Contextual Dissonance: Design Bias in Sensor-Based Experience Sampling Methods

Neal Lathia[1], **Kiran K. Rachuri**[1], **Cecilia Mascolo**[1], **Peter J. Rentfrow**[2]
[1]Computer Laboratory, [2]Department of Social and Developmental Psychology
University of Cambridge, United Kingdom
neal.lathia, kiran.rachuri, cecilia.mascolo@cl.cam.ac.uk, pjr39@cam.ac.uk

ABSTRACT

The Experience Sampling Method (ESM) has been widely used to collect longitudinal survey data from participants; in this domain, smartphone sensors are now used to augment the context-awareness of sampling strategies. In this paper, we study the effect of ESM design choices on the inferences that can be made from participants sensor data, and on the variance in survey responses that can be collected. In particular, we answer the question: are the behavioural inferences that a researcher makes with a trigger-defined subsample of sensor data biased by the sampling strategy's design? We demonstrate that different single-sensor sampling strategies will result in what we refer to as *contextual dissonance*: a disagreement in how much different behaviours are represented in the aggregated sensor data. These results are not only relevant to researchers who use the ESM, but call for future work into strategies that may alleviate the biases that we measure.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation: Miscellaneous

General Terms

Human Factors; Design; Mobile; Psychology

INTRODUCTION

As the usage of sensor-rich smartphones continues to grow, so too does the opportunity to leverage these devices in order to gain insight into our daily lives. Smartphones are an ideal platform for conducting Experience Sampling Method (ESM) based studies, where participants respond to short questionnaires when notified to do so by the device [3, 9]. Similarly, smartphones are increasingly used to unobtrusively sense peoples' contexts [18]; the intersection of these methodologies allows researchers to augment their insight into behaviour by leveraging both granular sensor data about participants' contexts and subjective survey responses about their views, thoughts, or feelings.

At the heart of any sensor-augmented ESM study lies the design decision about when, or under which conditions, a notification to complete a survey should be fired. Studies have, for example, been designed to collect data at random intervals [1] or using sensor states as triggers [9]. While the latter is often motivated by directly tying a device state with a device-related assessment (e.g., plugging in the phone triggering questions about phone charging [9]), both of these methods have been used by researchers to make inferences and test hypotheses about broad, non-device specific aspects of participants' behaviours, such as daily events and moods [4] and sustainable transportation choices [10].

This methodology assumes that the design choice of which trigger to use will not affect (or, indeed, will even augment the accuracy of) the contextual data that can be used to learn about participants. In doing so, these studies do not take into account the effect that the designed sampling strategy has on the conclusions they infer about participants' behaviours. However, these behaviours are likely to be habitual or, more broadly, variantly distributed across each day. For example, since people may split the majority of their time between home and work, sampling randomly is likely to fail capturing participants in new locations. While this could easily be solved by using a location-based sensor trigger, it is not clear how doing so affects sampling from the broader set of sensors that researchers may be collecting data from (i.e., how would location-based sampling bias the data about participants activity levels?).

In this paper, we study the effect of ESM design choices on the inferences that can be made from participants' sensor data, and on the variance in survey responses that can be collected from them. In particular we answer the question: *are the behavioural inferences that a researcher makes with a time or trigger-defined subsample of sensor data biased by the sampling strategy's design?* We demonstrate that different single-sensor sampling strategies will result in what we refer to as *contextual dissonance*: a disagreement in how much different behaviours are represented in the aggregated sensor data.

To do so, we designed a prototype smartphone-based system that we used to collect sensor data from 22 people who participated in a 1-month ESM-based study about momentary mood. We built an Android application that can be remotely reconfigured (in terms of survey questions, survey triggers,

UbiComp'13, September 8–12, 2013, Zurich, Switzerland.
Copyright © 2013 ACM 978-1-4503-1770-2/13/09...$15.00.
DOI string from ACM form confirmation

sensors to sample, and sampling parameters) in order to unobtrusively change our ESM study's structure on the go: we used this to vary the triggering mechanisms that notified users to answer the questionnaire. Using the results of this deployment, we empirically quantify the extent that sensor-triggered ESM designs would limit the breadth of behavioural data that researcher's may capture about each individual in their study. We then directly measure participants' compliance alongside the quality and variance in the survey response data that we obtained, and found that changing our study design not only changed the inferences we would have made about participants' behaviours, but also produced statistically significant differences in their responses. We close by discussing how this affects the design of future smartphone-based sensor-augmented ESM studies.

PROBLEM FORMULATION AND ASSUMPTIONS

The experience sampling method is traditionally used as a means of capturing survey responses from participants in naturalistic settings [16, 24]. Inherently, designing an ESM study requires defining a cue that will be used to notify that a survey should be completed: in this setting, sensors have been used to augment the context-awareness[1] of automated signals to complete surveys, as well as provide an additional source of behavioural data to researchers [21]. In this section, we formalise our assumptions about how this design may result in the bias introduced above. In particular, we focus on social-psychology and behavioural research scenarios, and assume that the following key requirements hold:

1. **Subjective Responses**. ESM studies often seek to obtain a diverse set of responses that facilitate both within- and between-participant/context analysis [16]. We focus on those studies where the surveys solicit *subjective* responses about a time-varying target variable related to the participant; these include, for example, studies related to mood and well-being [6, 13].

2. **Context Sensing**. The goal of collecting smartphone sensor data is to augment the responses that participants provide by means of an unobtrusive measurement of the broad context surrounding the device. For example, sensing could include polling data from (at least) the location, microphone, and accelerometer sensors. However, due to the energy constraints of modern-day smartphones, we assume that sensing from the full set of sensors will be limited to those moments when the ESM trigger is fired, and that data will only be regularly polled from the pre-defined trigger sensor.

