Investigating Self-Reporting Behavior In Long-Term Studies

Andreas Möller [1], **Matthias Kranz** [2], **Barbara Schmid** [1], **Luis Roalter** [1], **Stefan Diewald** [1]
[1] Technische Universität München, Distributed Multimodal Information Processing Group
Munich, Germany
[2] Universität Passau, Lehrstuhl für Eingebettete Systeme, Passau, Germany
andreas.moeller@tum.de, matthias.kranz@uni-passau.de, barbara.elisabeth.schmid@mytum.de,
roalter@tum.de, stefan.diewald@tum.de

ABSTRACT
Self-reporting techniques, such as data logging or a diary, are frequently used in long-term studies, but prone to subjects' forgetfulness and other sources of inaccuracy. We conducted a six-week self-reporting study on smartphone usage in order to investigate the accuracy of self-reported information, and used logged data as ground truth to compare the subjects' reports against. Subjects never recorded more than 70% and, depending on the requested reporting interval, down to less than 40% of actual app usages. They significantly overestimated how long they used apps. While subjects forgot self-reports when no automatic reminders were sent, a high reporting frequency was perceived as uncomfortable and burdensome. Most significantly, self-reporting even changed the actual app usage of users and hence can lead to deceptive measures if a study relies on no other data sources.
With this contribution, we provide empirical quantitative long-term data on the reliability of self-reported data collected with mobile devices. We aim to make researchers aware of the caveats of self-reporting and give recommendations for maximizing the reliability of results when conducting large-scale, long-term app usage studies.

Author Keywords
Self-reporting; survey; long-term study; application usage

ACM Classification Keywords
H.5.m Miscellaneous

General Terms
Experimentation; Measurement; Reliability.

INTRODUCTION AND BACKGROUND
For many research questions in HCI, user studies in the lab are not sufficient. Such controlled experiments can provide initial feedback on a system or investigate an individual, tightly focused usability question, e.g. the comparison of two interaction types or interfaces. By contrast, systems that interact with the environment, social software, location-based services and ubiquitous computing systems must in many cases be evaluated 'in the wild' in order to get an impression of how they work and how users work with them. Researchers are often interested in usage patterns, adaptation processes and learning curves – briefly, in users' behavior with a system in context. Long-term studies, lasting over weeks, months, or even years, are required to answer such questions Researchers have just begun to carry out 'Research in the Large' [14, 24], to be able to deploy and test systems, e.g. for mobile phones, on a significant number of devices and with a large number of users to obtain statistically significant feedback.

Among the numerous established techniques for acquiring information in user studies [21], *data logging*, the *experience sampling method (ESM)* and the *diary* particularly have proven useful for long-term data gathering, as no experimenter needs to be present for those methods.

With *data logging*, a device collects data or context information automatically and without user intervention [12]. This technique can record information which is difficult to gather otherwise efficiently in cost and time [25]. Examples are all sorts of quantitative measures like usage data of applications, or fine-grained context information. Its unobtrusiveness [17] entails high data validity (ideally, subjects do not notice logging at all and consequently do not change their habits). However, researchers cannot apprehend users' intentions through logging alone [12], so that interpretation of the recorded data and the combination with other techniques is often required. We explicitly here leave out privacy issues involved in automatic logging as well as sense making (e.g. machine learning) in the vast desert of automatically collected data sets.

With *experience sampling*, participants actively collect in-situ data upon request [5, 26], that are scheduled randomly, time-based, or triggered by specific events [20, 7]. Data can comprise photos, videos, audio recordings, sensor readings, but also annotations and questionnaires that clarify subjects' thoughts. However, *experience sampling* can be highly interruptive and burdensome if the sampling rate is high and the study is conducted over a long time.

A *diary* diminishes this problem, as it allows users to decide on their own when to record data [5, 26, 23, 21]. As a side effect, researchers implicitly learn about the importance of

CHI 2013, April 27–May 2, 2013, Paris, France.

events to subjects, based on whether they postponed note-taking or not. While diaries can be unstructured or structured [26, 21], they bear the risk that subjects do not remember all events throughout the day if they, e.g., only make one entry every evening, or that they might refrain from writing down certain events, such as, e.g. intimidating information.

