springer, Berlin.

A. Gaggioli, G. Pioggia, G. Tartarisco, G. Baldus, D. Corda, P. Cipresso, and G. Riva. 2013. A mobile data collection platform for mental health research. *Pers. Ubiq. Computing*, 17, 2, 241–251. DOI: https://doi.org/10.1007/s00779-011-0465-2

S. Ghosh, V. Chauhan, N. Ganguly, B. Mitra, and P. De. 2015. Impact of experience sampling methods on tap pattern based emotion recognition. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the International Symposium on Wearable Computers (Adjunct)*. ACM, 713–722. DOI: https://doi.org/10.1145/2800835.2804396

P. Gomes, M. Kaiseler, C. Queirós, M. Oliveira, B. Lopes, and M. Coimbra. 2012. Vital analysis: Annotating sensed physiological signals with the stress levels of first responders in action. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 6695–6698. DOI: https://doi.org/10.1109/EMBC.2012.6347530

A. L. Gonzales. 2014. Text-based communication influences self-esteem more than face-to-face or cellphone communication. *Comput. Human Behavior*, 39, 197–203. DOI: https://doi.org/10.1016/j.chb.2014.07.026

R. Gouveia and E. Karapanos. 2013. Footprint tracker: Supporting diary studies with lifelogging. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 2921–2930. DOI: https://doi.org/10.1145/2470654.2481405

S. Grandhi and Q. Jones. 2010. Technology-mediated interruption management. *Int. J. Hum.-Comput. Stud.* 68, 5. 288–306. DOI: https://doi.org/10.1016/j.ijhcs.2009.12.005

S. A. Grandhi and Q. Jones. 2015. Knock, knock! Who's there? Putting the user in control of managing interruptions. *Int. J. Hum.-Comput. Stud.* 79, 35–50. DOI: https://doi.org/10.1016/j.ijhcs.2015.02.008

GSMA. 2016. Mobile Economy 2016. Retrieved 20 February 2017 from http://gsmamobileeconomy.com/global/.

S. Guha and S. B. Wicker. 2015. Spatial subterfuge: An experience sampling study to predict deceptive location disclosures. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1131–1135. DOI: https://doi.org/10.1145/2750858.2804281

M. Gustarini, K. Wac, and A. K. Dey. 2016. Anonymous smartphone data collection: Factors influencing the users' acceptance in mobile crowd sensing. *Pers. Ubiq. Comput.* 20, 1, 65–82. DOI: https://doi.org/10.1007/s00779-015-0898-0

J. P. Haas and E. L. Larson. 2007. Measurement of compliance with hand hygiene. *J. Hosp. Inf.* 66, 1, 6–14. DOI: https://doi.org/10.1016/j.jhin.2006.11.013

A. A. Haedt-Matt and P. K. Keel. 2011. Revisiting the affect regulation model of binge eating: A meta-analysis of studies using ecological momentary assessment. *Psychol. Bull.* 137, 4, 660–681. DOI: https://doi.org/10.1037/a0023660

M. Harbach, E. von Zezschwitz, A. Fichtner, A. De Luca, and M. Smith. 2014. It's a hard lock life: A field study of smartphone (un)locking behavior and risk perception. In *Proceedings of the Symposium on Usable Privacy and Security*. 213–230.

D. H. Hareva, K. Tomoki, O. Hisao, N. Takao, O. Hiroki, and K. Hiromi. 2007. Development of real-time biological data collection system using a cellular phone. In *Proceedings of the SICE Annual Conference 2007*. 316–321. DOI: https://doi.org/10.1109/SICE.2007.4420999

S. S. Hasan, R. Brummet, O. Chipara, Y. H. Wu, and T. Yang. 2015. In-situ measurement and prediction of hearing aid outcomes using mobile phones. In *Proceedings of the International Conference on Healthcare Informatics*. 525–534. DOI: https://doi.org/10.1109/ICHI.2015.101

S. S. Hasan, O. Chipara, Y.-H. Wu, and N. Aksan. 2014. Evaluating auditory contexts and their impacts on hearing aid outcomes with mobile phones. In *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare*. ICST, 126–133. DOI: https://doi.org/10.4108/icst.pervasivehealth.2014.254952

S. S. Hasan, F. Lai, O. Chipara, and Y.-H. Wu. 2013. Audiosense: Enabling real-time evaluation of hearing aid technology in-situ. In *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems*. 167–172. DOI: https://doi.org/10.1109/CBMS.2013.6627783

J. M. Hektner, J. A. Schmidt, and M. Csikszentmihalyi. 2007. *Experience Sampling Method: Measuring the Quality of Everyday Life*. Sage (2007).

