Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being

LARRY CHAN, VEDANT DAS SWAIN, CHRISTINA KELLEY, Georgia Institute of Technology
KAYA DE BARBARO, University of Texas at Austin
GREGORY D. ABOWD, Georgia Institute of Technology
LAUREN WILCOX, Georgia Institute of Technology

Ecological Momentary Assessment (EMA) methods have emerged as an approach that enhances the ecological validity of data collected for the study of human behavior and experience. In particular, EMA methods are used to capture individuals' experiences (e.g., symptoms, affect, and behaviors) in real-world contexts and in near-real time. However, work investigating participants' experiences in EMA studies and in particular, how these experiences may influence the collected data, is limited. We conducted in-depth focus groups with 32 participants following an EMA study on mental well-being in college students. In doing so, we probed how the elicitation of high-quality, reflective responses is related to the design of EMA interactions. Through our study, we distilled three primary considerations for designing EMA interactions, based on observations of 1) response strategies to repeated questions, 2) the perceived burden of EMA prompts, and 3) challenges to the validity and robustness of EMA data. We present these considerations in the context of two microinteraction-based EMA approaches that we tested: lock-screen EMA and image-based question prompts. We conclude by characterizing design tensions in the presentation and delivery of EMA prompts, and outline directions for future work to address these tensions.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; *Ubiquitous and mobile computing design and evaluation methods*;

Additional Key Words and Phrases: Ecological Momentary Assessment, Experience Sampling Method, Emotion, User Interface Design, Focus Groups, Qualitative Study

ACM Reference format:
Larry Chan, Vedant Das Swain, Christina Kelley, Kaya de Barbaro, Gregory D. Abowd, and Lauren Wilcox. 2018. Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 3 (March 2018), 20 pages.
https://doi.org/10.1145/3191735

1 INTRODUCTION

Ecological momentary assessment (EMA) methods are used to capture someone's behaviors or experiences through periodic, in-situ, self-report. They call for participants to make their reports repeatedly, in a variety of contexts, each of which are natural for them and close in time to the experience being reported [29, 34]. While EMA paradigms have existed for many decades, improvements in computing portability and sensor technologies

This work was supported by a gift from AIG Science, the National Natural Science Foundation of China under Grant No.: 61273304_a, and the Young Scientists' Support Program.

2474-9567/2018/3-ART3 $15.00
https://doi.org/10.1145/3191735

have renewed interest in EMA data collection in the UbiComp community. In particular, a number of recent studies combine “passive" sensor data, collected from smartwatches and smartphones, with “active" EMA self-reports. These studies span a wide range of domains and goals, from understanding the proximal behaviors that contribute to failures to quit smoking [23] to studies examining the well-being and academic performance of students [38, 39].

As these high-density sociobehavioral data continue to emerge, a number of researchers in the HCI and UbiComp communities have begun to develop new tools to enhance traditional forms of EMA data collection on mobile and wearable devices. These interaction techniques explore how to exploit microinteractions on a smartphone [1, 35, 36, 41] or smartwatch [16, 31]. **We focus in this paper on how the design of specific microinteractions for prompting an EMA question and eliciting the response (i.e., through an unlock gesture on a smartphone) affect the collection of high-quality response data over time.** Past work has revealed that the unlock gesture is a prime opportunity to get users to respond quickly, and often, to EMA-like questions. What has *not* been explored is whether those answers are high-quality answers. In this paper, we investigate the importance of balancing the need to decrease user burden for an individual EMA, the strength of a microinteraction EMA approach, and the quality of the elicited response.

We explain how we adopted a mixed-methods approach in order to understand this balance between ease and quality of EMA response. We present findings from a multi-week study with 32 college students, who participated in a larger mobile-sensing study that included both passive (through the on-board smartphone sensors) and active data collection (through a microinteraction-based EMA platform). Participants in our study responded to an average of 10 EMA prompts per day over an average of 12 days. Question prompts included items assessing participants' mood, stress, and current activity. This paper does not present any results about student wellness or mental health; rather, its focus is on uncovering a design tension between the desire to get many self-assessments from an individual each day and the quality of each of those assessments. By studying experiences with EMA in this domain, we could explore this tension fully, since question prompts often require participants to reflect on their subjective internal experience in order to respond. By using a mixed-methods approach for the study, we derived insights from in-depth focus groups, supported by EMA log analyses, to shed light on how people experience answering EMA questions that sometimes require a period of reflection in order to answer accurately.

As such, we make the following contributions:

- We distill the strategies applied by study participants to respond to repeated EMA questions, and identify how these strategies challenge the validity and robustness of the resulting data, with special emphasis on microinteraction-based EMA for emotion.
- We identify attitudes toward use of microinteraction-based EMA technology in daily life, including preferences for sharing emotions when prompted, and perceived burden of EMA prompts.
- We characterize design tensions in the presentation and delivery of EMA prompts, emphasizing a conflict between the time required to report a current emotion and the EMA design objective of minimizing the time it takes to answer a question, and we outline directions for future work to address these tensions.

2 BACKGROUND AND RELATED WORK

2.1 EMA Methods

The family of methods that fall under “EMA” are diverse and have often gone by other names, such as “experience sampling method” and “diary study.” Our use of the term “EMA” encompasses these methods [34]. *Momentary* reporting of an experience helps to overcome problems with memory. This capability is important, because people can be disproportionately influenced by the "best", "worst", and most recent events when summarizing their experiences over a period [18] and they have trouble integrating and summarizing all relevant past information in a “global report”. As such, EMA is particularly well suited to the task of observing in-the-moment patterns of variability or stability in one's experience.

However, EMA methods can also introduce challenges to users and researchers. Prompts to report can arrive at inopportune moments or in inconvenient places for a user or study participant. The need to report repeatedly implies that participants will expend effort on a recurring basis, which can lead to *burden* [33, 34]. A frustrated user, reporting with insufficient effort, could introduce poor-quality data, or even leave a study altogether.

Even when the user interaction burden is reduced, EMA-collected data can be susceptible to other types of biases. Repeated use of a scale may cause its anchor points to shift if a participant begins to report current ratings only in relation to their previous ratings [33]. Contextual biases can be introduced if some contexts (e.g., activities) are sampled less frequently than others. This can occur when prompts, randomly-generated or on a schedule, are timed such that they do not coincide with a user's experience of those contexts (e.g., sampling smoking activity only on weekdays would fail to capture the behavior of someone who smokes at bars on weekends).

Finally, it is possible for behaviors and attitudes to be altered by the experience of mere measurement, and this risk may be exacerbated when they are measured repeatedly over extended periods of time. For example, evidence from consumer marketing research suggests that the measurement of intentions to make a purchase can increase the likelihood of various purchase patterns [8, 28].

Still, EMA provides many advantages over conventional surveys and interview methods, especially when there is a need to overcome recall bias [34]. Stone et al. [22] situate EMA within the larger context of survey design methodology, pointing out the need for the assessment of data quality in future work. Of particular relevance to our study, *satisficing theory* [20] provides a framework for understanding how a range of human strategies can influence the validity of survey data. In particular, Krosnick distinguishes conditions that encourage thoughtful and optimal response strategies from those in which participants reduce their effort and give a response that they deem satisfactory (satisficing). Little work has been done to apply this framework to EMA methodology, whose implementation introduces conditions that have not been thoroughly addressed in terms of satisficing—a focus of our work.

Krosnick [20] identified three factors that predict whether a person will satisfice or respond optimally when responding to questions: *task difficulty*, *respondent ability*, and *respondent motivation*. Krosnick found that the greater the task difficulty and the lower the respondent's ability and motivation, the more likely satisficing will take place. For example, when an interviewer asks questions too rapidly, the task becomes more difficult and encourages satisficing. These factors have been studied in terms of traditional surveys, but EMA methodology–and the interaction design facilitating EMA–present new contexts and circumstances that bear directly on factors such as task difficulty. In this paper, we elucidate these concerns as they relate to EMA, by exploring **how the design of EMA prompts can affect response accuracy, respondent motivation, *and* data quality.** We present accounts given by participants of what it was like to take part in a multi-week EMA study and to be asked questions in a wide array of circumstances that include walking, driving, talking to friends, and studying.