The reliability of self-reporting through experience sampling or diaries is influenced by several factors: participants' own perception (one's own behavior is perceived differently than from outside), memory (subjects may not remember correctly, and forget events or actions), sluggishness, enervation (especially when reporting would be interruptive to the current task), and privacy (reporting can be embarrassing or make subjects feel to appear in a bad light, consider e.g. media consumption, sports or food intake habits).

A structured exploration of how reliable self-reported information is under different conditions has not been conducted so far to our knowledge. In our work, we systematically investigate the reliability of self-reporting, in comparison to logged data as ground truth. In a six-week study, participants reported on their usage of Facebook and Mail on the smartphone (we assumed that these applications are used regularly by a majority of smartphone users and thus familiarity with them is given). We chose mobile app usage as a scenario for our investigation, since ground truth data can here easily be obtained. We thereby extend current research trends to conduct research in the large e.g. via digital marketplaces [14, 24], aiming at increasing the user population for extended studies on human-computer interaction (HCI). We add to this quantitative data on the accuracy of self-reported data.

We were thereby interested not only in the overall accuracy of self-reported data (i.e., how well people evaluate their own behavior), but also in time-dependent effects (i.e., how self-reporting behavior changes in the course of the study). We compared different self-reporting intensities with 3 groups, from voluntary reports over daily reminders to regularly presented automated questionnaires, and thereby provide first quantitative results for the reliability of self-reporting.

The paper is structured as follows: We begin with an overview on related work, focusing on self-reporting and logging with smartphones. We then introduce SERENA, our questionnaire framework, and subsequently describe our study in which we compare the accuracy of self-reporting and logging. The results are presented and discussed in detail, before we summarize the lessons learned and give an outlook to future work. As an additional contribution to the community, we offer the software we developed for our experiment, SERENA, as free-to-use framework (URL in the SERENA section of this paper). It consists of a study app and a web service, that support both user data logging and self-reporting through questionnaires. We hope that other researchers find our tools handy and can use them in their own research and studies.

RELATED WORK

Self-reporting has been used in countless long-term studies (see an overview of diary methods [4] and experience sampling [2]), under employment of a variety of techniques. For example, Robinson and Godbey [28] used time diaries to explore how Americans use their work and leisure time. Palen et al. [26] collected phone call data using voice mail diaries. Consolvo et al. [7] used surveys on mobile phones for context-triggered experience sampling. *Mobile Probes* [16] added the possibility to include images to experience sampling records. O'Hara et al. [16] investigated video usage on mobile devices using diaries and ethnographic interviews. Siek et al. [29] had subjects scan the barcodes of products to analyze food intake.

As introduced earlier, self-reporting potentially does not represent the actual behavior of subjects and can be subject to various distorting effects. For example, Lester et al. [22] mention forgetfulness and intentional misreporting as problems. Therefore, self-reporting is often combined with data logging. Eagle et al. [10] investigated the structure of friendship networks with a combination of self-reported data and location and proximity information which was logged by mobile phones. The authors found significant biases in self-reports, e.g. recency effects. Other scenarios where logging has been employed are mobile device interaction [11], and usage patterns [3] or life logging [19]. Although logging potentially provides information with higher accuracy than self-reporting, it can easily be perceived as a privacy threat by subjects [19] or result in a useless amount of log data.

Smartphones have evolved to a convenient tool for self-reporting, as many people use them on a regular basis and have them readily available all the time. A meta analysis conducted by Hufford and Shields [15] revealed that electronic diaries result in a higher compliance with subjects and in better results. Meanwhile, there exists a battery of research applications for collecting user-reported data. MyExperience [12] is an *in situ* data capturing tool for experience sampling. It logs over 140 event types related to the context, the environment and the usage of the phone. Based on previously specified triggers, it is able to make screenshots or to display questionnaires. Through the observation of the user's context, questionnaires can be prevented in inconvenient situations, and correlations between response behavior and certain events can be observed. Momento [6] is a mobile application that allows both study participants and experimenters to gather diary or experience sampling data. Collected information (text, audio, photos or sketches) can be sent over the web, or using SMS/MMS to the desktop counterpart of the system. The mobile client communicates bidirectionally with participants and can receive surveys. However, no simple way to design questionnaires is integrated into the tool. With the ContextPhone platform [27], developers can build context-sensing applications out of pre-existing modules. The tool turned out helpful to researchers for the creation of study applications, e.g. for investigating mobility patterns. The commercial application droid Survey[1] displays questionnaires to participants that can be generated with different templates. mQuest Survey[2]