J. Hernandez, D. McDuff, C. Infante, P. Maes, K. Quigley, and R. Picard. 2016. Wearable ESM: Differences in the experience sampling method across wearable devices. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 195–205. DOI: https://doi.org/10.1145/2935334.2935340

K. E. Heron and J. M. Smyth. 2010. Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behavior treatments. *Brit. J. Health Psychol.* 15, 1, 1–39. DOI: https://doi.org/10.1348/135910709X466063

S. E. Hormuth. 1986. The sampling of experiences in situ. *J. Pers.* DOI: https://doi.org/10.1111/j.1467-6494.1986.tb00395.x

G. Hsieh, I. Li, A. Dey, J. Forlizzi, and S. E. Hudson. 2008. Using visualizations to increase compliance in experience sampling. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 164–167. DOI: https://doi.org/10.1145/1409635.1409657

S. Ickin, K. Wac, and M. Fiedler. 2013. QoE-based energy reduction by controlling the 3G cellular data traffic on the smartphone. In *Proceedings of the ITC Specialist Seminar on Energy Efficient and Green Networking*. 13–18. DOI: https://doi.org/10.1109/SSEEGN.2013.6705396

S. Ickin, K. Wac, M. Fiedler, L. Janowski, J.-H. H. Hong, and A. K. Dey. 2012. Factors influencing quality of experience of commonly used mobile applications. *IEEE Communications Magazine*, 50, 4, 48–56. DOI: https://doi.org/10.1109/MCOM.2012.6178833

M. Iida, P. E. Shrout, J.-P. P. Laurenceau, and N. Bolger. 2012. Using diary methods in psychological research. *APA PsycNET*. DOI: https://doi.org/10.1037/13619-016

S. Intille, C. Haynes, D. Maniar, A. Ponnada, and J. Manjourides. 2016. μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1124–1128. DOI: https://doi.org/10.1145/2971648.2971717

S. S. Intille, E. M. Tapia, J. Rondoni, J. Beaudin, C. Kukla, S. Agarwal, L. Bao, and K. Larson. 2003. Tools for studying behavior and technology in natural settings, in ubicomp 2003: ubiquitous computing. UbiComp 2003. Lecture Notes in Computer Science, vol. 2864, A. K. Dey, A. Schmidt, and J. F. McCarthy (Eds.). Springer, Berlin, Heidelberg.

E. Isaacs, A. Konrad, A. Walendowski, T. Lennig, V. Hollis, and S. Whittaker. 2013. Echoes from the past: How technology mediated reflection improves well-being. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 1071–1080. DOI: https://doi.org/10.1145/2470654.2466137

L. A. Jelenchick, J. C. Eickhoff, and M. A. Moreno. 2013. "Facebook depression?" Social networking site use and depression in older adolescents. *J. Adoles. Health.* 52, 1, 128–130. DOI: https://doi.org/10.1016/j.jadohealth.2012.05.008

S. K. Johansen and A. M. Kanstrup. 2016. Expanding the locus of control: Design of a mobile quantified self-tracking application for whiplash patients. In *Proceedings of the Nordic Conference on Human-Computer Interaction*, ACM, 1–10. DOI: https://doi.org/10.1145/2971485.2971497

D. Kahneman, A. B. Krueger, D. A. Schkade, N. Schwarz, and A. A. Stone. 2004. A survey method for characterizing daily life: The day reconstruction method. *Science (New York, N.Y.)*, 306 (5702), 1776–1780. DOI: https://doi.org/10.1126/science.1103572

C. Karr. 2015. Purple robot. Retrieved 20 February 2017 from http://tech.cbits.northwestern.edu/purple-robot/.

M. Kay, G. L. Nelson, and E. B. Hekler. 2016. Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the Conference on Human Factors in Computing Systems*, ACM, 4521–4532. DOI: https://doi.org/10.1145/2858036.2858465

I. Ketykó, K. D. Moor, W. Joseph, L. Martens, and L. D. Marez. 2010. Performing QoE-measurements in an actual 3G network. In *Proceedings of the International Symposium on Broadband Multimedia Systems and Broadcasting*. IEEE, 1–6. DOI: https://doi.org/10.1109/ISBMSB.2010.5463132

A. Khalil and K. Connelly. 2006. Context-aware telephony: Privacy preferences and sharing patterns. In *Proceedings of the Conference on Computer Supported Cooperative Work*. ACM, 469–478. DOI: https://doi.org/10.1145/1180875.1180947