2.2 Categories of EMA

Wheeler et al. [40] describe three categories of EMA methods: *interval-contingent*, *signal-contingent*, and *event-contingent*. In interval-contingent designs, participants make reports after a pre-determined time interval. For example, reporting on the hour how many glasses of water they consumed for that hour. Signal-contingent designs call for the participant to respond to prompts from the researcher. Researchers determine when participants will be prompted, and participants make a report once they receive the prompt. In event-contingent designs, participants report data when a relevant event is taking place. This taxonomy is useful for distinguishing many EMA projects, but the categories are not orthogonal, so some projects that combine features may not be fully described with these terms. Froelich et al. [9] introduced the term *context-triggered* to distinguish a method that makes it unnecessary for participants to watch out for the events that call for reporting. A context-aware system could prompt (signal) the participant whenever it senses the appropriate context (event) [17]. As an alternative to the three-type taxonomy, researchers can describe: the conditions for *when* to collect a sample, *whether to prompt* the participant, whether the *participant can initiate* a report (instead of or as well as being prompted), and, if there are prompts, the *requisite conditions* for sending a prompt. This would capture studies that have conditional prompts and sample only during certain times *and* in a specific context (e.g., sample any time after 9:00 PM whenever the participant is in a hotel, but send a prompt only if the participant is awake). In the following sections, we describe EMA considerations on mobile devices, in the emotional domain, and in the domain of sampling student well-being—the primary areas of work that informed our study.

2.3 Reducing Burden in EMA Instruments for Emotion

Researchers have been working to reduce the specific burden of responding to EMA prompts. In particular, validated measures of mood and stress can be onerous because of the time and effort required to respond. For example, the Positive and Negative Affect Schedule (PANAS), which assesses a person's positive affect (PANAS PA subscore) and negative affect (PANAS NA subscore), requires answering 20 questions. In order to assess positive affect with a single question, Pollak et al. [30] developed the Photographic Affect Meter (PAM), an image-based prompt which only requires answering one question. PAM presents a 4 x 4 grid of images and asks respondents to choose the image that best captures their current mood. The images are ordered so that the highest arousal images are on top and the lowest arousal images are on the bottom, and so that images on the right represent the most positive valence and those on the left the most negative valence. Participants can shuffle the images to get a different combination of 16. Choosing an image yields a score between 1 and 16, which was shown to have good convergent validity with PANAS PA.

2.4 Reducing Burden in EMA Platforms

While EMA platforms are now commonly implemented on smart phones, research has also investigated the use of smart watches and other wearable devices [25, 35, 37]. For mobile devices, the burden of an interaction can be defined by two stages [3]. *Access time* refers to the time it takes to retrieve a device and prepare it for the intended use (e.g., retrieve phone, unlock it, and navigate to an application). *Usage time* refers to the amount of time spent carrying out the intended use.

To circumvent the access time related to phone use, Intille et al. [16] implemented a watch-based interface and compared it to a phone-based interface. Even though the smaller, watch-based interface interrupted people eight times more often, it outperformed its phone-based counterpart in terms of compliance, completion rate, and first-prompt response rate. Their strategy interrupted people more frequently, but made the interruptions brief. Intille et al. estimated that the access time and usage time added up to between three and four seconds, which meets the definition of a *microinteraction* [3].

Though devices worn on the head or wrist may reduce the time required to interact with EMA systems, it is a challenge to fit response options on worn displays. EMAs delivered for watch displays often translate question responses into separate portions that are asked 2-30 minutes apart. Phones may be preferable for ensuring that users report information accurately, as well as for ease of use, and potential future use [14].

Eliminating the burden of unlocking and navigating to a relevant application, Zhang et al. [41] developed and assessed LogIn, a prototype journal application with the EMA interface on the lock screen of mobile phones. When LogIn's lock screen interface replaced the effort of unlocking, participants responded to notifications more frequently, and the interaction was perceived as less intrusive.

A commercial application, called Quedget [1] (Figures 1 and 2) presents EMA questions on the lock screen and reduces “friction” in a way similar to LogIn and Slide-to-X [35]. However, Quedget and LogIn differ in how they prompt an assessment. LogIn presents the questionnaire interface every time the phone is activated. Participants can report at will instead of making reports contingent on intervals, events, or signals. To prompt assessment using LogIn, Zhang et al. used a mobile-phone-platform notification. Thus the user experiences LogIn as a personal journal tool, coupled with reminder prompts. In contrast, participants using Quedget can only report in response to a prompt that has been scheduled to appear on the lock screen. When no question is scheduled, the lock screen is free of any intrusive interface (unlike LogIn where the interface permanently replaces the native interface) In addition, Quedget imposes a context condition on prompts: they will only appear when the participant activates the phone (whereas prompts can arrive at any time in the case of LogIn). We used Quedget to schedule the delivery of EMA prompts while taking advantage of lock screen interaction.

2.5 Qualitative Work in EMA

Consolvo et al.[5] conducted an assessment of *My experience (Me)*, a platform that enabled researchers to use a participant's location to decide when to prompt them for a report. Their qualitative feedback showed that their participants did not want to receive notifications at inopportune moments, such as while eating or in conversation. We focus more on EMA prompts about mood and emotion [12, 15, 24, 27] and specifically investigate participants' lived experiences with EMA.

In a field study, Morris et al. [27] deployed a mobile phone application that both collected mood information through EMA and delivered therapeutic exercises. The researchers conducted interviews with their participants and gained qualitative insights into the EMA experience. The qualitative feedback revealed difficulties disentangling the influence of interventions (such as therapy or visualization of past emotions) on the experience of responding to repeated questions. Their study combined the experience of reporting mood information with the experience of mobile therapy. Other qualitative work in EMA of well-being incorporates the experience of reviewing recorded data, so participants'feedback does not necessarily describe the experience of answering questions. **We focus in this paper on how students experienced EMA for well-being without the effects of feedback or specific interventions, and we concentrate on the experience of answering EMA questions.**

Capturing the subjective aspects of emotion—so that they can be connected with environmental and physiological context—is one of the major applications of EMA in UbiComp [4]. As EMA is increasingly applied to help draw correlations between behaviors and well-being [12, 15, 24, 27, 38], it is now more important than ever to understand—and begin to address—tensions inherent in the interaction design of EMA technologies to capture subjective experiences.

2.6 EMA in the Study of College Student Well-being

As a type of questionnaire, EMAs are applicable to the study of numerous domains in health and social science. A number of recent efforts have supplemented traditional EMA studies with smartphone platforms that passively collect mobile phone sensor data (e.g. motion, GPS coordinates and audio snippets [11]). The StudentLife study at Dartmouth College [38] pioneered the combined use of EMA with passive mobile sensing data to infer markers of various student behaviors (e.g., attending a class or a party) in order to draw relationships between behaviors, academic performance, and mental well-being, empirically. Their findings include behavioral and contextual predictors of student depression and stress (e.g., strong negative correlation between stress and evening conversations).

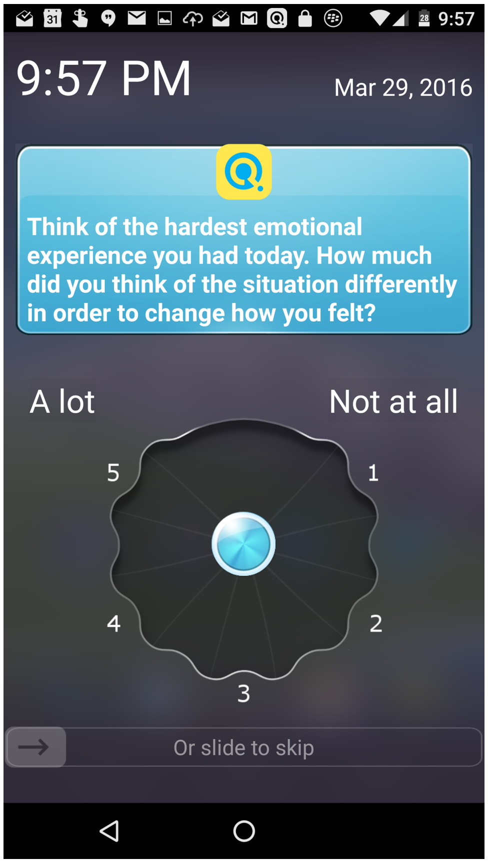
Yet, prior work on the design and study of EMA has approached the design of EMA technology from the researcher's perspective, with the goal to collect data that grows in volume, variety, and quality. Less work has been done probing the experience of answering questions and sustaining ongoing participation in EMA studies—an important focus of our work. By focusing on this dimension to better understand underlying EMA response behavior, we gain insights that can help improve response validity.

3 STUDY

We conducted focus groups at the conclusion of a larger mobile-sensing study. As in the StudentLife project, we collected passive and active data through students'Android phones. Passive data were collected through the sensors on board the phone, and active data (self-report) were collected through the Quedget's context-and-signal-contingent EMA platform (Figures 1 and 2).