[1] https://play.google.com/store/apps/details?id=com.contact.droidSURVEY, https://www.droidsurvey.com/

[2] https://play.google.com/store/apps/details?id=de.cluetec.mQuestSurvey, http://www.mquest.eu/

additionally offers the diary study method, photographs and audio recordings. A third survey app is called SurveyToGo[3]. It offers the additional functionality to record videos and to ask 13 different question types. It is offered for Android and Windows Mobile. EpiCollect[4] is a free data collection tool for Android and iOS. It offers gathering data through questionnaires and to view it online or on the phone. The tool offers GPS functionality and four question types. The entries made can be reviewed on the phone. Moreover, a possibility to communicate with participants via Google Talk was added.

SERENA – A SELF-REPORTING APPLICATION

We developed a toolbox supporting data collection in long-term studies that we call SERENA (**SE**lf-**R**eporting and **E**xperie**N**ce sampling **A**ssistant). SERENA combines self-reporting through questionnaires and automated data logging in a singular smartphone application. The Android app can be customized exactly to the experimental needs through remote survey management, adaption of logging and automated data upload to a backend. Each participant just needs to install SERENA on his or her smartphone at the beginning of the experiment, which greatly simplifies setting up studies. While we initially built SERENA for the study described later in this paper, we designed it flexible enough to support a variety of possible study designs and anticipate to use it in future work as well. The SERENA software and a brief tutorial is available at our website at `https://vmi.lmt.ei.tum.de/serena`. By making SERERA available to the community, we encourage also other researchers to use it for their work. We assume that the required consensus for data acquisition is obtained a priori by the researchers using this tool.

Backend

SERENA is highly configurable through a web application (see Figure 2) that fulfills the tasks of preparing the smartphone app for in-field use and of analyzing the results.

Configuration

Prior to the experiment, the experimenter can create sets of questionnaires that will be used to collect information in the study. Each questionnaire consists of a series of pages that can have one of seven types (see Figure 1, left).

Questionnaires are configured to be *voluntary*, *interval-based* or *event-based*. While voluntary questionnaires serve, e.g., for ESM-like methods, interval-based surveys facilitate also the conduction of diary studies. Additionally, the experimenter can assign a group ID to each questionnaire and specify in which timespan it is valid (e.g., a two-week interval). With these features, SERENA supports multiple conditions in within-subjects and between-subjects study designs. A within-subjects study can be realized by multiple questionnaires with different time intervals. For example, participants receive questionnaire A in the first part and questionnaire B in the second part of the study. For a between-subjects study, questionnaires can be assigned different group IDs. Group 1 could then receive questionnaire A, and group 2 could receive questionnaire B. Group IDs also allow running completely independent studies in parallel.

[3]https://play.google.com/store/apps/details?id=dooblo.surveytogo, http://www.dooblo.net/stgi/surveytogo.aspx

[4]http://www.epicollect.net/

Question Type	Usage Description
Single Choice	Singular answer using radio buttons
Multiple Choice	Multiple answers using checkboxes
Drop-Down	Single choice from a large numer of elements using a drop-down menu
Likert	Choice from a number of steps on a labeled Likert scale
Free Text	Single- or multi-line text entry
Range	Integer or float selection using a slider
Text-Only Page	Messages and instructions (e.g. an *intro/thank you* page at the end of the questionnaire)