We depict these requirements via the following example: a study (like below) that aims to measure how self-reported mood compares to *a set of* smartphone sensor streams. ESM surveys are therefore used to assess the participants' mood, and context sensing is used to measure facets of their behaviour. More formally, we define a smartphone's context c_t in a time period t as data from the set S of n streams the device can sense from:

$$c_t = \bigcup_{i \in S} (s_{i,t}) = (s_{1,t}, s_{2,t}, ..., s_{n,t}) \quad (1)$$

Naturally, c_t may be sparse: data from a particular sensor may not be available or have been sensed in the given time window t. Experience sampling studies are typically designed to use time or the state of a single contextual item ($s_i \in S$) in order to trigger a survey notification. For the latter, the probability that a notification is triggered is proportional to the probability that the sensor i's state at time t meets a trigger condition α_i, which will be subject to a particular distribution. In other words, given a classifier f_i that outputs one of a set of states from i's sensor data, we define the underlying distribution of α_i as X:

$$P(f_i(s_{i,t}) = \alpha_i) \sim X \quad (2)$$

which we assume will vary between sensors. In practice, we would like a collection of contexts C that is representative of the participant's overall behaviour alongside multiple survey responses. However, using a sensor-based trigger preemptively imposes restrictions on when notifications can appear. The question we ask is related to the extent that one sensor's underlying distribution affects the quality of data that can be collected from other distributions. More formally, if we only sample when $f_n(s_{n,t}) = \alpha_n$, then for every other sensor i/state α_i we can, at best, only learn the conditional distribution Y:

$$P(f_i(s_{i,t}) = \alpha_i | f_n(s_{n,t}) = \alpha_n) \sim Y \quad (3)$$

In the following, we set out to investigate the extent that $X \sim Y$, and how different Ys will have variant survey responses: any differences between these two distributions will unveil the extent that sampling from users' contexts by sensor-triggers introduces a design bias into ESM studies that aim to make inferences about participants' behaviour from the aggregated sensor data.

A sampling strategy that only captured the participant when they were inactive, at home, in the evening would produce data that could only speak to how mood fluctuates in this well-defined context. A location-based sampling strategy, instead, may increase the geographic variance of responses, but would it continue to trigger notifications only when the user is inactive? Finally, a random sampling strategy may mostly capture data corresponding to dominant contexts: the ones that occur the most frequently.

SYSTEM OVERVIEW

This section gives an overview of an Android application we built in order to conduct ESM studies that address the research challenge above. Like other systems [9], the app contains a set of generic survey-question interfaces, configurable survey triggers, and a sensor manager that collects data. However, we also designed the system to be remotely reconfigurable without requiring the app to be manually updated. In the following, we provide further details about each of these open-sourced components (for details, see [14]):

Collecting Sensor Data. We designed a library that abstracts the complexity of managing sensors and provides a unified

[1] E.g., see http://web.mit.edu/caesproject/

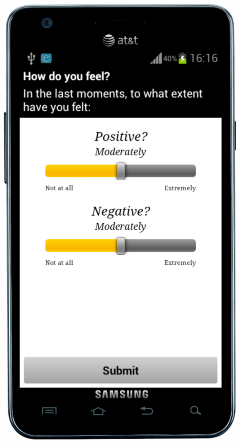
way to access all sensor data [14]. This library starts, pauses, and stops polling from sensors according to reconfigurable parameters and provides a publish/subscribe service to access the data. It supports two types of sensors: *physical sensors* and *software sensors*. Physical sensors include the accelerometer, microphone, proximity, screen status, and radios such as Bluetooth, GPS, and Wi-Fi fingerprint (i.e., in the Android operating system, those that must be actively polled for data). Software sensors, instead, capture phone calls, screen on/off events and SMS activity (or, those events that are broadcast by the Android operating system with assistance from the phone's hardware).

Clock, Sensor, and Hybrid Triggers. A request to complete a survey is given to a user via a notification, which vibrates and/or makes the phone ring (based on the user's volume settings) and places an icon in the notification bar, much like receiving an SMS. The control of these requests is defined by a set of *triggers*: we implemented three types. The first are *time-based*, such as random-choice and on a set interval. The second group are *sensor-based*; triggers that notify the user to complete a survey when a particular sensor event occurs. Finally, we implemented *hybrid* triggers; these use a clock-based trigger to start and stop a sensor-based trigger. This kind of trigger may be used, for example, to sample from the accelerometer once an hour for five minutes and notify the user if the sensor data indicates that the device is not moving.

User Preferences and Data Collection Awareness. We added constraints on the application's ability to notify the user to complete surveys: all triggers must comply with a set of user preferences, which users could edit via the application's menu. These settings include the maximum number of surveys that can be triggered in one day, and the times when users were available to respond (by default, set to maximum 2 notifications between 08:00 AM and 10:00 PM). Additionally, users could "pause" all sensing for 30 minutes, and could renew this time-out by selecting the pause option as many times as they wanted to.

Remote Reconfiguration. The sensor parameters, surveys and triggers can be configured remotely to allow for automatic updating without the user's involvement. The native application periodically downloads the configuration files of the ongoing experiment from the remote server and updates each of the previous components. The primary objective of the design, as described above, was to allow for experiments to be conducted without requiring participants to reinstall or otherwise manually update the application. In the following section, we describe how we used this feature in order to address our research question by collecting data across a variety of ESM protocols.

Storing and Uploading Data. A separate component receives and stores data from all the sensors, which is then periodically uploaded to a remote server using a background process. This helps in avoiding manual transfer of collected sensor and survey data, thereby reducing the involvement of users in the administration of any experiment. We configured the service to upload data whenever the phone is connected to a Wi-Fi network, to avoid incurring any data-transfer cost on the participants. This assumes that participants will, at some point, connect their phone to a Wi-Fi network, which we believe is a reasonable assumption.



How do you feel?

In the last moments, to what extent have you felt:

Positive?
Moderately

Negative?
Moderately

Figure 1. Survey Question Interface: showing the first questions of the mood survey, which asked for ratings in answer to the question "In the last few moments, to what extent have you felt ..." (a) positive and (b) negative.

Survey Interfaces. We implemented a set of generic interfaces that support a variety of survey question-types, including rating, single-choice, multiple-choice, and free-text questions; the questions themselves, as well as the sequence of questions contained within a survey, are also defined in a JSON-formatted configuration file that our retrieved by the application from our server.