3.0.1 *Implementation of Self-Report on the Quedget Platform.*

As mentioned above, Quedget is an EMA service that employs the lock screen to minimize the interruption burden in a way similar to LogIn [41], Slide-to-X [35], and Twitch Crowdsourcing [37].



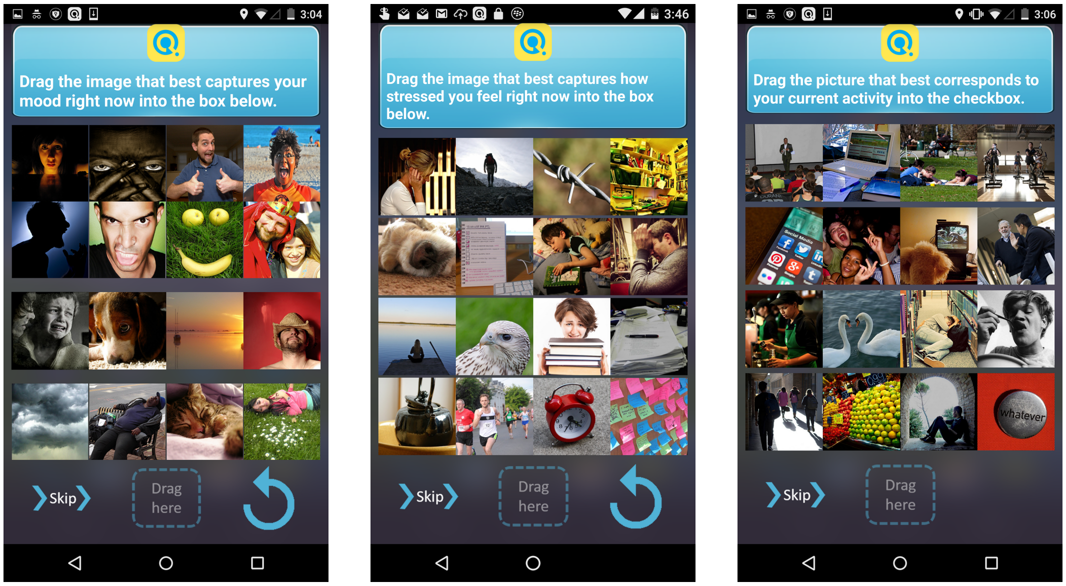
Think of the hardest emotional experience you had today. How much did you think of the situation differently in order to change how you felt?

Fig. 1. The lock-screen interface with a Likert-based question prompt, shown when prompted a participant to answer one of the emotion regulation (ER) questions.

It provides a full, back-end Web application for defining EMA questions and analyzing the responses. Quedget is available only on Android (this platform allows developers to build interactions that appear between the activation of the phone and before the lock screen appears). Quedget enables multiple question types and we used two types in our study: Likert-scale rating questions (Figure 1) and image-based multiple-choice questions (Figure 2).

We used the Quedget platform to ask a variety of questions, including Photographic Affect Meter (PAM), and Likert questions derived from established measures, such as the State Self Esteem Scale (SSES) [13] and Patient Health Questionnaire (PHQ-4) [19]. Only one item was asked at a time. Quedget only enables assessments as responses to a *scheduled signal*. Researchers define windows of time during which a question should be presented to the participant. Within that window, Quedget calculates a random time to trigger a prompt. Once a prompt has been triggered, it will appear on the phone the next time a lock screen would appear. The prompt itself contains both the questionnaire item and all response choices.

Thus, the presentation of the EMA interface is itself the signal (in the taxonomy used by Wheeler et al.[40]). Unlike pure notification-based, signal-contingent systems, the timing of the signal cannot be precisely defined, because the participants enact a triggered signal whenever they next activate their phones. Prompts on the Quedget platform thus combine signal-contingent and event-contingent EMA design.



Drag the image that best captures your mood right now into the box below.

Drag the image that best captures how stressed you feel right now into the box below.

Drag the picture that best corresponds to your current activity into the checkbox.

(a) PAM (b) MSM (c) PAR

Fig. 2. The lock screen of the participant when prompted by an image-based question. Photographic Affect Meter (PAM), Mobile Stress Meter (MSM) and Photographic Activity Recognition (PAR) are the three different image-grid questions, each of which uses a set of 16 images to capture a response from the participant.

When the prompt is displayed, participants can use an option in the Quedget interface to skip the question, or they can answer the question. Answering the question causes the Quedget interface to disappear, exposing the lock screen and enabling normal use of the phone. Skipping the question causes it to appear again the next time the phone is activated. If participants skip a question and do not activate their phones before the window for that prompt is over, that prompt will not appear until it has been triggered again.

3.0.2 Adapting Likert Scales for Emotion Regulation. To study Emotion Regulation (ER)[1] we asked participants how much they employed each of seven regulation strategies, including acceptance of a feeling and cognitive reappraisal (thinking differently about something to change how one feels about it) [2]. We adapted the questions to use a five-point scale (see Figure 1).

We scheduled the ER questions to be presented at the end of the day and modified the wording so that participants could choose any moment since the day began: "Think of the hardest emotional experience you had today." This retrospective process carries a greater risk of recall bias than a momentary formulation of the same question, but we made this choice so that there would be a better chance that the "hardest emotional experience" chosen by the participant would indeed correspond to the day's greatest emotional challenge.

3.0.3 Image-based Multiple-Choice Questions. We included three image-based questions: we used the Photographic Affect Meter (PAM) to measure affect; we adopted the Mobile Stress Meter (MSM); and we developed our own question for capturing current activity, Photographic Activity Recorder (PAR), which we describe in detail below.

[1] "Emotion regulation refers to the way individuals influence which emotions they experience, as well as how and when they experience and express them." [10]

MSM resembles PAM in a number of ways. Images are arranged in a 4 x 4 grid, and participants respond by choosing the image that best captures their current state—in this case, *stress*. Images are arranged so that the one corresponding to the lowest stress score (1) is at the bottom left position on the grid. Each additional increment of stress is represented either by the image directly above or the image at the bottom of the column to the right. The image at the top right of the grid corresponds to the highest stress score in the scale (16). As with PAM, there are a total of 48 different images, three associated with each position on the grid. The participant can shuffle the images to get a different combination of 16 images. Using PAR, participants choose one image to indicate an activity that they are currently engaged in. Unlike PAM and MSM, these images are arranged in no particular order. Each image was labeled to indicate the activity depicted ("Class," "Studying/Homework," "Socializing," "Exercise," "Social Media," "Party," "TV/Leisure," "Talking with a professor," "Paid Job," "Date/ Romance," "Sleeping," "Eating," "Walk/Commute," "Errands," "Contemplation," "Other"). The label for each image appeared when participants touched the image corresponding to that label. Releasing the image before dragging it to the target square enabled the participant to explore other labels by touching other images.

3.1 Likert-Style Questions

3.1.1 Prompt Triggering. Quedget requires a researcher to define windows of time during which a question should be presented to the participant. Within that window, Quedget calculates a random time to trigger the prompt. To cover each day, we defined four windows of time. We scheduled PAR and PAM to be triggered every day, within each of these four windows (Table 1 shows a sample schedule from one of the weeks). We scheduled one of the four SSE questions to be asked within each window. We scheduled MSM to be triggered once per day. We scheduled one of the four PHQ-4 questions to be asked within each window on Saturday, with no question repeated that day. We scheduled four questions about burden, partying, studying, and productivity to be asked on Sunday, each in separate windows (these are tagged as WR, for "Weekly Review"). Finally, we scheduled ER questions such that each of the seven is asked twice between Monday and Sunday with no repetition of the same on the same day.

3.1.2 Interaction Pattern. Once a prompt has been triggered, it will appear on the phone the next time a lock screen would appear. The prompt itself contains both the questionnaire item and all response choices. Each prompt presents a single question. Because responding or skipping required only one gesture, this means that participants could receive a prompt, respond to it (or skip it), and be ready for normal phone use after a single drag gesture.

Interfaces that become responsive on a phone before the owner is aware of it can facilitate inadvertent interactions. An owner may touch the screen while the phone is pocketed and unknowingly register a tap or swipe. To prevent participants from doing this when a prompt is displayed, Quedget requires participants to make specific drag-and-drop gestures that are less likely to qualify inadvertently (as does LogIn). In the example of image-based questions, participants register an answer by choosing an image and dragging it from the grid to a target square below the grid.