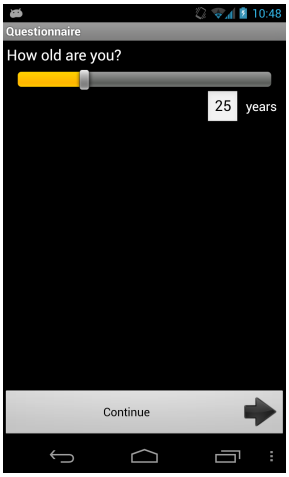

Figure 1. ***Left:*** **Supported question types with the SERENA framework.** ***Right:*** **Example screenshot of the self-reporting app that subjects installed on their smartphone during the long-term study. The survey items have been specified using the backend (see Figure 2).**

Moreover, event logging can be activated. Currently, SERENA records app usage and location information. A comma-separated list of Android app activity and package names can be specified (strings containing wildcards are possible). Thereby, we can adjust logging on sub-application level and limit observations to applications that are interesting for the study, preserving subjects' privacy as much as possible.

The configuration tool generates and exports a model describing the setup and questionnaires, which can either be used to build a ready-to-use app that is distributed to participants, or to remotely configure app instances that are already installed on subjects' devices. We used an extensible XML-based description format to specify SERENA's functionality.

Analysis

Logs and survey data received from subjects' phones are saved to a database from which they can be reviewed in the backend and exported to a spreadsheet for further evaluation.

Mobile Application

From the main menu of the SERENA smartphone application, users can view and answer questionnaires which are currently active. Questionnaires that are added or removed by the experimenter during the study appear or disappear automatically in this list. *Interval*-triggered questionnaires are presented to the user in regular, previously specified intervals (e.g., once a day or every two hours) using a notification that appears on the top of the screen. Upon a click on the message, the questionnaire opens up. With the notification, the user is given the possibility to postpone answering the questionnaire in order to minimize disturbance in an ongoing task. *Event*-triggered questionnaires are shown automatically after a certain event has occurred, e.g. after closing an app. The questionnaire pops up directly afterwards.

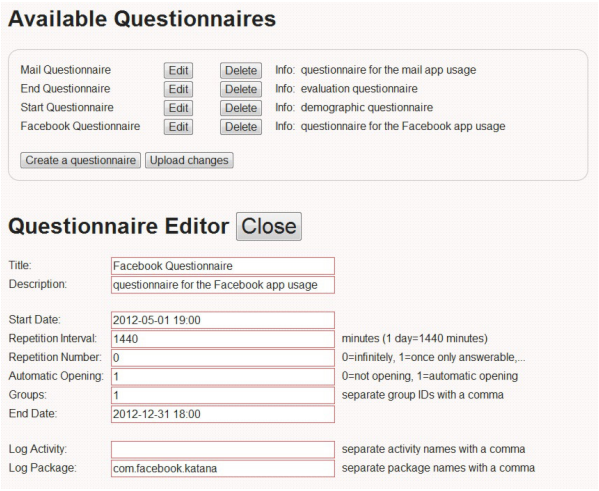

Figure 2. The backend of our self-reporting toolkit where questionnaires can be created and managed. The experimenter can, e.g., specify the interval after which subjects are notified to answer a questionnaire, and specify filters for event logging on the smartphone. Questionnaires can then be sent to the mobile SERENA app (see Figure 1, right).

A background service logs the currently active app package and activity name with timestamp information, according to the previously specified filters by the researcher. Log files are saved to the SD card and are regularly uploaded to the server. In case an application to be logged is not installed on the device, an *Intent* to the application's download page in the Google Play store is sent, so that the user has the possibility to download it. If communication with the server fails, the user is notified to restore connectivity and to upload questionnaires manually using a menu item within the app. All logs are saved locally, so that the researcher can collect the data from participants' phones manually in case automatic upload is not possible at all. For each user, a unique ID is created. This ID (as pseudonym) allows the experimenter to uniquely identify from which device data was received. Questionnaire replies and logged information can thereby be related to individuals without revealing the identity of the person.