EXPERIMENT DESIGN AND IMPLEMENTATION

Using the system described above, we designed a study that encoded the conditions required for our research goals to be met. To describe these, we revisit the assumptions described in the 'Problem Formulation and Assumptions' section:

1. **Survey Questions**. We focused on momentary mood assessment: in this case, the role of sensor sampling is to unobtrusively collect data about the participant's current context prior to asking them, primarily, about their feelings. The survey contained the following 4 questions: (1) *Mood Rating*, which asked for Likert ratings for how *positive* and *negative* they felt, each ranging from "Not at all" to "Extremely," (2) *Location*, selected from categories like home or work, (3) *Social Setting* or "people around you," such as friends, family, or nobody, and (4) a *Mood Tag*, an optional free-text word describing their current mood. Figure 1 shows a screen shot for the first questions in the survey.

2. **Context Sensing**. Given the geographic dispersion of our participants (details below), we opted to not collect data from the Bluetooth radio; we sensed from the accelerometer, location, proximity, and microphone sensors, and logged all screen, SMS, and call events. We sensed data continuously, using similar parameters as those in previous work [22]; a pre-trial feasibility study with 7 members

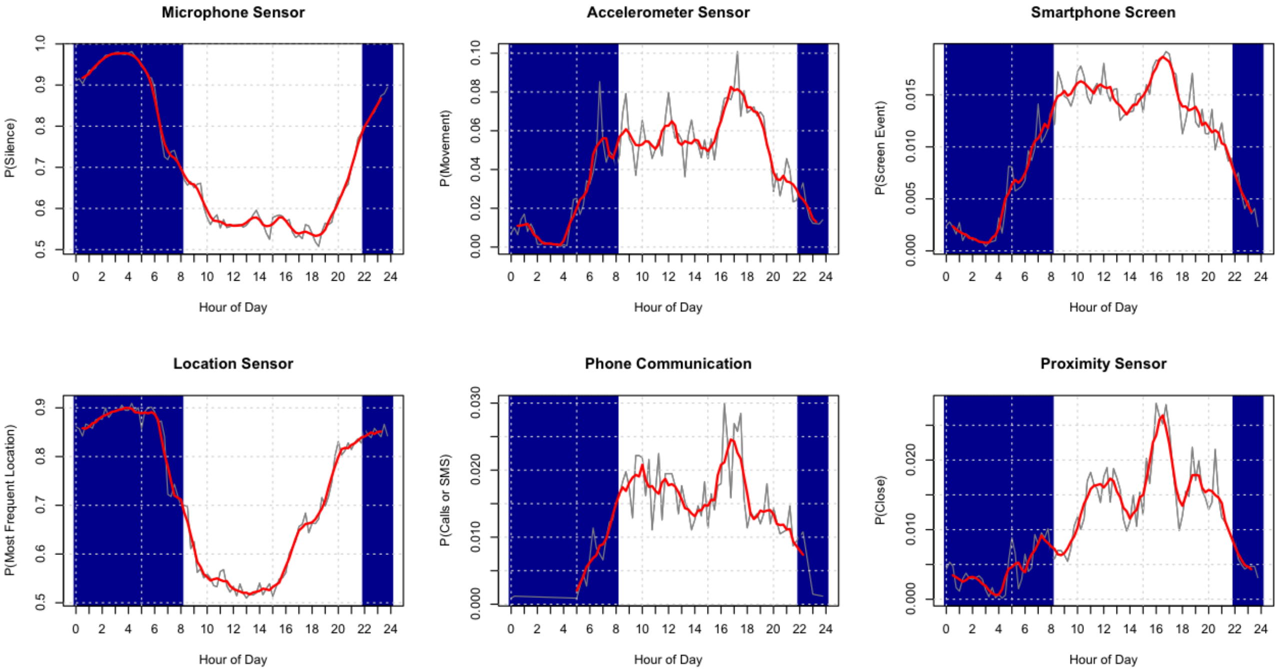


Figure 2. Aggregate Distribution of Sensors' Data: the red line is a 2-sided moving average. Since notifications could (by default) only be triggered between 8:00AM and 10:00PM, we added a blue shading during out-of-bound times.

of our research group, indicated that the effect of this was that devices now required daily recharging, which has been reported to be the routine of most users [20].

Participant Recruitment and Deployment Details

To avoid guiding participants' response rates, we sought for volunteers who would use their own device to participate in an experiment without being financially rewarded for doing so. We searched for volunteers by circulating calls for participation on Twitter and Facebook, the application's website[2], and sending advertisements across a number of universities' mailing lists. Enrolling entailed completing an online form (name, email address, location, Android device type) and agreeing with an informed consent statement required by the ethical approval granted to our study. A total of 36 people registered to join the trial.

We used participants' email addresses in order to create trigger ordering subgroups, send weekly updates about the experiment, and solicit any feedback or questions that they may have had. Once the trial finished, we emailed all participants a final, online survey. This set of questions included demographics and solicited open ended feedback about the trial. We opted to email participants this survey rather than deliver it via the mobile application in order to allow participants to input longer free-text responses.

We conducted a month-long data collection trial throughout August 2012. In order to minimise the effect of the ordering of triggers, we randomly divided the volunteers who signed up into two groups, who each obtained the four triggers in a different, randomised order. The participants were given a group-identifier that they used to activate the application after they downloaded it from the Google Play Market. During each of the four weeks, the application was remotely re-configured in order to use a different trigger. We ultimately collected valid data from 22 unique devices which reported locations from the United Kingdom (15), France (2), Spain (1), Italy (1), Portugal (1), Slovenia (1), and Estonia (1). The post-trial demographic survey was completed only 12 times (54.54% of the participants): those who answered are between 22 and 40 years of age (average: 30.91 ± 5.16), and one-third of them are female.

ANALYSIS OF RESULTS

We now detail the results of analysing the data that was collected throughout the trial. First, a brief overview of the its aggregate statistics: as above, we received data from 22 devices. As participants were free to leave the study at any time, we measured the extent that they continued contributing sensor data by having the application installed. The time differences between the earliest and latest data received from participants ranged from 2.26 days to 30.57 days (the study lasted 30 days), with an aggregate average of 17.99 ± 10.44 days of data per participant. A total of 13 (59.09%) participants contributed sensor data for more than half of the month, and 8 (36.36%) persisted for over 90% of the trial's duration. Some participants who left the trial directly contacted us to let us know that they were uninstalling the application. The reasons they gave included leaving their country of residence (to go on holiday) and the degradation of battery life. However, we note that other participants emailed us with similar concerns (both travel and battery) but opted to remain in the study.