V.-J. J. Khan, P. Markopoulos, B. Eggen, W. IJsselsteijn, and B. de Ruyter. 2008. Reconexp: A way to reduce the data loss of the experiencing sampling method. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 471–476. DOI: https://doi.org/10.1145/1409240.1409316

V. J. Khan, P. Markopoulos, and B. Eggen. 2009. An experience sampling study into awareness needs of busy families. In *Proceedings of the Conference on Human System Interactions*. 338–343. DOI: https://doi.org/10.1109/HSI.2009.5091002

J. Kim, T. Nakamura, H. Kikuchi, and Y. Yamamoto. 2015a. Psychobehavioral validity of self-reported symptoms based on spontaneous physical activity. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 4021–4024. DOI: https://doi.org/10.1109/EMBC.2015.7319276

J. Kim, T. Nakamura, H. Kikuchi, K. Yoshiuchi, T. Sasaki, and Y. Yamamoto. 2015b. Covariation of depressive mood and spontaneous physical activity in major depressive disorder: Toward continuous monitoring of depressive mood. *IEEE J. Biomed. Health Informat.* 19, 4, 1347–1355. DOI: https://doi.org/10.1109/JBHI.2015.2440764

J. Kim, J. J. Tran, T. W. Johnson, R. Ladner, E. Riskin, and J. O. Wobbrock. 2011. Effect of mobileasl on communication among deaf users. In *Proceedings of the Conference on Human Factors in Computing Systems (Extended Abstracts)*, ACM, 2185–2190. DOI: https://doi.org/10.1145/1979742.1979872

P. Klasnja, B. L. Harrison, L. LeGrand, A. LaMarca, J. Froehlich, and S. E. Hudson. 2008. Using wearable sensors and real time inference to understand human recall of routine activities. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 154–163. DOI: https://doi.org/10.1145/1409635.1409656

C. Korunka, R. Prem, and B. Kubicek. 2012. Diary studies as a macro-ergonomic evaluation tool: Development of a shift diary and its application in ergonomic evaluations. In *Proceedings of the Southeast Asian Network of Ergonomics Societies Conference*. 1–6. DOI: https://doi.org/10.1109/SEANES.2012.6299552

S. Kvale. *Doing Interviews*, SAGE (2007).

R. Larson and M. Csikszentmihalyi. 1983. The experience sampling method, In *Flow and the Foundations of Positive Psychology*, M. Csikszentmihalyi (Eds.). Wiley Jossey-Bass.

N. Lathia, K. K. Rachuri, C. Mascolo, and P. J. Rentfrow. 2013. Contextual dissonance: Design bias in sensor-based experience sampling methods. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 183–192. DOI: https://doi.org/10.1145/2493432.2493452

J. A. Lee, C. Efstratiou, and L. Bai. 2016. OSN mood tracking: Exploring the use of online social network activity as an indicator of mood changes. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (Adjunct)*. ACM, 1171–1179. DOI: https://doi.org/10.1145/2968219.2968304

S. Lee, J. Seo, and G. Lee. 2010. An adaptive speed-call list algorithm and its evaluation with ESM. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 2019–2022. DOI: https://doi.org/10.1145/1753326.1753632

B. Lepri, J. Staiano, G. Rigato, K. Kalimeri, A. Finnerty, F. Pianesi, N. Sebe, and A. Pentland. 2012. The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations. In *Proceedings of the International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*. 623–628. DOI: https://doi.org/10.1109/SocialCom-PASSAT.2012.71

G. Liang, J. Cao, and W. Zhu. 2013. CircleSense: A pervasive computing system for recognizing social activities. In *Proceedings of the International Conference on Pervasive Computing and Communications*. IEEE, 201–206. DOI: https://doi.org/10.1109/PerCom.2013.6526733

M. Linnap and A. Rice. 2014. The effectiveness of centralised management for reducing wasted effort in participatory sensing. In *Proceedings of the International Conference on Pervasive Computing and Communication (Adjunct)*. IEEE, 68–73. DOI: https://doi.org/10.1109/PerComW.2014.6815167

Y. Liu, J. Goncalves, D. Ferreira, B. Xiao, S. Hosio, and V. Kostakos. 2014a. CHI 1994-2013: Mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the Conference on Human Factors in Computing Systems*. 3553–3562. DOI: https://doi.org/10.1145/2556288.2556969

Z. Liu, J. Shan, R. Bonazzi, and Y. Pigneur. 2014b. Privacy as a tradeoff: Introducing the notion of privacy calculus for context-aware mobile applications. In *Proceedings of the Hawaii International Conference on System Sciences*. 1063–1072. DOI: https://doi.org/10.1109/HICSS.2014.138