3.2 EMA Study

We conducted the focus group research as part of a larger, IRB-approved, multi-modal sensing project. In the following sections we describe the protocol for the entire project and give an overview of participant participation.

3.2.1 Study Protocol. Enrollment for the study began on 3/30/2016 and continued until 4/20/2016. Participants completed an online battery of questionnaires during the enrollment session, another online battery at the end of the session, and throughout the study they answered EMA questions. We also used AWARE [7], a mobile instrumentation research framework, to capture "passive" data (i.e., accelerometry, GPS, ambient noise levels, call and message meta-data, screen state, application history, and physical activity provided through GoogleDetectedActivity API). (We do not focus on analysis of the passive sensor data in this paper.)

Undergraduate and graduate students at Georgia Institute of Technology who used Android phones as their primary phone were eligible to participate in the study. In enrollment sessions, we described the aim of the study, all sensor data collected, the procedure for responding to various kinds of EMA questions, and the compensation scheme.

Table 1. A sample weekly question prompt schedule used in the first week of our mobile sensing study. The time-windows on the left column are subject to shift, while the overall daily composition of questions remains constant.

	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
9:00-12:30	PAR	PAR	PAR	PAR	PAR	PAR	PAR
	MSM				MSM		
	PAM	PAM	PAM	PAM	PAM	PAM	PAM
	SSE	SSE	SSE	SSE	SSE	SSE	SSE
							PHQ-4
						WR	
12:30-16:00	PAR	PAR	PAR	PAR	PAR	PAR	PAR
		MSM					MSM
	PAM	PAM	PAM	PAM	PAM	PAM	PAM
	SSE	SSE	SSE	SSE	SSE	SSE	SSE
							PHQ-4
						WR	
16:00-19:30	PAR	PAR	PAR	PAR	PAR	PAR	PAR
			MSM			MSM	
	PAM	PAM	PAM	PAM	PAM	PAM	PAM
	SSE	SSE	SSE	SSE	SSE	SSE	SSE
							PHQ-4
						WR	
19:30-23:00	PAR	PAR	PAR	PAR	PAR	PAR	PAR
				MSM			
	PAM	PAM	PAM	PAM	PAM	PAM	PAM
	SSE	SSE	SSE	SSE	SSE	SSE	SSE
							PHQ-4
						WR	
18:00-22:00	ER	ER	ER	ER	ER	ER	ER
	ER	ER	ER	ER	ER	ER	ER

Participants could earn a maximum of $80 for participating in the EMA study (reaching the maximum as the number of responses they provided approached the number of questions asked).

After providing written consent, participants installed Quedget and AWARE on their phones, answered sample EMA questions in the presence of researchers. Participants could ask questions about the user interface and receive more training if needed during the sessions.

In this study, we separate the experience of responding to prompts from that of receiving explicit interventions, including visual feedback of the recorded data. Our EMA data collection platform minimizes interruption and employs a combination of traditional and image-based EMA instruments.

On a standard weekday schedule, participants could receive up to 15 questions per day. On Saturday and Sunday, they could receive up to 19 per day. Table 1 shows the schedule for one study week. The mean number of questions answered by each participant per day varied considerably (see Table 2). At the end of the study, we invited participants to schedule a focus group session.

3.2.2 Focus Groups. Of those who participated in the larger study, 32 graduate and undergraduate students (44% Female) participated in the focus groups. The moderator used a topic guide in all focus groups (described below). The topic guide addressed four categories: *respondent burden*, *self-observation*, *privacy*, and *social considerations*.

To examine respondent burden, we invited participants to talk about the number of times they were asked EMA questions, how long it took to answer, and what they thought of the questions. To examine self-observation, we asked whether certain experiences caused participants to think about or remember information. For privacy and social considerations, we asked what students were comfortable disclosing through EMA. Each focus group—eight in all—lasted approximately one hour and included, on average, four participants. The sessions were recorded using both audio and visual equipment and transcribed verbatim.

3.3 Data Analysis

Focus group transcripts were segmented by speaker turn. Once segmented, two authors analyzed speaker data, independently coding the responses using constant comparison to iteratively arrive at themes—in an inductive, bottom-up fashion—until consensus was reached. Analysis conformed to conventional approaches associated with conducting a basic interpretive qualitative study [26]. Some a priori codes were generated by the topic guide that was constructed before the focus groups were conducted.

4 FINDINGS AND DISCUSSION

We first provide descriptive quantitative data to provide the necessary backdrop for interpreting the qualitative results that follow. We then describe salient findings from our qualitative data analysis and discuss their implications for future research in this space.

Table 2. Summary of participation in our mobile sensing study with college students. ("Qu." = Quartile.)

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Days in Study per Participant	1.00	9.00	10.00	12.09	15.00	29.00
Responses per Participant	5.00	63.00	112.50	114.40	149.80	403.00
Daily Responses per Participant	0.26	6.04	9.91	10.35	13.34	33.00

4.1 Overview of Response Activity

Participants regularly answered fewer than the 15 or 19 questions possible for the day, answering an average of 10 EMA prompts per day. Occasionally, participants received extra questions on days when new participants were being enrolled. These questions were meant to familiarize new participants with the question format on the day of enrollment, but all participants already enrolled received the extra questions as well. However, in order to receive all questions, a participant would need to activate her phone frequently enough within each time window. For example, she would need to activate her phone at least seven times between 7:30 PM and 11:00 PM to get all of the questions scheduled to appear during that window of time. Most of the responses took place between 9am and 10pm. Table 2 shows the daily and total numbers of responses obtained per participant.

4.2 Time Required to Thoughtfully Answer Prompts

We asked one image-based question (PAM) to measure mood and a different one (MSM) to measure stress. In the focus groups, we learned that it was **common for participants to first take time to identify how they were feeling in order to then map their emotional experience onto the images**.

> *"[I]'m not sure about the girl just lying down there[...] I was just feeling, 'Meh.' [...] Not happy, not sad, not stressful."* -P12

When describing how they reported stress using MSM, participants indicated that they used image features to express aspects of stress, but first reflected on how their stress manifested in different ways . One participant

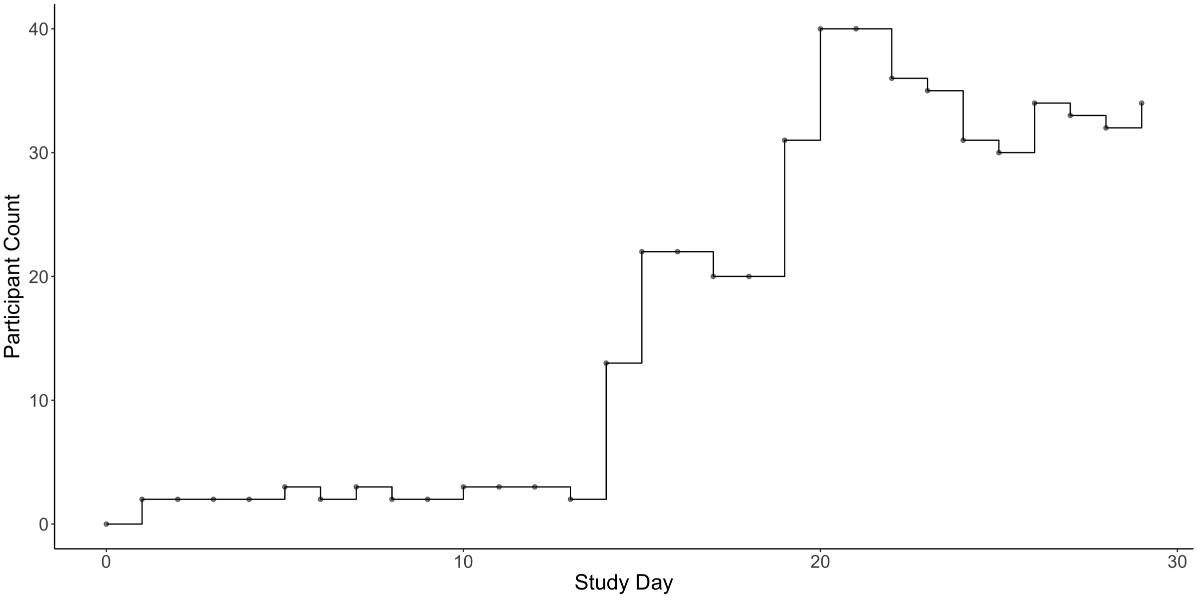


Fig. 3. The study had a rolling admission of participants and thus had a differing number of participants during different days of the study. The Y axis depicts the total number of active participants—those who contribute EMA data—on a given day of the study.

mentioned that, *"...study-stress is different from I'm-fighting-with-my-best-friend stress ... "*). When participants responding to PAM prompts found the process of identifying their emotions to be straight-forward, so as not to require much momentary reflection, they still spent time viewing images on which to map their emotions.