Implementation

The backend is built with the Python Pyramid framework[5] and a MySQL database. The web pages for creating questionnaires and analyzing the results are created using jQuery[6] and Ember.js[7]. The mobile SERENA app is implemented in Android, supporting API level 10 (version 2.3.3 and newer). App and server communicate using JSON over HTTP.

Initial Evaluation

Prior to the actual experiment, a three-day pilot study was conducted with five participants in order to test the functionality of SERENA with different smartphone models and the communication with the backend. As a scenario for this pilot, we set the system to log the usage of the Facebook app, and additionally presented a short questionnaire to subjects after they had used Facebook.

In the pilot, we identified and fixed some technical and usability issues in the first version of SERENA. For example, we decided to keep all log files on the device in addition to uploading them to the server, since logs of one subject had arrived incomplete on the server. The *Free Text* and *Range* questionnaire pages were redesigned for better usability.

USER STUDY: SELF-REPORTING BEHAVIOR

In order to investigate the reliability of the data gained through self-reporting, we conducted a six-week user study using the SERENA framework.

Experimental Design

Since smartphone usage can very well be assessed automatically through logging [1, 18, 9, 11, 3], we chose this scenario to gain reliable reference data that we used to compare against self-reported information. While SERENA is not limited to any specific app, we constrained our analysis to two applications, in order to make self-reporting not too excessive. We chose Facebook and Mail, since we assumed that they are frequently used by the majority of smartphone owners.

Self-reports were collected through *Facebook questionnaires* and *Mail questionnaires*, which were filled out according to the study condition. In our study, self-reports were used only as vehicle to assess subjects' reporting behavior, so that we kept the questions simple. Subjects were asked to estimate how long they had just used Facebook or Mail, and how often they had used Facebook or Mail without filling out a questionnaire. Given that a user has started e.g. Facebook three times and answered three questionnaires on those usages, we call those *direct self-reports*. If she only fills out a questionnaire after the third usage, indicating that she has used Facebook three times, the former two app usages are considered to be reported indirectly; we hence call them *indirect self-reports*.

The study consisted of three conditions (*Voluntary*, *Interval* and *Event*) that correspond to different 'intensity levels' of self-reporting.

Voluntary

In the *Voluntary* condition, users were reminded only once prior to the study to report on their application usage. Every time they used either Facebook or Mail, they were instructed to fill out a short questionnaire. However, they were never actively reminded throughout the study (six weeks) to do so.

Interval

In this condition, subjects were, similar to *Voluntary*, not actively reminded to fill out a questionnaire after each application use. However, a reminder notification appeared once a day (scheduled to 7:00 PM for Facebook and to 9:30 AM for Mail). The reminder only showed up if reporting has been missed at least once since the previous reminder.

Event

In this condition, subjects were actively reminded to report on their Facebook and Mail usage right after they had used one

[5]http://www.pylonsproject.org/projects/pyramid/about
[6]http://jquery.com
[7]http://emberjs.com

of these applications. Approximately one second (depending on the phone model) after returning to the home screen, a questionnaire opened automatically, why we call this the *event-based* condition. In some cases, no questionnaire appeared. This was the case when a user did not quit the application using the home button, but started another application using the application switcher, when another application was started immediately after quitting the previous one (before the one-second grace time), or when another application was started automatically from Facebook or Mail using an Android Intent (e.g., the camera to upload a photo).

Task and Procedure

Participants were asked to install SERENA on their personal smartphone and to use the phone as usual during the study period. Participants could come into our lab for installation assistance. All other communication throughout the study was conducted via email. Prior to the study, participants filled out a short questionnaire regarding their own smartphone usage. Another questionnaire was filled out at the end of the study. All questionnaires were sent to the experimenters through SERENA and could be matched to individual participants only through a unique ID. In the course of the six-week study period, a reminder email was sent every two weeks, thanking subjects for their participation so far and indicating how long the study would still last.

At the end of the study, participants could choose a small gift for compensation (a sweets package or a cinema tickets voucher). We deliberately decided for a modest compensation in order not to influence results through the incentive.