V. López, L. Ahumada, S. Galdames, and R. Madrid. 2012. School principals at their lonely work: Recording workday practices through ESM logs. *Computers & Education*, 58, 1. 413–422. DOI: https://doi.org/10.1016/j.compedu.2011.07.014

T. Lovett and E. O'Neill. 2012. *Mobile Context Awareness*, Springer (2012).

P. Lynn. 2001. The impact of incentives on response rates to personal interview surveys: Role and perceptions of interviewers. *Int. J. Pub. Opin. Res.* 13, 3, 326–336. DOI: https://doi.org/10.1093/ijpor/13.3.326

T. Maekawa, N. Yamashita, and Y. Sakurai. 2016. How well can a user's location privacy preferences be determined without using GPS location data? *IEEE Trans. Emerg. Topics Comput.* DOI: https://doi.org/10.1109/TETC.2014.2335537

C. Mancini, K. Thomas, Y. Rogers, B. A. Price, L. Jedrzejczyk, A. K. Bandara, A. N. Joinson, and B. Nuseibeh. 2009. From spaces to places: Emerging contexts in mobile privacy. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. 1–10. DOI: https://doi.org/10.1145/1620545.1620547

P. Markopoulos, N. Batalas, and A. Timmermans. 2015. On the use of personalization to enhance compliance in experience sampling. In *Proceedings of the European Conference on Cognitive Ergonomics*. ACM. DOI: https://doi.org/10.1145/2788412.2788427

A. Maxhuni, A. Matic, V. Osmani, and O. M. Ibarra. 2011. Correlation between self-reported mood states and objectively measured social interactions at work: A pilot study. In *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare (Adjunct)*. 308-311. DOI: https://doi.org/10.4108/icst.pervasivehealth.2011.246136

J. M. Mayer, S. R. Hiltz, L. Barkhuus, K. Väänänen, and Q. Jones. 2016. Supporting opportunities for context-aware xocial matching: An experience sampling study. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 2430–2441. DOI: https://doi.org/10.1145/2858036.2858175

K. O. McCabe, L. Mack, and W. Fleeson. 2011. A guide for data cleaning in experience sampling studies, In *Handbook of Research Methods for Studying Daily Life*, M. R. Mehl, and T. S. Conner (Eds.). Guilford Press, New York.

A. Mehrotra, M. Musolesi, R. Hendley, and V. Pejovic. 2015. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 813–824. DOI: https://doi.org/10.1145/2750858.2807544

A. Mehrotra, V. Pejovic, J. Vermeulen, R. Hendley, and M. Musolesi. 2016. My phone and me: Understanding people's receptivity to mobile notifications. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 1021–1032. DOI: https://doi.org/10.1145/2858036.2858566

A. Meschtscherjakov, A. Weiss, and T. Scherndl. 2009. Utilizing emoticons on mobile devices within ESM studies to measure emotions in the field. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services (Adjunct)*.

K. Mihalic and M. Tscheligi. 2007. 'Divert: Mother-in-law': Representing and evaluating social context on mobile devices. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 257–264. DOI: https://doi.org/10.1145/1377999.1378016

T. R. Mitchell, L. Thompson, E. Peterson, and R. Cronk. 1997. Temporal adjustments in the evaluation of events: The "rosy view". *J. Experim. Soc. Psychol.* 33, 4, 421–448. DOI: https://doi.org/10.1006/jesp.1997.1333

T. Miu, P. Missier, and T. Plötz. 2015. Bootstrapping personalised human activity recognition models using online active learning. In *Proceedings of the Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*. IEEE, 1138–1147. DOI: https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.170

M. A. Moreno, L. Jelenchick, R. Koff, J. Eikoff, C. Diermyer, and D. A. Christakis. 2012. Internet use and multitasking among older adolescents: An experience sampling approach. *Comput. Hum. Behav.* 28, 4, 1097–1102. DOI: https://doi.org/10.1016/j.chb.2012.01.016

S. Motahari, S. Ziavras, and Q. Jones. 2009a. Preventing unwanted social inferences with classification tree analysis. In *Proceedings of the International Conference on Tools with Artificial Intelligence,* IEEE, 500–507. DOI: https://doi.org/10.1109/ICTAI.2009.15

S. Motahari, S. Ziavras, M. Naaman, M. Ismail, and Q. Jones. 2009b. Social inference risk modeling in mobile and social applications. In *Proceedings of the International Conference on Computational Science and Engineering*, 125–132. DOI: https://doi.org/10.1109/CSE.2009.237