[2] `http://emotionsense.org/`

Trigger	Sensor	Description (α_i)
Time	None	All times eligible
Movement	Accelerometer	Non-stationary sample
Location	GPS/Coarse Loc.	At most frequent location
Sound	Microphone	Non-silent audio sample
Device	Screen	Screen events received
Social	Phone/SMS	Call/SMS sent/received

Table 1. Trigger Types, covering a range of conditions related to location, movement, device usage, and time.

Individual Distributions of Sensor Contexts

To investigate how the distributions of sensor streams vary by sensor, we first split the data from all users, by sensor, into 15 minute bins. We then aggregated each set as follows; Figure 2 shows the resulting raw (grey) and smoothed (2-sided moving average, red) distributions, and blue shading indicates where surveys were (by default) not allowed to be triggered.

- **Microphone**. We used an amplitude-threshold to classify audio samples into "silent" and "non-silent." We then compute, for each bin, the proportion of collected samples that are silent. As expected, silent samples are most prevalent at night time, while during the day samples are less than 60% silent.
- **Accelerometer**. We compare the magnitude of acceleration to a threshold that classifies a sample as "moving" or "stationary." For each bin, we compute the proportion of samples classified as "moving." Overall, a very small proportion of samples indicate movement, and those that do tend to be found in the later hours of the day.
- **Location**. First, we clustered GPS latitude/longitude pairs based on geographic distance: we assigned a sample to an existing cluster if it was less than 1 km from the cluster's centroid; otherwise, we created a new cluster. Next, we picked the cluster that had the highest number of samples and use this as the most frequent location. Finally, for each bin, we compute the proportion of location samples where the user is at the most frequent location. At night time, over 85% of samples found the participant in this location, indicating that it is likely to be their home. However, more than 50% of the day-time samples were also within the bounds of this location.
- **Device Interaction**. By logging those times that the users turns on their screen, we can compute the distribution for when users interact with their device. We count the number of times a user turns on their phone screen in a given time bin, and normalise this by the sum of times that the screen.
- **Communication**. We logged SMS events (both sent, received, and when the SMS inbox was edited), as well as phone state events; we aggregate by counting the number of call/SMS events in each time bin, and normalising them by the sum of events.
- **Proximity**. The proximity sensor detects when an object is within a short vicinity of the phone's screen. As per the above two entries, we count the number of times that this sensor indicates a near object, and plot the normalised distribution over time.

This aggregated data gives a first glimpse into the temporal non-uniformity of sensor streams that may be used to trigger an ESM survey notification. In the following, we investigate the extent that using this data to cue triggers will provide variant views of participants' behaviour.

Effect of Context Triggers on Data Collection

The analysis above shows how the state of different sensors have varying distributions. To what extent does designing an ESM study to trigger surveys based on one sensor influence researchers' ability to collect varying data from each other contextual dimension? To investigate this, we selected a set of 6 different triggers related to movement, location, ambient sound, and both device and social interaction (Table 1); these cover the set of sensors that we collected data from and therefore also cover a diverse set of behaviours.

We first combined the sensor data that was collected into a matrix M, where each row represents a participant's current context (Equation 1). More specifically, each row is a 15-minute time window and each column is a sensor's state. A given row c_t of M is of the form:

$$c_t = (acc_t, loc_t, mic_t, scr_t, sms_t, cal_t) \qquad (4)$$

Which captures the aggregated sensor readings for the accelerometer (acc), location (loc), audio amplitude (mic), and device interaction (scr) sensors, as well as the number of text message (sms) and calls (cal) events made within that time period. We pre-processed this matrix by removing all rows that were ineligible for survey notifications (i.e., in time windows before 8AM or after 10PM). Given this representation, we repeat, for each trigger (in Table 1) a two-step process:

1. **Filter by trigger condition**. We create a matrix $T \subseteq M$, which contains all rows from M where a particular trigger condition is met. For example, if a notification is set to trigger when the device is used for social interaction (i.e., making or receiving a call or SMS), then we keep only those rows where $(sms_t + cal_t) > 0$. What remains is a set of *candidate* times for surveys to be triggered, or the set of contexts that an ESM study will sample from: comparing the relative size of T and M allows us to see how often each particular trigger is likely to occur.
2. **Evaluate for Bias** in matrix T across all contextual dimensions by counting the number of rows where other trigger conditions are met (Equation 3). For example, given the set of rows where, as above, a social interaction trigger may fire, we count the proportion of times where $acc_t =$"non-stationary." More formally, for each sensor, we compute:

$$P(f_i(s_{i,t}) = \alpha_i|T) = \frac{|c_t \in T : f_i((s_{i,t}) = \alpha_i)|}{|T|} \qquad (5)$$

This exercise exposes the extent that triggering based on a single sensor can vary the extent that other sensor patterns are observed (Table 2). For example, consider the *Sound* trigger, which would fire a notification when a non-silent audio

Trigger	Time	Movement	Location	Sound	Device	Social
Time	100	10.61	41.40	37.78	23.14	4.60
Movement	10.61	100	39.24	95.23	64.73	14.43
Location	41.40	10.06	100	43.01	30.14	4.85
Sound	37.78	26.75	47.12	100	46.03	9.48
Device	23.14	29.69	53.92	75.15	100	18.71
Social	4.60	33.30	43.62	77.90	94.16	100

Table 2. Contextual Variance Across Triggers: each row is a sensing trigger; each column is the proportion of the resulting sensor data that contains the given column feature. For example, 10.61% of the all data contains movement, and 95.23% of that subset also contains non-silent audio samples.