Hypotheses

Hypotheses 1 and 2 address the reliability of self-reporting compared to the actual behavior of subjects, with relation to different self-reporting conditions. We thereby look at the self-reporting ratio, i.e., the ratio of reported and actually occurred app usages, and the estimated app usage durations of subjects in relation to actual durations.

(H1a) Self-reporting ratios are higher in the *Event-based* than in the *Interval-based* condition.
(H1b) Self-reporting ratios are higher in the *Interval-based* than in the *Voluntary* condition.
(H2) Estimated app usage durations are higher than actual usage durations.

Furthermore, there might be an effect with relation to time. There are different possible patterns, such as an increase or decrease of self-reporting rates, or a decrease at the beginning and a rise towards the end ('U-shape'). Reporting could also influence actual app usage, i.e., subjects could use apps less because they have to answer a questionnaire afterwards. This leads us to the following hypotheses.

(H3) Self-reporting ratios decrease in the course of the study.
(H4) Actual app usage remains constant in the course of the study.

Finally, we address users' perception and acceptance of the self-reporting process with the following hypotheses:

(H5a) Subjects perceive *Event* more effortful than *Interval*.
(H5b) Subjects perceive *Interval* more effortful than *Voluntary*.

Participants

We required that participants own an Android smartphone and use email and the Facebook application with it. 30 subjects between 18 and 32 years (average age: 25, standard deviation = 2.8), most of them students, participated in the study. 8 were females, 22 were males. Participants were randomly assigned to one of the three conditions (*Voluntary*, *Event*, *Interval*), so that n = 10 for each condition (between-subjects design).

Most subjects stated to use Facebook and Mail one or several times a day (Facebook usage per day: 21 several times, 6 once, 3 fewer. Mail usage per day: 25 several times, 2 once, 3 fewer).

Measurements

In all three conditions, SERENA logged participants' actual app usage, filtered by the package names of the official Android Facebook application (com.facebook.katana) and the name of the mail application the subject used. Since a variety of different mail clients exist on Android (e.g., the built-in Android mail clients, manufacturer-specific mail clients, K9 Mail, ...) we asked prior to the study which application(s) subjects use in order to include them to the logging whitelist. If a participant used multiple mail clients on his or her device (e.g., for Googlemail and Exchange mail), both were aggregated and referred to as Mail usage.

In case subjects switched back and forth between applications within a short period of time, we aggregated subsequent usages of the same app and counted them as singular app usage, summing up individual usage times. We assumed a single task if users returned to the original app within 60 seconds. This was the case, e.g., when subjects were composing an email, looked something up in another app, and switched back to finish the email. Often, those other applications were launched programmatically using an *Intent*, e.g. for choosing an email attachment or sharing an image in Facebook.

RESULTS

Number of App Usages

In total, 3,631 Mail usages and 3,181 Facebook usages were logged during the study. For the following statistics, we do not average over participants, but look at individual usages. Figure 3 illustrates the ratios of self-reported app usages in relation to the logged usages. The bottom, darker-colored portions of the columns in Figure 3 represent direct reports (i.e., filled out questionnaires). The top, light-colored portions illustrate indirect reports, so that the columns in total represent the amount of all reported usages.

Facebook usages were in total reported at 37.6% in *Voluntary*, at 63.8% in *Interval* and at 54.3% in *Event*. Mail usages were reported at 54.6% in *Voluntary*, at 68.4% in *Interval* and at 53.9% in *Event*. As previously described, not every application usage in the *Event* condition entailed a questionnaire notification. Only when subjects returned to the home