S. T. Moturu, I. Khayal, N. Aharony, W. Pan, and A. Pentland. 2011a. Sleep, mood and sociability in a healthy population. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 5267–5270. DOI: https://doi.org/10.1109/IEMBS.2011.6091303

S. T. Moturu, I. Khayal, N. Aharony, W. Pan, and A. Pentland. 2011b. Using social sensing to understand the links between sleep, mood, and sociability. In *Proceedings of the International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*. IEEE, 208–214. DOI: https://doi.org/10.1109/PASSAT/SocialCom.2011.200

H. Muukkonen, K. Hakkarainen, M. Inkinen, K. Lonka, and K. Salmela-Aro. 2008. CASS-methods and tools for investigating higher education knowledge practices. In *Proceedings of the International Conference on International Conference for the Learning Sciences*. International Society of the Learning Sciences, 107–114.

I. Myin-Germeys, M. Oorschot, D. Collip, J. Lataster, P. Delespaul, and J. van Os. 2009. Experience sampling research in psychopathology: Opening the black box of daily life. *Psychol. Med.* 39, 9, 1533–1547. DOI: https://doi.org/10.1017/S0033291708004947

T. Nguyen, S. Gupta, S. Venkatesh, and D. Phung. 2014. A Bayesian nonparametric framework for activity recognition using accelerometer data. In *Proceedings of the International Conference on Pattern Recognition*. 2017–2022. DOI: https://doi.org/10.1109/ICPR.2014.352

T. Nguyen, S. Gupta, S. Venkatesh, and D. Phung. 2016. Nonparametric discovery of movement patterns from accelerometer signals. *Patt. Recognition Lett.* 70, 52–58. DOI: https://doi.org/10.1016/j.patrec.2015.11.003

E. Niforatos and E. Karapanos. 2014. EmoSnaps: A mobile application for emotion recall from facial expressions. *Pers. Ubiq. Comput.* 19, 2, 425–444. DOI: https://doi.org/10.1007/s00779-014-0777-0

PACO. 2016. PACO - The personal analytics companion. Retrieved 20 February 2017 from https://www.pacoapp.com/.

W. K. Park. 2005. *Mobile Phone Addiction, in Mobile Communications*. R. Ling, and P. E. Pedersen (Eds.). Springer, London.

J. Pärkkä, J. Merilahti, E. M. Mattila, E. Malm, K. Antila, M. T. Tuomisto, A. V. Saarinen, M. V. Gils, and I. Korhonen. 2009. Relationship of psychological and physiological variables in long-term self-monitored data during work ability rehabilitation program. *IEEE Trans. Inf. Tech. Biomed.*, 13, 2, 141–151. DOI: https://doi.org/10.1109/TITB.2008.2007078

S. Patil, R. Hoyle, R. Schlegel, A. Kapadia, and A. J. Lee. 2015. Interrupt now or inform later?: Comparing immediate and delayed privacy feedback. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 1415–1418. DOI: https://doi.org/10.1145/2702123.2702165

S. Patil, R. Schlegel, A. Kapadia, and A. J. Lee. 2014. Reflection or action?: How feedback and control affect location sharing decisions. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 101–110. DOI: https://doi.org/10.1145/2556288.2557121

V. Pejovic, N. Lathia, C. Mascolo, and M. Musolesi. 2016. Mobile-based experience sampling for behaviour research, In *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, and A. Košir (Eds.). Springer International Publishing, Cham.

V. Pejovic and M. Musolesi. 2014. InterruptMe: Designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 897–908. DOI: https://doi.org/10.1145/2632048.2632062

E. Pergler, R. Hable, E. Rico-Schmidt, C. Kittl, and R. Schamberger. 2014. A context-sensitive tool to support mobile technology acceptance research. In *Proceedings of the Hawaii International Conference on System Sciences*. 1015–1022. DOI: https://doi.org/10.1109/HICSS.2014.133

T. D. Pessemier, K. D. Moor, A. Juan, W. Joseph, L. D. Marez, and L. Martens. 2011. Quantifying QoE of mobile video consumption in a real-life setting drawing on objective and subjective parameters. In *Proceedings of the International Symposium on Broadband Multimedia Systems and Broadcasting*. IEEE, 1–6. DOI: https://doi.org/10.1109/BMSB.2011.5954937

M. Pielot, K. Church, and R. de Oliveira. 2014. An in-situ study of mobile phone notifications. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices & Services*. ACM, 233–242. DOI: https://doi.org/10.1145/2628363.2628364

M. Pielot, T. Dingler, J. S. Pedro, and N. Oliver. 2015. When attention is not scarce - detecting boredom from mobile phone usage. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 825–836. DOI: https://doi.org/10.1145/2750858.2804252

PsyMate. 2016. PsyMate [EN]. Retrieved 20 February 2017 from http://www.psymate.eu/psymate–en-.html.