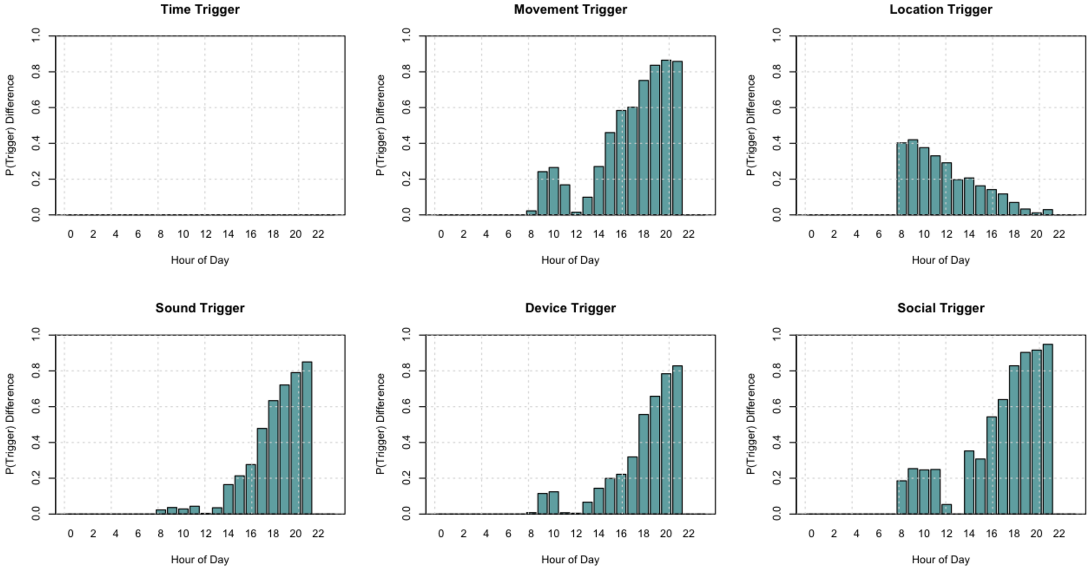


Figure 3. Time Biases: The underlying distribution of sensor data (Figure 2) affects the probability of sampling at different times of day. These plots show the difference between a given trigger's sampling probability and the uniform distribution.

sample is recorded. On aggregate, only 37.78% of participants' time bins (i.e., rows in M) met this condition. When it was met, the data indicated movement 26.75% of the time (i.e., rows in T), while unfiltered data from M indicated movement in only 10.61% of the bins. Similarly, when the sound sensors detected non silent samples, participants were marginally more often at home (47.12% vs. 41.40%), were using their device more often (46.03% vs. 23.14%), and were communicating more frequently (9.48% vs. 4.60%). *Naturally, this implies that a researcher who makes behavioural inferences based on a non-silent audio ESM trigger would overestimate the extent that the participant is active, using the device, and communicating with others*. To give another example, the movement-based trigger would keep only 10.61% the available rows in M. In this subset, users are less frequently at home (39.24% vs. 41.40%), more frequently in non-silent contexts (95.23% vs. 37.78%), and both using their device (64.73% vs. 23.14%) and communicating with others (14.43% vs. 4.60%) more often. *Again, inferences about participants' behaviour would over estimate the ambient sound, device usage, and communication patterns*.

Trigger	Euclidean Distance
Time	0
Location	0.907
Device	1.513
Movement	1.976
Sound	1.629
Social	2.095

Table 3. Temporal Sampling Similarity: Euclidean distance between the uniform time-based trigger and each sensor-based trigger.

We also examined the extent that sensor triggers will *temporally* bias the probability of collecting data. To do so, we repeated the procedure above: for each trigger, we generate T by filtering the matrix M based on when trigger conditions are met, and then use T to compute a normalised distribution of the probability that a survey will be triggered in any given hour of the day (as per Figure 2). If a trigger were not temporally biased, this distribution would be uniform across the day; since sensor events have non-uniform distributions, this is not the case. In order to uncover those hours of the day that are less likely to be sampled from, Figure 3 shows the difference between a uniform distribution and the distribution of each trigger's hourly sampling probability. We quantified these differences by computing the Euclidean distance between the normalised hourly notification probability vectors. For two vectors p and q, the Euclidean distance is:

$$d(p,q) = \sqrt{\sum_i (q_i - p_i)^2} \tag{6}$$

This similarity metric gives higher values for pairs of vectors that are more dissimilar; it indicates that a sampling strategy based on when participants use their phone for communication purposes (i.e., the social trigger) is the least similar to the uniform (time) trigger. In general, the majority of the sensor triggers are biased away from sampling in the later hours of the day. The frequent location-based trigger is less likely to sample from the hours near the middle of the day and afternoon; sampling based on sound, instead, would diminish the likelihood of sampling from the early morning and late evening.

Distributions of Survey Response Data

The analysis above, which was based on data that had been unobtrusively and regularly sampled from participants' smartphones, allowed us to quantify how sampling with different triggers *would have* provided different views of the users' contexts. In this section, we investigate the extent that we were able to collect variant response data from our participants during the trial in August 2012. An exhaustive exploration of all triggers is beyond the scope of our work: instead, we selected four triggers that were deemed to be sufficiently different from one another to warrant testing:

1. **Hybrid Microphone Triggers**. Past smartphone-based studies have used audio samples to analyse emotional expressivity [17, 23]. We thus implemented two hybrid triggers that were based on the participants' microphones. More specifically, the triggers would select N random moments of the day to begin sampling audio data (where N is the user-defined maximum survey cap, with default value 2), and would then ask participants to complete a questionnaire when a non-silent audio data sample had been obtained, or if an extended period of silence (60 minutes) had passed. The only difference between the two triggers was how quickly after obtaining a non-silent reading the notification would be triggered: the first trigger would send it immediately (Mic_i), while the second would wait for a silent moment after that before triggering (Mic_w); while we posit that the latter trigger may be less intrusive on participants' conversations, this choice allows us to see how contexts vary across minor sensor-trigger changes.
2. **Device Interaction Triggers**. The first device-based trigger, $Comms$, asked participants to complete a self-report just after having used the phone for communication purposes (i.e., after hanging up the phone or after receiving a text message). The second trigger ($Screen$) took a broader view on device interaction, and pinged the user for a self-report with a set probability after the device's screen had been on for 30 - 90 seconds. These triggers assume that it would be less intrusive to ask participants about their mood during or right after moments that they have been interacting with their smartphone, rather than interrupting them from a different task as the previous triggers may do.