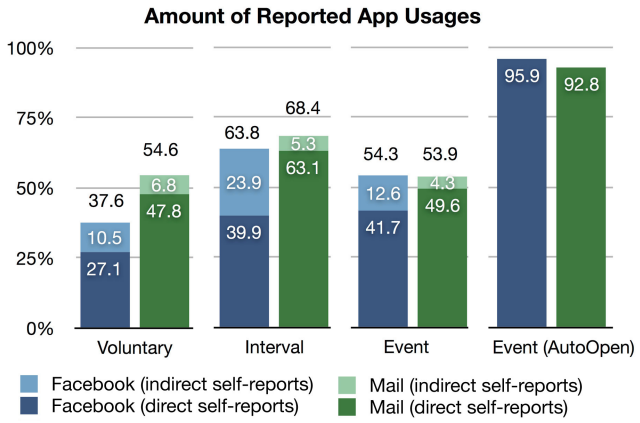

Figure 3. Ratio of self-reported and measured app usage in the conditions *Voluntary*, *Interval* and *Event*. *AutoOpen* denotes the amount of filled out questionnaires that have been opened automatically in the *Event* condition. Direct self-reports represent the number of filled out questionnaires, while direct and indirect self-reports also comprise app usages that have been caught up in a subsequent questionnaire.

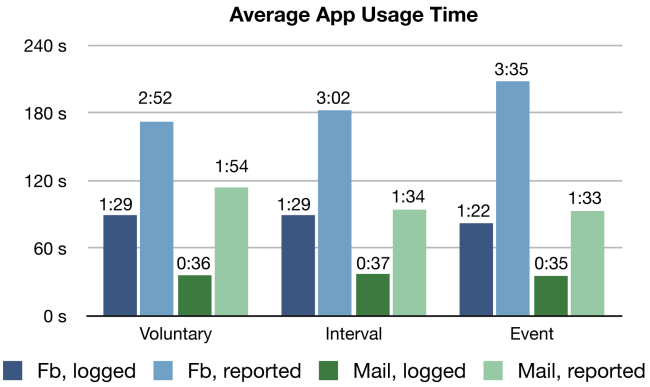

Figure 4. Self-reported and logged usage times of Facebook (Fb) and Mail in the different conditions. Participants overestimated the durations of their actual app usage sessions often by more than 100%.

screen from the application, a questionnaire notification was shown. This was the case in 1,224 of 2,488 app usages (49%). Of those *AutoOpen* questionnaires, 95.9% were answered for Facebook and 92.8% for Mail.

The portion of indirect self-reports was higher for Facebook than for Mail, and it was particularly high in the *Interval* condition. Presumably, Facebook is more often used at a more unconscious level and has so much passed into subjects' habits and naturally integrated in their phone interaction that subjects did not remember filling out a questionnaire right afterwards. This also explains why particularly in *Interval*, where only one reminder a day was sent, so many Facebook reports were forgotten.

A Student's t-test (α = 0.05, two-tail) showed no significant differences between the reported ratios in *Voluntary*, *Interval*, and *Event*. The differences between *AutoOpen* and the other conditions were significant (P(T $\leq$ t) = 0.0045 for *Interval–AutoOpen*, P(T $\leq$ t) = 0.00034 for *Voluntary–AutoOpen*, P(T $\leq$ t) = 0.00001 for *Event–AutoOpen*). Although the number of self-reporting reminders rose from *Voluntary* over *Interval* to *Event*, the number of actually reported events did not. In *Interval*, subjects reported more than in *Voluntary*, which we hypothesized in Hypothesis 1 (**H1b**). Interestingly, they reported less in *Event* than in *Interval*, which does not confirm Hypothesis **H1a**. However, Hypothesis **H1a** could be accepted if we look only at the report rate where the questionnaire had appeared automatically.

App Usage Duration

Figure 4 summarizes logged and reported Facebook and Mail usage durations. Facebook has generally been used at least twice as long as Mail. In *Voluntary*, Facebook has been used averagely for 1:29 minutes and Mail for 33 seconds. In *Interval*, Facebook has been used averagely for 1:29 minutes at a time, Mail only for 37 seconds. The average usage times in *Event* were 1:22 minutes for Facebook and 35 seconds for Mail.

Subjects overestimated the usage duration in their self-reports in all conditions. Subjects estimated their Facebook usage in *Voluntary* at 2:52 minutes (92% more), in *Interval* with 3:02 minutes (104% more) and in *Event* with 3:35 minutes (153% more). Mail usage durations were overestimated even further. Subjects reported usage times of 1:54 minutes in *Voluntary* (217% more than the actual ones), 1:34 minutes in *Interval* (153% more), and 1:33 minutes in *Event* (163% more).