M. Raento, A. Oulasvirta, and N. Eagle. 2009. Smartphones: An emerging tool for social scientists. *Sociol. Meth. Res.* 37, 3, 426–454. DOI: https://doi.org/10.1177/0049124108330005

W. M. Randall and N. S. Rickard. 2013. Development and trial of a mobile experience sampling method (m-ESM) for personal music listening. *Music Perception: An Interdisciplinary Journal*. 31, 2, 157–170. DOI: https://doi.org/10.1525/mp.2013.31.2.157

H. T. Reis and S. L. Gable. 2000. Event sampling and other methods for studying everyday experience, In *Handbook of Research methods in Social and Personality Psychology*, H. T. Reis, and C. M. Judd (Eds.). Cambridge University Press, New York, NY, US.

S. Reyal, S. Zhai, and P. O. Kristensson. 2015. Performance and user experience of touchscreen and gesture keyboards in a lab setting and in the wild. In *Proceedings of the Conference on Human Factors in Computing Systems*, ACM, 679–688. DOI: https://doi.org/10.1145/2702123.2702597

A. Rieger, S. Neubert, S. Behrendt, M. Weippert, S. Kreuzfeld, and R. Stoll. 2012. 24-Hour ambulatory monitoring of complex physiological parameters with a wireless health system: Feasibility, user compliance and application. In *Proceedings of the International Multi-Conference on Systems, Sygnals & Devices*, 1–3. DOI: https://doi.org/10.1109/SSD.2012.6198122

S. Rosenthal, A. K. Dey, and M. Veloso. 2011. Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In *Proceedings of the International Conference on Pervasive Computing*. Springer-Verlag, 170–187.

D. Rough and A. Quigley. 2015. Jeeves – a visual programming environment for mobile experience sampling. In *Proceedings of the Visual Languages and Human-Centric Computing (Symposium)*. IEEE, 121–129. DOI: https://doi.org/10.1109/VLHCC.2015.7357206

M. Sabatelli, V. Osmani, O. Mayora, A. Gruenerbl, and P. Lukowicz. 2014. Correlation of significant places with self-reported state of bipolar disorder patients. In *Proceedings of the International Conference on Wireless Mobile Communication and Healthcare*. 116–119. DOI: https://doi.org/10.1109/MOBIHEALTH.2014.7015923

J. B. Sabra, H. J. Andersen, and K. Rodil. 2015. Hybrid cemetery culture: Making death matter in cultural heritage using smart mobile technologies. In *Proceedings of the International Conference on Culture and Computing*. 167–174. DOI: https://doi.org/10.1109/Culture.and.Computing.2015.16

S. Saeb, Z. Mi, M. Kwasny, C. J. Karr, K. Kording, and D. C. Mohr. 2015. The relationship between clinical, momentary, and sensor-based assessment of depression. In *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare*. 229–232. DOI: https://doi.org/10.4108/icst.pervasivehealth.2015.259034

A. Sahami Shirazi, N. Henze, T. Dingler, M. Pielot, D. Weber, and A. Schmidt. 2014. Large-scale assessment of mobile notifications. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 3055–3064. DOI: https://doi.org/10.1145/2556288.2557189

M. C. Sala, K. Partridge, L. Jacobson, and J. B. Begole. 2007. An exploration into activity-informed physical advertising using PEST, In *Pervasive Computing*. A. LaMarca, M. Langheinrich, and K. N. Truong (Eds.). Springer, Berlin.

C. E. Schwartz, M. A. G. Sprangers, A. Carey, and G. Reed. 2004. Exploring response shift in longitudinal data. *Psychology & Health*, 19, 1, 51–69. DOI: https://doi.org/10.1080/0887044031000118456

C. N. Scollon, C. Kim-Prieto, and E. Diener. 2003. Experience sampling: promises and pitfalls, strengths and weaknesses. *J. Happiness Stud.* 4, 1, 5–34. DOI: https://doi.org/10.1023/A:1023605205115

E. Seto, J. Hua, L. Wu, A. Bestick, V. Shia, S. Eom, J. Han, M. Wang, and Y. Li. 2014. The Kunming Calfit study: Modeling dietary behavioral patterns using smartphone data. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 6884–6887. DOI: https://doi.org/10.1109/EMBC.2014.6945210