	Mic$_i$	**Mic$_w$**	**Comms**	**Screen**
Notifications Sent				
Total	138	144	92	113
Users	13	15	18	14
Average	10.62	9.60	5.11	8.07
Std. Dev	6.50	7.21	3.69	7.24
Surveys Received				
Total	127	116	78	85
Average	9.77	7.73	4.33	6.07
Std. Dev	6.57	7.45	3.48	7.23
Compliance (%)				
Average	67.88	73.21	62.22	71.22
Std. Dev	37.04	25.32	42.53	34.82
Delay (Minutes)				
Average	20.68	56.95	42.23	48.07
Std Dev.	21.48	52.48	137.36	56.10

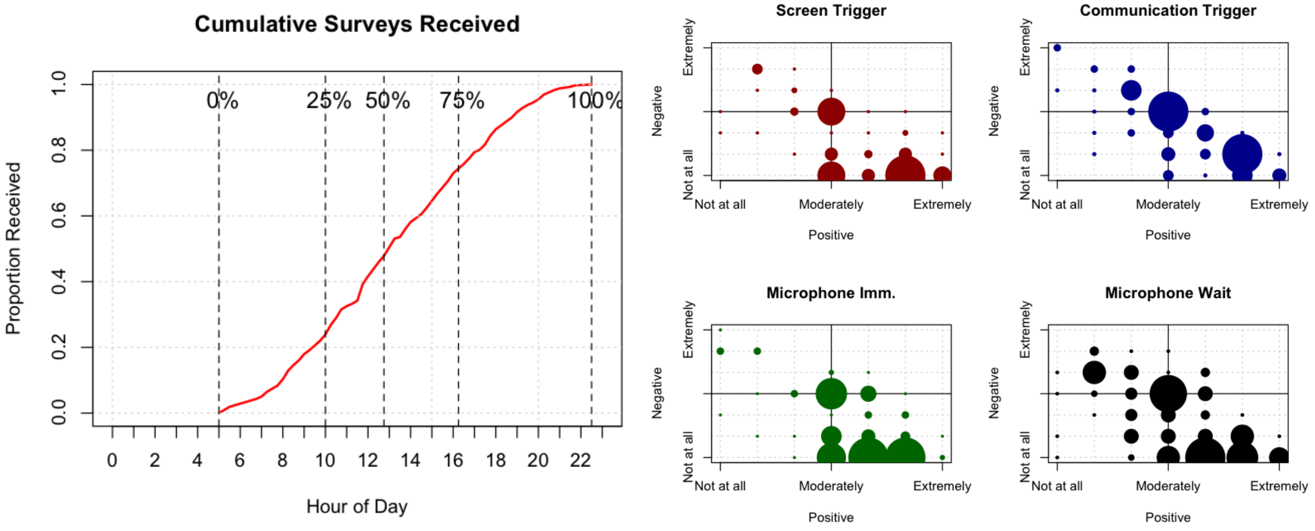
Table 4. A comparison of the variability in number of notifications sent, survey responses received, compliance, and delay between the four triggers we tested.

	Positive		**Negative**	
Trigger	**Mean**	**Median**	**Mean**	**Median**
Screen	3.69±1.50	3	1.39±1.59	1
Comms	3.40±1.54	3	2.05±1.59	2
Mic$_i$	3.25±1.49	3	1.73±1.60	1
Mic$_w$	3.56±1.30	4	1.39±1.59	1

Table 5. Affect Rating Responses: Mean, Standard Deviation, and Median Per Trigger Group.

Table 4 summarises the aggregate counts of all the data related to ESM responses: the number of notifications, responses received, compliance, and delay, across all four different triggers. Overall, the microphone-based triggers resulted in a higher number of notifications being sent; this is due to the fact that they do not rely on any device-usage conditions being met. Since, as above, not all of our participants remained in the study throughout the entire month, we also counted the number of users who received each type of notification. The microphone-based triggers produced a higher per-user average number of notifications; conversely, the communication-based trigger produced the lowest average number of notifications per participant (5.11 ± 3.69), indicating that participants were not using their phone for its call/SMS functionality throughout the day. This result agrees with the previous analysis and shows how a trigger selection will directly impact the amount of data that researchers can collect.

Given this set of responses, we first analysed the extent that they had been uniformly sampled across the day. Figure 4(a) shows the cumulative distribution of the response set: 25% of surveys were completed before 10AM; 50% before approximately 1PM, and 75% just after 4PM. While we do find that surveys are approximately uniformly split between morning and afternoon, the afternoon subset is skewed towards the earlier hours (1PM - 4PM) over the later ones. This result agrees with the temporal bias in sampling that we observed above: Figure 3 indicated that sampling strategies would be biased



(a) Response Time CDF

(b) Response Ratings Across Trigger Groups

Figure 4. Survey Responses: time of day cumulative distribution (left) and frequency plots for each pair of positive/negative affect ratings (right).

away from the later hours of the day, and indeed our collected data also indicates this feature.

Recall that the "Experiment Design and Implementation" section describes the full set of survey questions we used: the first questions of the survey asked for ratings related to the participant's current positive and negative affect (e.g., "In the last few moments, to what extent have you felt positive?" where ratings went from "Not at all" to "Extremely"). Figure 4(b) plots the frequency of each pair of ratings: the x-axis is the positive rating, the y-axis is the negative one, and the size of each dot is proportional to the normalised frequency of that pair of ratings; Table 5 gives the resulting sample means, standard deviations, and medians.

If the design of our experiments did not have any influence on the data that we collected, we would expect that, on aggregate, the ratings sets from each trigger group would be consistent with one another. However, the summary statistics indicate differences between each rating sample, in terms of varying means and medians. We thus conducted further tests to determine whether the differences between each set of ratings was statistically significant; we evaluate these separately since the literature reports that they tend to be independent from one another [7].

After determining that the ratings were not normally distributed, we opted for the unpaired Wilcoxon rank-sum test to verify whether pairs of samples were distributed significantly differently from one another. In this case, the null hypothesis is that the ratings come from the same distribution, or that the *design* of survey triggers does not bias the resulting sample of affect data that is collected. We reject this, with varying levels of confidence, if the resulting p-values are small. Table 6 summarises the results of these tests. Most notably, 4 of the 6 tests found that the negative affect ratings (and 2 of 6 for the positive ratings) were significantly different from one another with at least 90% confidence. A further test between the Mic_i and Mic_w positive ratings confirmed an alternative hypothesis: that is, that those positive ratings from the Mic_w trigger group were greater than the Mic_i sample ($p = 0.02968$). Uncovering why this result has emerged is beyond the scope of our work: it may be explained by the fact that the Mic_i triggers were likely to be more obtrusive than the Mic_w ones, thus affecting responses. In either case, we observe that our design parameters influence the outcome, and any inferences made on such data should take into account that design will influence the view that researchers will build of their participants.