> *"I feel like [PAM] was the more straight-forward [...] 'I'm feeling like this,' and just slide [...] if I saw a picture and I'm like, 'Eh, I don't really feel like that,' I [would] hit the refresh button a couple of times."* -P17

Some participants said that their awareness of their emotions changed while participating, and they frequently noticed this during EMA prompts. Taking time to cultivate and reflect on this awareness was integral to the process of answering the prompt, as P23 described.

> *"[In] the beginning I was just answering to answer them. Over time I started to actually think about my answers and actually notice on a spectrum of one to five, like it could be different: me putting one or two ... some days I really think about some things; when it came up I was actually like, 'do I feel this way, like, this past hour?'"* -P23

Noting that the EMA prompt itself became part of a reflective intervention has important implications for the design of EMA technology. In particular, **where EMA prompts have previously been designed to elicit information as quickly as possible and with as little effort as possible, our findings suggest that the designers of EMA technology should take into account the importance of purposeful momentary reflection** as part of the EMA prompt–response process.

4.3 Perceived Burden and Time

When we asked the participants how they felt about the number of prompts they received (up to 19 in one day), not one said that there were too many. Participants said that they would continue responding to the current rate or to an even higher rate of prompts. For example, some participants suggested that sampling mood and stress even more frequently than we did would be beneficial.

> *"Yeah, the how do you feel right now and what are you doing right now, that could be asked a bit more frequently."* -P18
>
> *"And stress level too."* -P19
>
> *"Yeah, and how stressed you are."* -P18

As mentioned above, Intille et al. kept interruptions short to increase the number of prompts that would be tolerated [16]. **We intentionally designed the study to keep interruptions short as well**. In addition, our context-and-signal contingent design also avoided interrupting participants at a specific level of *activity*. While participants were interrupted at what is called the *action* level in activity theory (e.g., checking the weather on the phone or looking up the definition of a word), we avoided interrupting them at the *operations* level (well-defined, routine actions like tapping "send" on an email). We attribute their tolerance for frequent prompts to brevity of interactions and uninterrupted operations. Future research could reveal how each feature contributed to the tolerance participants showed for frequent prompts.

5 DESIGN TENSIONS

In the following sections, we identify and describe conflicts emerging from our focus group findings, illustrating tensions between competing design objectives. After introducing these tensions, we discuss how researchers and user interface designers can consider them and weigh trade-offs in designing EMA for emotion.

5.1 Response Rate Versus Quality of Response

In many cases, the intended data are not collected when respondents decline to answer a question, or choose responses such as "don't know" or "no opinion." To avoid sparsity in the self-report data collected, researchers may try to *increase the likelihood that respondents will provide an answer to every question asked.* One way to do that is to de-emphasize —or make less prominent in the user interface—the option to skip a question. In our study, the option to skip was indeed de-emphasized: in order to skip a question presented by our system on the lock screen, respondents needed to notice icons at the bottom left of the interface. In focus groups, we learned that task overload situations motivated skipping, but the interaction required for skipping a question became a challenging task in itself:

> *"I was driving ... then the question popped up, but then it was really awkward trying not to get into an accident, but skip the question ... "* -P16

In fact, participants in *all* focus groups reported that being occupied by a different task affected their motivation to answer prompts thoughtfully.

> *"... if I'm late for class or something, I'm checking the phone to make sure I'm not really late. . . I'm going to answer that question faster and not give as much thought as if I was [in] my dorm or something."* -P21

One important explanation participants gave for *not* skipping in these situations is related to incentives for answering EMA prompts. Compensating participants in proportion to the number of responses can maximize the response rate [6, 12], but at a cost. Our **attempts to maximize the response rate had elicited invalid responses in some contexts**. If there is no way of distinguishing such responses, the data are compromised in general.

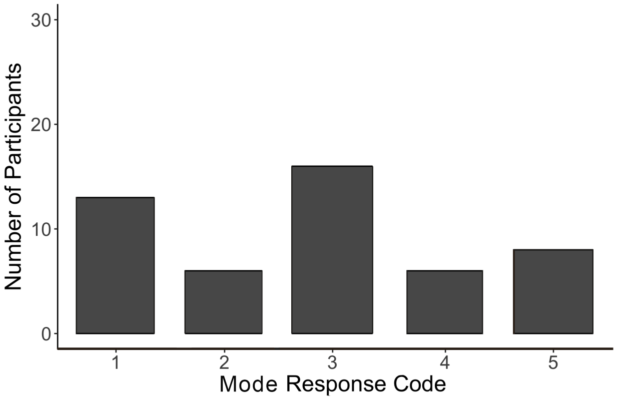


Fig. 4. A majority of our sample of participants chose "3" more often than any other response when answering Emotion Regulation (ER) questions.

Even when sufficiently motivated to give valid answers, participants improvised responses to questions that they felt unable to answer at a particular time or in a particular setting. For example, without a way to indicate that an EMA question did not apply, our participants tried to *approximate* the effect of giving a *neutral* answer or taking a "not applicable" option. When we asked what *response values* they used to do this, respondents in a majority of focus groups reported using "3"–the midpoint of a five-point scale:

> *"Three."* -P26
>
> *"It was pretty much three, yeah."* -P18
>
> *"Three–neutral."* -P26

One participant called this "midpoint" (3) a "neutral" answer. In the minds of participants, choosing a neutral answer could do the *least harm* to the study and does not bias the data in either direction. Such good intentions may or may not hold true for a study, and this case is an example of how such data can be harmed.

For example, the scale for our ER question is unipolar, with 1 representing "not at all." There is no neutral answer, and a value of "3" elevates the frequency with which an ER strategy was employed. Figure 4 shows in a histogram that the mode is "3". We are unable to interpret this figure with confidence, because some responses are "3" where we would expect "1". Such confounds could be avoided through instruction during enrollment, or through the addition of user interface elements to provide a layer of reference information that could be progressively revealed.

In summary, we found that we cannot expect that providing question prompts through microinteractions will, by their nature, elicit a natural response that is representative of a participant's mental or emotional state. An unexpected side-effect of such an interaction is the *ease with which a participant can answer arbitrarily*, so as to get on with their primary task, i.e. unlocking the phone. We recommend the following design considerations to minimize such an effect:

- Consider providing a skipped/neutral/"don't know" response option and accepting it as a form of active response. In addition, we recommend facilitating such options by designing them with equal or greater accessibility (or ease) relative to the content-laden response options. We expect such designs to 1) reduce invalid responses and 2) shed light on the comprehensibility of questions as well as the appropriateness of questions in given contexts.
- Improving the contextual prompting mechanism of the EMA, such that it only interrupts users during appropriate opportunity intervals. Therefore, on sensing the presence of a situation with a high cognitive load primary task (for e.g. driving) the prompt is not displayed and is deferred.

5.2 Response Speed Versus Response Quality

We chose the lock screen design because it enabled our study to avoid interrupting the primary phone tasks. Having removed that form of burden, we further sought to mitigate other sources of burden. First, we made choices that minimized the time required to respond to a question, reasoning that this would facilitate resuming activities in progress as quickly as possible. Second, we chose simple interface designs over complex ones, reasoning that simplicity would reduce cognitive burden, the learning required, and, ultimately, the time expended on responding. However, the choices we made in pursuit of these design goals conflicted with other study objectives.

Studies that rely on participants to reflect before giving an answer may suffer when responses are given too quickly. While the response times in our study were not as short as those reported by Intille et al. [16], the median response time for PAM was not much more than the four seconds that define microinteractions (Table 3). Answers to questions about observable matters of fact ("Are you alone or with others?") may be compatible with very quick responses, but the same is not true for questions that ask participants to make subtle judgments of mood and emotion.

Table 3. We grouped all of our EMA questions into question families and compared the response times across the families.