S. Shiffman. 2009. Ecological momentary assessment (EMA) in studies of substance use. *Psychol. Assess.* 21, 4, 486–497. DOI: https://doi.org/10.1037/a0017074

S. Shiffman, A. A. Stone, and M. R. Hufford. 2008. Ecological momentary assessment. *Ann. Rev. Clin. Psychol.*, 4, 1–32.

F. Shih, I. Liccardi, and D. Weitzner. 2015. Privacy tipping points in smartphones privacy preferences. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 807–816. DOI: https://doi.org/10.1145/2702123.2702404

J. Smith, A. Lavygina, J. Ma, A. Russo, and N. Dulay. 2014. Learning to recognise disruptive smartphone notifications. In *Proceedings of the International Conference on Human-computer Interaction with Mobile Devices & Services*. ACM, 121–124. DOI: https://doi.org/10.1145/2628363.2628404

J. M. Smyth and K. E. Heron. 2016. Is providing mobile interventions "just-in-time" helpful? An experimental proof of concept study of just-in-time intervention for stress management. In *Proceedings of the Wireless Health*. IEEE, 1–7. DOI: https://doi.org/10.1109/WH.2016.7764561

G. Spanakis, G. Weiss, B. Boh, and A. Roefs. 2016. Network analysis of ecological momentary assessment data for monitoring and understanding eating behavior, In *Proceedings of the Internationsl Conference on Smart Health (ICSH 2015)*. (Phoenix, AZ, Nov. 17–18, 2015). Revised Selected Papers, X. Zheng, D. D. Zeng, H. Chen, and S. J. Leischow (Eds.). Springer, Cham.

A. Stone, R. Kessler, and J. Haythomthwatte. 1991. Measuring daily events and experiences: Decisions for the researcher. *J. Pers.* 59, 3, 575–607. DOI: https://doi.org/10.1111/j.1467-6494.1991.tb00260.x

A. A. Stone and S. Shiffman. 2002. Capturing momentary, self-report data: A proposal for reporting guidelines. *Ann.Behav. Med.* 24, 3, 236–243. DOI: https://doi.org/10.1207/S15324796ABM2403_09

C. B. B. Taylor, L. Fried, and J. Kenardy. 1990. The use of a real-time computer diary for data acquisition and processing. *Behav. Res. Therapy* 28, 1, 93–97. DOI: https://doi.org/10.1016/0005-7967(90)90061-m

N. Tejani, T. R. Dresselhaus, and M. B. Weinger. 2010. Development of a hand-held computer platform for real-time behavioral assessment of physicians and nurses. *J. Biomed. Informat.* 43, 1, 75–80. DOI: https://doi.org/10.1016/j.jbi.2009.08.011

G. H. ter Hofte. 2007. Xensible interruptions from your mobile phone. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 178–181. DOI: https://doi.org/10.1145/1377999.1378003

S. Teso, J. Staiano, B. Lepri, A. Passerini, and F. Pianesi. 2013. Ego-centric graphlets for personality and affective states recognition. In *Proceedings of the International Conference on Social Computing*. 874–877. DOI: https://doi.org/10.1109/SocialCom.2013.132

K. Tollmar and C. Huang. 2015. Boosting mobile experience sampling with social media. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 525–530. DOI: https://doi.org/10.1145/2785830.2785894

P. Totterdell and S. Folkard. 1992. In situ repeated measures of affect and cognitive performance facilitated by use of a hand-held computer. *Behav. Res. Meth., Instrum., Comput.* 24, 4, 545–553. DOI: https://doi.org/10.3758/BF03203603

C. C. Tsai, G. Lee, F. Raab, G. J. Norman, T. Sohn, W. G. Griswold, and K. Patrick. 2006. Usability and feasibility of PmEB: A mobile phone application for monitoring real time caloric balance. In *Proceedings of the Pervasive Health Conference and Workshops*. 1–10. DOI: https://doi.org/10.1109/PCTHEALTH.2006.361659

H. Väätäjä and V. Roto. 2010. Mobile questionnaires for user experience evaluation. In *Proceedings of the Conference on Human Factors in Computing Systems (Extended Abstracts)*. ACM, 3361–3366. DOI: https://doi.org/10.1145/1753846.1753985

N. van Berkel, J. Goncalves, S. Hosio, and V. Kostakos. 2017. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*. 1, 3, 107:1–107:21. DOI: https://doi.org/10.1145/3130972

N. van Berkel, S. Hosio, T. Durkee, V. Carli, D. Wasserman, and V. Kostakos. 2016a. Providing patient context to mental health professionals using mobile applications. In *Proceedings of the CHI workshop on Computing and Mental Health*. 1–4. DOI: https://doi.org/10.13140/RG.2.1.3793.1923