Positive				
	Screen	**Comms**	**Mic$_i$**	**Mic$_w$**
Screen	1	0.218	0.039**	0.595
Comms		1	0.498	0.378
Mic$_i$			1	0.059*
Mic$_w$				1
Negative				
	Screen	**Comms**	**Mic$_i$**	**Mic$_w$**
Screen	1	0.003**	0.086*	0.991
Comms		1	0.144	0.001**
Mic$_i$			1	0.059*
Mic$_w$				1

Table 6. The p-values from the Wilcoxon rank-sum test results. Stars indicate: * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$.

Differences also emerged from the other questions in the survey: Figure 5 shows how the average *reported* time at home varies between trigger groups. Similarly, participants reported being alone varying amounts, ranging from 33.33% ($Screen$), to 42.31% ($Comms$), 46.67% (Mic_w), to 60.77% (Mic_i). Once again, the choice of sensor-trigger produces a varying set of responses: a researcher who opts for a device-

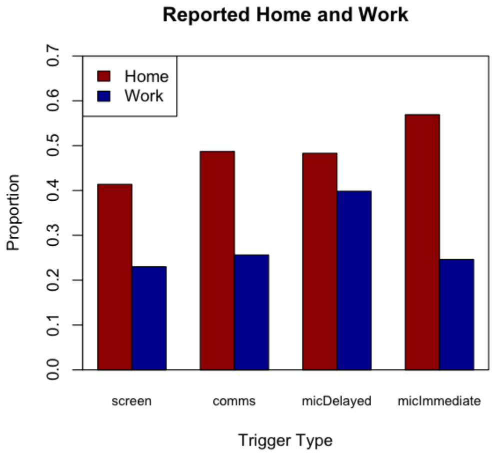


Figure 5. Reported Locations. Demonstrates how the implicit design bias in the sensor-trigger selection influences the reported time that participants are at home.

related trigger will underestimate the extent that participants are co-located with others.

Overall, the results show that the trigger design choice would, beyond influencing the inferences researchers would make about participants' behaviours, also affect the responses that are collected and indeed paint differing pictures of participants' moods, locations, and social settings. These differences may arise due to one or many intersecting reasons. First, as above, trigger design determines the contexts when surveys will be sent. In doing so, they will also inherently change the conditions under which they 'interrupt' participants for responses; doing so may further affect their expectations, engagement, and overall experience of participating in the study (and, consequently, their responses). We leave a full exploration of the relationship between participant's experiences and quantitative outcomes as a question for future research.

DISCUSSION AND RELATED WORK

In this section, we discuss the results above within the frame of relevant *experience sampling* and *mobile sensing* research literature. Broadly speaking, the related work we reviewed confirms that the methodologies historically used to study daily life [16] using sensor-triggered ESM have not tended to investigate the extent that design decisions can alter the data that researchers collect about their participants; we discuss the extent that differing research questions may incur this bias, as well as open research questions to mitigate it where possible.

ESM studies are conducted in order to "capture life as it is lived" [2, 16], by asking participants to answer a set of questions at given moments. Merging ESM and smartphone sensing gives researches an unprecedented opportunity to collect complementary data about how people live and experience their daily lives. Combining these methodologies also promises to tackle challenges that each approach, when used alone, faces. These include, for example, ESM's bias to positive experiences [1], collecting data at a level of granularity and accuracy that goes well beyond people's memory [8], and sensors' lack of subjective feedback [23].

Research Questions and Potential Bias

Naturally, each study is driven by a different research question, and not all experiments may have or need to account for the measurement biases we have uncovered above. Researchers have used the ESM to perform situated studies of ubiquitous technology [3, 9] as well as measure facets of participants' lives; for example, happiness and the environment [15] and emotional reactions to music [12]. Bias in the former scenarios, which may be driven by those precise moments when participants interact with devices, will be determined by the extent that researchers can measure behaviours related to the device. For example, given the growing trend in ESM-based studies towards using participants' own smartphones [11] and the particular emphasis on conserving battery while sensing, not all interactions may be continuously detectable.

Many of the latter studies centre on using the ESM to access data about the participants' lives outside of the domain of ubiquitous technology (e.g., moods or physical activities [5]). Similar to the above, the results for the subset of these studies that collect sensor data will be potentially biased by the mechanism used to collect the data; for example, a study of locations that people frequent based on a random-time trigger will tend to miss locations that are not visited often; as above, location-sensor sampling would skew the representativeness of other sensor data. In fact, many historical studies have been designed around time-based triggers [16], such as random or interval based assessments; the results we present here indicate that, as a consequence of this design, these studies may strongly underrepresent infrequently occurring events (whether these be locations or affective states), and are thus suitable for investigations that explicitly do not seek to measure these cases.

Mitigating Bias: Open Research Questions

Our experiments' results demonstrate that time-based triggers will skew data collection towards those contexts that occur more frequently, while sensor-based triggers (by virtue of being dependent on sensor events occurring) generate a different view of behaviour than more a complete sampling would provide. Moreover, we showed that these design decisions also produced statistically significant differences in the responses we captured from participants. Underlying all sampling strategies is some knowledge (or, at least, assumptions) about the distribution of events that is being sampled from. This leads to two open research questions:

Can multi-sensor sampling be more representative? Our results showed that sampling using a single sensor as a trigger produced skewed results. How can multiple strategies be combined? What role will learning algorithms, that adapt their sampling strategy based on the data collected to date about the participant, play in mitigating this problem? In this work's evaluation, we did not delve into layering sampling strategies together as this may both compound or mitigate sampling bias. Indeed, future ESM studies may not be guided by a single protocol, but be conducted using systems that automatically personalise their behaviour to each participant's behaviour, using machine learning algorithms that make inferences from the sensor and survey response data.