	ER	PAM	PAR	PHQ-4	MSM	SSE	WR
1st Qu. (seconds)	3.25	3.62	3.49	3.59	4.44	2.68	3.60
Median (seconds)	5.83	5.78	5.89	5.68	7.66	4.10	5.52
3rd Qu. (seconds)	10.52	10.42	11.49	9.13	13.85	6.36	8.97
No. of Days in Study	30.00	30.00	30.00	8.00	27.00	30.00	6.00
No. of Responses	787.00	1573.00	1673.00	206.00	445.00	1438.00	220.00

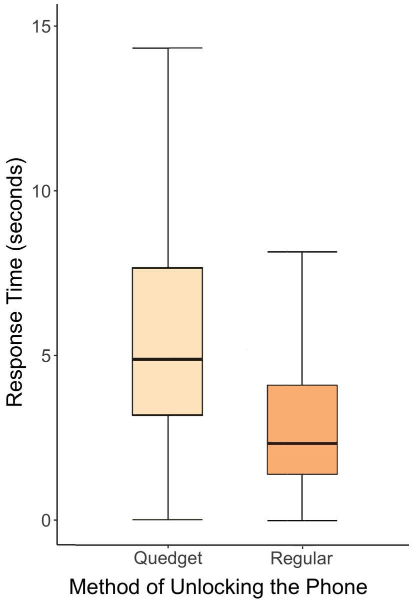


Fig. 5. Comparison of the time taken to unlock the phone with the standard smartphone lock screen interaction (Regular), with the time taken to unlock the screen with the Quedget tool. While the median unlock time with Quedget is slower by a few seconds, there is substantial overlap in our recorded interaction times for the two.

As a tactic for minimizing the usage time required to respond to questions, we designed interactions to require only one gesture. As a drawback to this *quick* interactive experience, we did not support complex question flows, where multiple questions could be presented at once, and branching questions are asked.

Using watch-sized displays and optimizing for brief interruption requires fragmenting multi-item measures so that each item can be asked separately[16]. Further research is needed to determine whether it is valid to combine items that were asked separately for comparison with a measure that was designed for the items to be asked together. To keep the interactions simple, we avoided using navigation. We reasoned that distributing information for one question across multiple views would require participants to learn how to activate views. However, the limited real-estate restricts our ability to provide navigational cues and other information labels. In *every* focus group, participants expressed the need for help interpreting images or mapping experiences onto them.

To address this, more fluid interactions need to be explored through which users can seamlessly traverse multiple points on the screen in order to provide active responses. Dissecting interactions like the pattern unlock or the number unlock (which requires multiple taps) could bring us closer to a design compatible with such requirements.

5.3 Unobtrusiveness Versus Participant Agency

In order to provide an unobtrusive experience, our signal is prompt-based as opposed to an active journal system that relies on the participant's adherence to the study protocol. However, we learned that the participants felt that this "push-down" interaction didn't allow for the capture of the entire spectrum of their experiences. Some participants had a desire to correct responses if they were reported disproportionately to others.

> *"Most of the time, for me at least – the activity questions – I seem to be either sleeping or watching TV. I think: 'Why didn't you ask me when I'm out with my friends, I want to say that socializing part?'"* -P26

Others, like P18, judged the sampling of their experiences to be erroneously homogeneous: *"On a single day I selected that commute like five times and it was like I'm commuting the whole day."*

We found that participants felt the need to communicate their experiences without the presence of prompts. In some cases, they also felt the way the prompts were scheduled did not holistically sample their experience. To circumvent these hurdles, the design considerations include:

- Provide users with agency that enables them to choose which question they want to answer during a particular prompt (similar to LogIn) via a multi-gestural interface where they can both select a question and answer it too, as part of the same interaction flow.
- Allow users to actively initiate assessment (in addition to the scheduled prompts). This method of journal-based reporting can give richer data and let users counteract the contextual inadequacies of the scheduled prompts. Combining the agency of a lock screen application like LogIn with the conditional notification paradigm of Quedget can meet such requirements.

6 DISCUSSION

In the following section, we discuss open questions for future research, grounding these in our study findings, and we propose some promising avenues for the design of EMA tools for eliciting subjective experiences.

6.1 Emotion Reporting Processes

There are two ways to judge one's subjective experiences such as emotions: one is associated with heuristics and memory while the other is associated with current feeling and information [32]. While semantic knowledge consists of beliefs one has about one's emotions, episodic knowledge is knowledge about one's emotions in a particular place at a particular time. Reporting how one experiences a feeling as it unfolds is not the same as summarizing emotions experienced during a two-week period. This is why memory-based, semantic reporting is less appropriate for EMA than episodic reporting.

Many reports in the focus groups suggested that participants expected to use *semantic judgments* about their emotions. This would explain why some did not understand why the same questions would be asked successively and why some participants thought that providing the same response repeatedly, without repeated reflection, would be a valid way to respond. This makes sense for people who consider their beliefs about their own emotions to be accurate predictors of actual patterns of feeling.

Our qualitative data provides initial support for the possibility that, in the context of EMA, the semantic reporting process can be faster and more attractive to a participant who does not want to spend time and effort judging emotion on a given occasion. Naturally, this behavior can compromise authenticity of the response. For example, in the case of the participant who simply chose from among the positive emotions when rushed, this strategy (to save time by consulting her belief that she tends to be positive) could qualify as a form of *satisficing* in emotional reporting.

Another common satisficing strategy we found entailed choosing to repeat the same emotion as was last reported, which results in a form of non-differentiation. The descriptions by students suggest that semantic reporting of emotion can save time by providing a response option that becomes well-practiced; the emotion judgment will have been established and its mapping to a response option will have been repeated over time. Likewise, non-differentiation can save time, because the required judgment is not about emotion; it is simply about what response was last given. These strategies may be patterns of satisficing that have emerged with the application of EMA methods, but user interface technology could be designed to attempt to disrupt that tendency.

6.2 Designing Against Haste

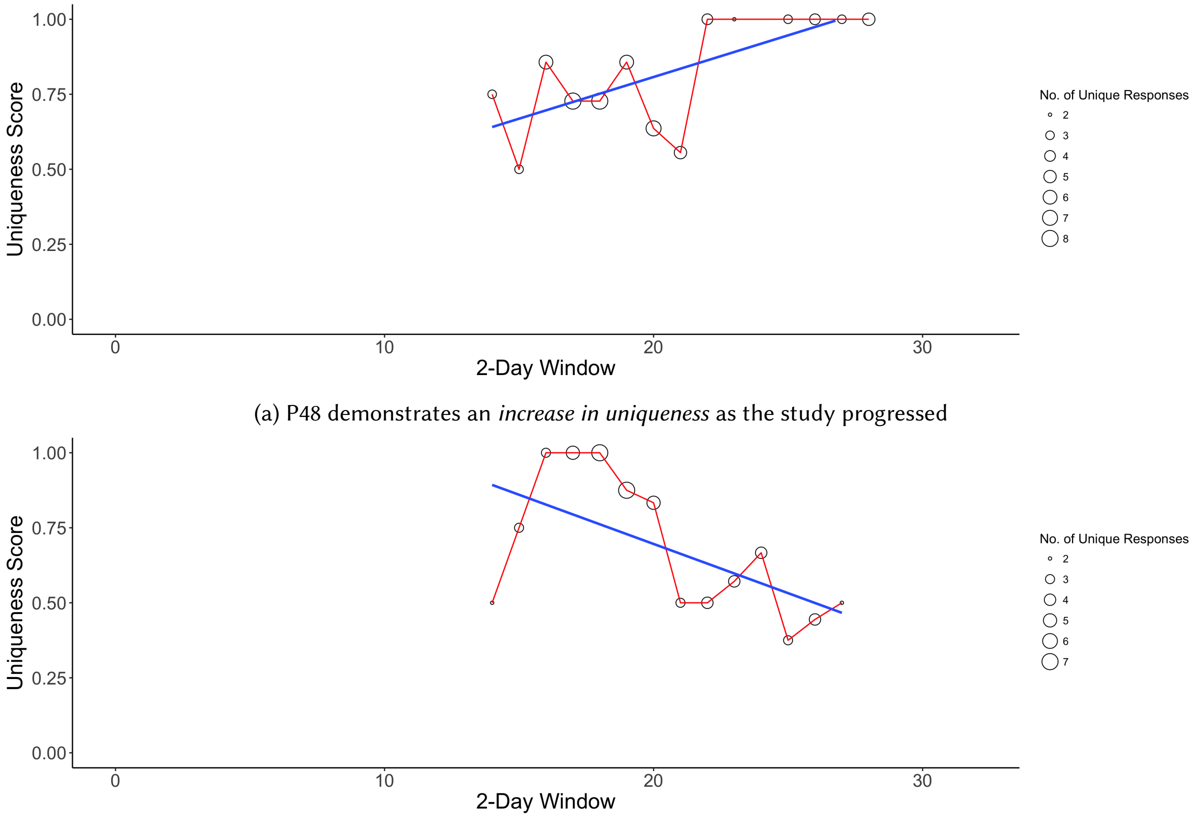
It is possible that our participants expected to be able to provide a proper response to any question quickly. EMA interactions are optimized for speed, so respondents could expect to spend little time reflecting on the response they give. While the lock-screen design does not, in itself, rush responses, the *facilitation of fast responses through microinteractions may reinforce the impression that only very brief interactions are appropriate.* A single-gesture lock-screen design could reinforce that expectation, causing frustration when responding takes longer. Future work is needed to investigate how to design user interfaces that counteract expectations that all logging interactions should be very quick, and even support users in taking the time they need. For example, UI elements could be designed to suggest that some considerable segment of time is appropriate. There could also be an option to indicate “I need more time” without skipping the question.