N. van Berkel, C. Luo, T. Anagnostopoulos, D. Ferreira, J. Goncalves, S. Hosio, and V. Kostakos. 2016b. A systematic assessment of smartphone usage gaps. In *Proceedings of the Conference on Human Factors in Computing Systems*. 4711–4721. DOI: https://doi.org/10.1145/2858036.2858348

N. van Berkel, C. Luo, D. Ferreira, J. Goncalves, and V. Kostakos. 2015. The curse of quantified-self: An endless quest for answers. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (Adjunct)*. 973–978. DOI: https://doi.org/10.1145/2800835.2800946

K. Van den Broucke, D. Ferreira, J. Goncalves, V. Kostakos, and K. De Moor. 2014. Mobile cloud storage: A contextual experience. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. 101–110. DOI: https://doi.org/10.1145/2628363.2628386

S. Vhaduri and C. Poellabauer. 2016. Human factors in the design of longitudinal smartphone-based wellness surveys. In *Proceedings of the International Conference on Healthcare Informatics*. IEEE, 156–167. DOI: https://doi.org/10.1109/ICHI.2016.24

C. R. Walker. 1956. *The Foreman on the Assembly Line*. Harvard University Press (1956).

E. I. Walsh and J. K. Brinker. 2016. Should participants be given a mobile phone, or use their own? effects of novelty vs utility. *Telemat. Inform.* 33, 1, 25–33. DOI: https://doi.org/10.1016/j.tele.2015.06.006

Z. Wang, J. M. Tchernev, and T. Solloway. 2012. A dynamic longitudinal examination of social media use, needs, and gratifications among college students. *Comput. Human Behav.* 28, 5, 1829–1839. DOI: https://doi.org/10.1016/j.chb.2012.05.001

D. Weber, A. Voit, P. Kratzer, and N. Henze. 2016. In-situ investigation of notifications in multi-device environments. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1259–1264. DOI: https://doi.org/10.1145/2971648.2971732

J. Weppner, P. Lukowicz, S. Serino, P. Cipresso, A. Gaggioli, and G. Riva. 2013. Smartphone based experience sampling of stress-related events. In *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. 464–467. DOI: https://doi.org/10.4108/icst.pervasivehealth.2013.252358

J. Westerink, M. Ouwerkerk, G. J. d. Vries, S. d. Waele, J. v. d. Eerenbeemd, and M. V. Boven. 2009. Emotion measurement platform for daily life situations. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–6. DOI: https://doi.org/10.1109/ACII.2009.5349574

L. Wheeler and H. T. Reis. 1991. Self-recording of everyday life events: Origins, types, and uses. *J. Pers.* 59, 3, 339–354. DOI: https://doi.org/10.1111/j.1467-6494.1991.tb00252.x

M. L. Wilson, D. Craggs, S. Robinson, M. Jones, and K. Brimble. 2012. Pico-ing into the future of mobile projection and contexts. *Pers. Ubiq. Comput.* 16, 1, 39–52. DOI: https://doi.org/10.1007/s00779-011-0376-2

R. Xiao, G. Laput, and C. Harrison. 2014. Expanding the input expressivity of smartwatches with mechanical pan, twist, tilt and click. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 193–196. DOI: https://doi.org/10.1145/2556288.2557017

D. Xu, L. Qian, Y. Wang, M. Wang, C. Shen, T. Zhang, and J. Zhang. 2015. Understanding the dynamic relationships among interpersonal personality characteristics, loneliness, and smart-phone use: evidence from experience sampling. In *Proceedings of the International Conference on Computer Science and Mechanical Automation*. 19–24. DOI: https://doi.org/10.1109/CSMA.2015.11

Y. Yang, G. D. Clark, J. Lindqvist, and A. Oulasvirta. 2016. Free-form gesture authentication in the wild. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 3722–3735. DOI: https://doi.org/10.1145/2858036.2858270

Z. Yue, E. Litt, C. J. Cai, J. Stern, K. Baxter, Z. Guan, N. Sharma, and G. Zhang. 2014. Photographing information needs: The role of photos in experience sampling method-style research. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 1545–1554. DOI: https://doi.org/10.1145/2556288.2557192

X. Zhang, L. R. Pina, and J. Fogarty. 2016. Examining unlock journaling with diaries and reminders for in situ self-report in health and wellness. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 5658–5664. DOI: https://doi.org/10.1145/2858036.2858360

Received September 2016; revised June 2017; accepted July 2017