What is the relationship between ESM engagement and responses? Historical studies vary in terms of how often (e.g. up to five [15], seven [12] or nine [19] times a day) and when (i.e., randomly or regularly—say, every 3 hours [2]) participants are signalled to complete questions, or use sensors (e.g., GSM tower changes [10], SMS-events [9]) as triggers for surveys. As we observed, the strategy used to engage with the participant affects the outcome: what is the relationship between engagement, participation, and quantifiable responses?

CONCLUSION

In this work, we examined the extent that design bias influences the response and sensor/behavioural data that researchers can collect from participants in context-aware ESM studies. We demonstrated that different single-sensor sampling strategies will result in contextual dissonance: a disagreement in how much different behaviours are represented in the aggregated sensor data. We based this conclusion on a 1-month, 22-participant, ESM study that solicited survey responses about participants' moods while collecting data from a set of sensors about their behaviour.

The system that we built allowed us to remotely reconfigure the ESM study parameters, which we used to dynamically update triggering conditions during the trial. An overview of the results highlights the temporal bias in sampling, the differences in behaviours that would emerge from varying sampling triggers, and a conflict between different metrics: for example, movement-based triggers would over-represent how often participants use their phone for communication purposes, and device-related triggers would over-represent movement. Similarly, sensor-based triggers produced a higher number of notifications and responses, but this data had a lower average compliance rate, and sensor-based triggers which used the microphone resulted in statistically significant differences in participants' moods.

ACKNOWLEDGMENTS

The authors would like to thank Alastair Beresford for his helpful comments. This research has been funded by the EPSRC Ubiquitous and Social Computing for Positive Behaviour Change project (EP/I032673) and the Gates Cambridge Trust.

REFERENCES

1. Barrett, L. F., and Barrett, D. J. An Introduction to Computerized Experience Sampling in Psychology. *Social Science Computer Review 19*, 2 (2001), 175–185.
2. Bolger, N., Davis, A., and Rafaeli, E. Diary Methods: Capturing Life as it is Lived. *Annu. Rev. Psychology* (2003).
3. Carter, S., Mankoff, J., and Heer, J. Momento: Support for Situated Ubicomp Experimentation. In *ACM CHI* (San Jose, California, 2007).
4. Clark, L., and Watson, D. Mood and the Mundane: Relations Between Daily Life and Self-Reported Mood. *Journal of Personality and Social Psychology 54*, 2 (1988), 296–308.
5. Consolvo, S., McDonald, D., Toscos, T., Chen, M., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., Smith, I., and Landay, J. Acitivity Sensing in the Wild: A Field Trial of Ubifit Garden. In *ACM CHI* (Florence, Italy, 2008).
6. Csikszentmihalyi, M., and LeFevre, J. Optimal Experience in Work and Leisure. *Journal of Personality and Social Psychology 56*, 5 (1989), 815–822.
7. Diener, E., and Emmons, R. The Independence of Positive and Negative Affect. *Journal of Personality and Social Psychology 47* (1984).
8. Eagle, N., and Pentland, A. Reality Mining: Sensing Complex Social Systems. *Personal and Ubiquitous Computing 10* (2006), 255–268.
9. Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., and Landay, J. MyExperience: A System for In situ Tracing and Capturing of User Feedback on Mobile Phones. In *ACM MobiSys* (Puerto Rico, 2007).
10. Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B., and Landay, J. UbiGreen: Investigating a Mobile Tool for Tracking and Supporting Green Transportation Habits. In *ACM CHI* (Boston, USA, 2009).
11. Intille, S., Rondoni, J., Kukla, C., Ancona, I., and Bao, L. Context-Aware Experience Sampling. In *ACM CHI Extended Abstracts* (Ft. Lauderdale, Florida, 2003).
12. Juslin, P., Liljestrom, S., Vastfjall, D., Barradas, G., and Silva, A. An Experience Sampling Study of Emotional Reactions to Music: Listener, Music, and Situation. *Emotion 8*, 5 (2008), 668–683.
13. Killingsworth, M., and Gilbert, D. A Wandering Mind is an Unhappy Mind. *Science 330* (2010).
14. Lathia, N., Rachuri, K., Mascolo, C., and Roussos, G. Open Source Smartphone Libraries for Computational Social Science. In *2nd ACM Workshop on Mobile Systems for Computational Social Science* (Zurich, Switzerland, 2013).
15. Mackerron, G. Happiness and Environmental Quality. *PhD Thesis, The London School of Economics and Political Science* (2012).
16. Mehl, M., and Conner, T., Eds. *Handbook of Research Methods for Studying Daily Life*. The Guildford Press, 2012.
17. Mehl, M., Robbins, M., and Deters, F. Naturalistic Observation of Health-Relevant Social Processes: The Electronically Activated Recorder (EAR) Methodology in Psychosomatics. *Psychosomatic Medicine 74* (2012), 410–417.
18. Miluzzo, E., Lane, N., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S., Zheng, X., and Campbell, A. Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of the CenceMe Application. In *ACM SenSys* (Raleigh, NC, 2008).
19. Nezlek, J. B., Vansteelandt, K., Mechelen, I., and Kuppens, P. Appraisal-Emotion Relationships in Daily Life. *Emotion 8*, 1 (2008), 145–150.
20. Oliver, E. A., and Keshav, S. An Empirical Approach to Smartphone Energy Level Prediction. In *ACM Ubicomp* (Beijing, China, 2011).
21. Patrick, K., Griswold, W., Raab, F., and Intille, S. Health and the Mobile Phone. *American Journal of Preventive Medicine 35* (2008).
22. Rachuri, K., Mascolo, C., Musolesi, M., and Rentfrow, P. SociableSense: Exploring the Trade-Offs of Adaptive Sampling and Computation Offloading for Social Sensing. In *ACM MobiCom* (Las Vegas, USA, 2011).
23. Rachuri, K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., and Aucinas, A. EmotionSense: A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research. In *ACM UbiComp* (Copenhagen, Denmark, 2010).
24. Stone, A., and Shiffman, S. Ecological Momentary Assessment (EMA) in Behavioral Medicine. *Annals of Behavioral Medicine 16*, 3 (1994).