Research is also needed to investigate techniques for presenting the data back to participants, while they are participating, in ways that raise their interest in the *quality* of their responses. Researchers could investigate systematically whether interventions could be designed to measurably lengthen the time devoted to EMA responses that benefit from pausing to collect episodic information. Previous research suggests that feedback can assist in response quality [27]. The effectiveness of this approach for encouraging high-quality responses will depend on how data presentation appeals to the curiosity and interest of that participant population. It will also be affected by the perceived quality of the data—one motivation for avoiding hasty responses.

6.3 Reducing Ambiguities

Our participants adopted strategies to cope with at least two different types of ambiguity. One type stems from the *difficulty of rating subjective constructs* like self-esteem, on a scale. What criteria should be used for choosing between two adjacent response options? One solution to this problem is to select only from the extremes and to ignore the options in between. Questionnaire design research suggests that reducing ambiguity is feasible and that it can be accomplished by labeling all of the response options (instead of anchoring only at the extremes)[21]. However, such a solution requires presenting a great deal more information in the user interface. Mobile user interfaces may require novel approaches to presenting that information to overcome spatial constraints on the display.

Another type of ambiguity was caused by the *need to interpret images.* Expressions of confusion in the focus groups indicated that, on the image-based questions, sometimes the meaning of response options was unclear. One coping strategy we witnessed for dealing with ambiguity in the image-based question included *avoiding* certain response options that were unintelligible, and only selecting from the immediately comprehensible ones. The data produced by this restricted procedure loses accuracy, as the respondent is essentially constraining the possible response set to include only those responses that they can immediately interpret. We saw this behavior manifest in our log data: Figure 6 shows a trend for increasing diversity in PAM responses for a participant. Using a sliding 2-day window, the graph shows the proportion of responses that are unique within that window. Such a trend of increasing diversity could be interpreted as genuinely increasing variability in mood, or it could be read as the result of increasing familiarity with more of the available responses. This confound prevents making confident interpretations of emotional dynamics, and would best be avoided by ensuring that all response options are equally comprehensible at the outset of the study. Paradigms of mood depiction that are more apparently dimensional could eliminate the problems with comprehensibility that we saw in the image-based paradigms.



(a) P48 demonstrates an *increase in uniqueness* as the study progressed

(b) P28 demonstrates a *decline in uniqueness* as the study progressed

Fig. 6. We use a 2-day rolling window within which we calculate the *Uniqueness Score* (depicted on the Y axis) of a participant. This can be represented as the *Number of Unique Responses/Total Number of Responses* during those two days. A score of "1" signifies different responses for every question a participant answered.

Like PAM, MoodMap captures mood along the two dimensions of valence and arousal, but unlike PAM, it presents a two-dimensional space with labels to indicate what meanings are associated with regions on the space: "high-energy," "low energy," "positive," and "negative."

Because the images in our EMA study varied in many ways, our participants reported using eccentric methods for mapping emotions to images.

> *"What did you pick when your mood was moderate?"* -Moderator

> *"[I] think I picked something without people in it. . . "*-P16

Difficulty mapping images can introduce error in the data in other ways: increased task difficulty can make satisficing *more* likely [20].

> *"I just wish it was, you know, just four things. Well, that way I wouldn't even mind if there were more pop-ups coming up."* -P14

> *"After a couple of times I just associated [a] few of the pictures with [a] few of my moods. If I was feeling this way I can just pick this picture. I really didn't bother about all of those [other] pictures"* -P27

> *"Also, there are too many options. Before you could answer a question, you have to look through all the options. Most of them are very similar."* -P9

Mapping emotions onto a 4 x 4 grid of randomized images requires that participants do a new visual search task with each presentation. The cognitive load for this task could be too high for some. Future work could investigate whether the user's cognitive load could be reduced by making the dimensions that underlie mood measurement more obvious to them.

7 CONCLUSIONS

We have taken a qualitative approach to studying how students participating in a multi-week mobile sensing study experienced reporting on their mental well-being through microinteraction-based EMA. After listening to the experiences of 32 students who participated in a mobile sensing study for an average of 12 days, answering an average of 10 EMA prompts per day, we have found that various forms of satisficing are likely to be a concern, due in part to the interation design of the EMA tool. We found that some aspects of emotion assessment are *not* compatible with a microinteraction approach to EMA design. When sensing current emotional state can take quite a bit more than a few seconds, the increase in usage time may undo some of the benefits of microinteraction. Our study suggests that lock-screen and wearable paradigms are indeed well-suited to giving participants the agency to initiate or defer a question through microinteraction, but that future work is needed to determine how to design EMA prompts that encourage reflection, convey the appropriateness of pausing, and give permission to take the time required. Lock-screen EMA questionnaires facilitate lightweight, context-opportunistic approaches to record what are in fact effortful assessments. When complex judgments are called for, EMA tools should support such effort. To avoid arbitrary answers, it is also important for EMA interactions to better support participants in skipping questions as a valid response, and to provide feedback to enable awareness of prior responses. To that same end, researchers should use caution when linking study incentives or rewards to the number of questions answered, and they should reconsider the use of *response time* as an optimal metric in EMA for emotion.

By listening to students describe how they answered EMA questions about their well-being, we learned that students faced difficulties mapping their experiences onto image options that differ from their neighbors on many dimensions. Future work could offer systematic comparisons between image-based designs, like PAM, and more explicitly dimensional designs, like Mood Map.

ACKNOWLEDGMENTS

The authors would like to thank our study participants and the members of the CampusLife consortium. This work was supported by a gift from AIG Science, the National Natural Science Foundation of China under Grant No.: 61273304_a, and the Young Scientists' Support Program.

REFERENCES

[1] G. Abowd, Y. Han, P. Abowd, W. Horner, S. Tengler, and J. Chen. 2014. *Systems and methods for utilizing micro-interaction events on computing devices to administer questions.* Google Patents. https://www.google.com/patents/US20140298260

[2] Amelia Aldao and Susan Nolen-Hoeksema. 2012. The influence of context on the implementation of adaptive emotion regulation strategies. *Behaviour research and therapy* 50, 7 (2012), 493–501. http://www.sciencedirect.com/science/article/pii/S0005796712000733

[3] Daniel Lee Ashbrook. 2010. *Enabling mobile microinteractions.* Ph.D. Dissertation. Georgia Institute of Technology. https://smartech.gatech.edu/handle/1853/33986

[4] Jakob E. Bardram, Mads Frost, KÃąroly SzÃąntÃş, and Gabriela Marcu. 2012. The MONARCA Self-assessment System: A Persuasive Personal Monitoring System for Bipolar Patients. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium (IHI '12)*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/2110363.2110370

[5] Sunny Consolvo, Beverly Harrison, Ian Smith, Mike Y. Chen, Katherine Everitt, Jon Froehlich, and James A. Landay. 2007. Conducting in situ evaluations for and with ubiquitous computing technologies. *International Journal of Human-Computer Interaction* 22, 1-2 (2007), 103–118. http://www.tandfonline.com/doi/abs/10.1080/10447310709336957

[6] Sunny Consolvo and Miriam Walker. 2003. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing* 2, 2 (2003), 24–31. http://ieeexplore.ieee.org/abstract/document/1203750/

[7] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT* 2 (2015), 6. http://journal.frontiersin.org/article/10.3389/fict.2015.00006/full

[8] Gavan J. Fitzsimons and Vicki G. Morwitz. 1996. The Effect of Measuring Intent on Brand-Level Purchase Behavior. *Journal of Consumer Research* 23, 1 (June 1996), 1–11. https://doi.org/10.1086/209462

[9] Jon Froehlich, James Landay, Mike Chen, Sunny Consolvo, Beverly Harrison, and Ian Smith. 2006. An overview of in situ self report and the my experience tool. (2006). http://www.cs.umd.edu/~jonf/publications/Froehlich_AnOverviewOfInSituSelfReportAndTheMyExperienceTool_Unpublished2006.pdf

[10] James J. Gross and Ross A. Thompson. 2007. Emotion regulation: Conceptual foundations. (2007). http://psycnet.apa.org/psycinfo/2007-01392-001

[11] Gabriella M Harari, Nicholas D Lane, Rui Wang, Benjamin S Crosier, Andrew T Campbell, and Samuel D Gosling. 2016. Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science* 11, 6 (2016), 838–854.

[12] Gillian R. Hayes, Karen G. Cheng, Sen H. Hirano, Karen P. Tang, Marni S. Nagel, and Dianne E. Baker. 2014. Estrellita: A Mobile Capture and Access Tool for the Support of Preterm Infants and Their Caregivers. *ACM Trans. Comput.-Hum. Interact.* 21, 3 (June 2014), 19:1–19:28. https://doi.org/10.1145/2617574

[13] Todd F Heatherton and Dylan D Wagner. 2011. Cognitive neuroscience of self-regulation failure. *Trends in Cognitive Sciences* (Jan. 2011). https://doi.org/10.1016/j.tics.2010.12.005

[14] Javier Hernandez, Daniel McDuff, Christian Infante, Pattie Maes, Karen Quigley, and Rosalind Picard. 2016. Wearable ESM: differences in the experience sampling method across wearable devices. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 195–205. http://dl.acm.org/citation.cfm?id=2935340

[15] John Hicks, Nithya Ramanathan, Donnie Kim, Mohamad Monibi, Joshua Selsky, Mark Hansen, and Deborah Estrin. 2010. AndWellness: an open mobile system for activity and experience sampling. In *Wireless Health 2010*. ACM, 34–43. http://dl.acm.org/citation.cfm?id=1921087

[16] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. 2016. μEMA: Microinteraction-based Ecological Momentary Assessment (EMA) Using a Smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 1124–1128. https://doi.org/10.1145/2971648.2971717

[17] Stephen S. Intille, John Rondoni, Charles Kukla, Isabel Ancona, and Ling Bao. 2003. A Context-aware Experience Sampling Tool. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. ACM, New York, NY, USA, 972–973. https://doi.org/10.1145/765891.766101

[18] Daniel Kahneman and Alan B. Krueger. 2006. Developments in the measurement of subjective well-being. *The journal of economic perspectives* 20, 1 (2006), 3–24. http://www.ingentaconnect.com/content/aea/jep/2006/00000020/00000001/art00001

[19] Kurt Kroenke, Robert L. Spitzer, Janet B. W. Williams, and Bernd LÃűwe. 2009. An Ultra-Brief Screening Scale for Anxiety and Depression: The PHQâĂŞ4. *Psychosomatics* 50, 6 (Nov. 2009), 613–621. https://doi.org/10.1016/S0033-3182(09)70864-3

[20] Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology* 5, 3 (1991), 213–236. http://onlinelibrary.wiley.com/doi/10.1002/acp.2350050305/full

[21] Jon A. Krosnick and Stanley Presser. 2010. Question and questionnaire design. *Handbook of survey research* 2, 3 (2010), 263–314. https://books.google.com/books?hl=en&lr=&id=mMPDPXpTP-0C&oi=fnd&pg=PA263&dq=krosnick+and+presser+2010&ots=i4YWE0GrGn&sig=_oQIENpfQWHC2wrbjEbdMaB3mnE

[22] Jon A. Krosnick, Stanley Presser, Kaye Husbands Fealing, Steven Ruggles, and David Vannette. 2015. The future of survey research: challenges and opportunities. *The National Science Foundation Advisory Committee for the Social, Behavioral and Economic Sciences Subcommittee on Advancing SBE Survey Research.* (2015). https://www.nsf.gov/sbe/AC_Materials/The_Future_of_Survey_Research.pdf

[23] Santosh Kumar, Gregory D. Abowd, William T. Abraham, Mustafa alâĂŹAbsi, J. Gayle Beck, Duen Horng Chau, Tyson Condie, David E. Conroy, Emre Ertin, Deborah Estrin, and others. 2015. Center of excellence for mobile sensor data-to-knowledge (MD2K). *Journal of the American Medical Informatics Association* 22, 6 (2015), 1137–1142. http://jamia.oxfordjournals.org/content/22/6/1137.abstract

[24] Nicholas D. Lane, Mu Lin, Mashfiqui Mohammod, Xiaochao Yang, Hong Lu, Giuseppe Cardone, Shahid Ali, Afsaneh Doryab, Ethan Berke, Andrew T. Campbell, and others. 2014. Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing. *Mobile Networks and Applications* 19, 3 (2014), 345–359. http://link.springer.com/article/10.1007/s11036-013-0484-5

[25] Nicholas D. Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2011. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*. 23–26. http://www.cs.cornell.edu/~ms2749/pubs/PervasiveHealth_BeWell.pdf

[26] Sharan B. Merriam and others. 2002. Introduction to qualitative research. *Qualitative research in practice: Examples for discussion and analysis* 1 (2002), 1–17. http://stu.westga.edu/~bthibau1/MEDT%208484-%20Baylen/introduction_to_qualitative_research/introduction_to_qualitative_research.pdf

[27] Margaret E. Morris, Qusai Kathawala, Todd K. Leen, Ethan E. Gorenstein, Farzin Guilak, William DeLeeuw, and Michael Labhard. 2010. Mobile Therapy: Case Study Evaluations of a Cell Phone Application for Emotional Self-Awareness. *Journal of Medical Internet Research* 12, 2 (2010), e10. https://doi.org/10.2196/jmir.1371

[28] Vicki G. Morwitz, Eric Johnson, and David Schmittlein. 1993. Does measuring intent change behavior? *Journal of consumer research* 20, 1 (1993), 46–61. http://jcr.oxfordjournals.org/content/20/1/46.abstract

[29] Debbie S. Moskowitz and Simon N. Young. 2006. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of psychiatry & neuroscience: JPN* 31, 1 (2006), 13. http://search.proquest.com/openview/ff77a6319609405a55b7480b2ee9d990/1?pq-origsite=gscholar&cbl=30201

[30] John P. Pollak, Phil Adams, and Geri Gay. 2011. PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 725–734. http://dl.acm.org/citation.cfm?id=1979047

[31] Aditya Ponnada, Caitlin Haynes, Dharam Maniar, Justin Manjourides, and Stephen Intille. 2017. Microinteraction Ecological Momentary Assessment Response Rates: Effect of Microinteractions or the Smartwatch? *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 92 (Sept. 2017), 16 pages. https://doi.org/10.1145/3130957

[32] Michael D. Robinson and Gerald L. Clore. 2002. Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology* 83, 1 (2002), 198–215. https://doi.org/10.1037//0022-3514.83.1.198

[33] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*. Springer, 157–180. http://link.springer.com/chapter/10.1007/978-90-481-2354-4_8

[34] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. Ecological Momentary Assessment. *Annual Review of Clinical Psychology* 4, 1 (April 2008), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

[35] Khai N. Truong, Thariq Shihipar, and Daniel J. Wigdor. 2014. Slide to X: unlocking the potential of smartphone unlocking. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3635–3644. http://dl.acm.org/citation.cfm?id=2557044

[36] Rajan Vaish, Keith Wyngarden, Jingshu Chen, Brandon Cheung, and Michael S. Bernstein. 2014. Twitch Crowdsourcing: Crowd Contributions in Short Bursts of Time. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3645–3654. https://doi.org/10.1145/2556288.2556996

[37] Rajan Vaish, Keith Wyngarden, Jingshu Chen, Brandon Cheung, and Michael S. Bernstein. 2014. Twitch crowdsourcing: crowd contributions in short bursts of time. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3645–3654. http://dl.acm.org/citation.cfm?id=2556996

[38] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14. http://dl.acm.org/citation.cfm?id=2632054

[39] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 295–306. http://dl.acm.org/citation.cfm?id=2804251

[40] Ladd Wheeler and Harry T. Reis. 1991. Self-Recording of Everyday Life Events: Origins, Types, and Uses. *Journal of Personality* 59, 3 (Sept. 1991), 339–354. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x

[41] Xiaoyi Zhang, Laura R. Pina, and James Fogarty. 2016. Examining Unlock Journaling with Diaries and Reminders for In Situ Self-Report in Health and Wellness. ACM Press, 5658–5664. https://doi.org/10.1145/2858036.2858360

Received May 2017; revised November 2017; accepted January 2018