A t-test showed significant differences between reported and logged time for Facebook in *Voluntary* (P(T $\leq$ t) = 0.0071, α = 0.05, single-tail), in *Interval* (P(T $\leq$ t) = 0.020, α = 0.05, single-tail), but not in *Event* (P(T $\leq$ t) = 0.056, α = 0.05, single-tail). For Mail, reported and logged durations were significantly different in all conditions. P(T $\leq$ t) = 0.0004, α = 0.05, single-tail (*Voluntary*), P(T $\leq$ t) = 0.0013, α = 0.05, single-tail (*Interval*), P(T $\leq$ t) = 0.0044, α = 0.05, single-tail (*Event*).

We hypothesized that subjects overestimate their actual app usage when reporting on their behavior, which was already suggested by previous findings [8, 13]. In fact, subjects overestimated app usage durations mostly by more than 100%, so that Hypothesis **H2** is likely to be correct.

Self-reporting over Time

Figure 5 illustrates the self-reporting behavior in the course of the study. The diagrams illustrate the direct self-report ratios (blue, squared graph) and direct plus indirect self-report ratios (green, circular graph), aggregated for each week.

The general trend of the self-reporting ratio is decreasing in all conditions, except for Mail usage reports in the *Interval* condition. We have no explanation for this exception, compared to the other conditions. In the first week, subjects filled out a questionnaire at between 38.6% (Facebook, *Voluntary*) and at 61.4% (Mail, *Interval*) of all usages. Considering also indirect reports, the highest reporting rate is 78.1% (Facebook, *Event*). In the second week, self-reporting rates decrease in average by 9.4% and remain almost constant between week 2 and 3 ($-$1.0%). Beginning from week 4, subjects reported slightly more again (+6.1% to week 3), but reduce reporting again towards the end of the study ($-$7.7% be-

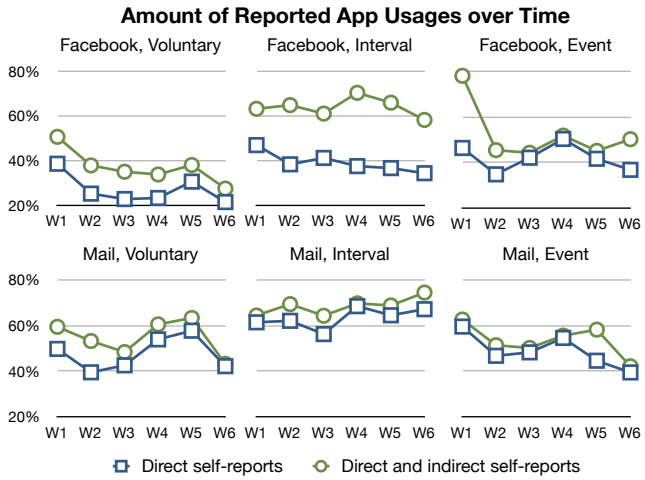

Figure 5. Amount of self-reported and logged application usage for each week in the conditions *Voluntary*, *Interval* and *Event*. Direct self-reports represent the number of filled out questionnaires, while direct and indirect self-reports also comprise app usages that have been caught up in a subsequent questionnaire.

Logged App Usage

		Overall	W1		W2		W3		W4		W5		W6
Voluntary	Fb	**1190**	241	-1.3 %	238	-32.2 %	180	21.1 %	228	-88.4 %	121	33.5 %	182
	Mail	**1018**	227	-83.1 %	124	20.0 %	155	-2.0 %	152	5.6 %	161	19.1 %	199
Interval	Fb	**934**	234	-55.0 %	151	-19.8 %	126	-0.8 %	125	15.0 %	147	2.6 %	151
	Mail	**1184**	236	-7.8 %	219	-10.1 %	199	-30.9 %	152	18.3 %	186	3.1 %	192
Event	Fb	**1057**	237	-19.7 %	198	-41.4 %	140	-2.2 %	137	4.9 %	144	28.4 %	201
	Mail	**1429**	324	-50.0 %	216	16.0 %	257	-31.1 %	196	8.0 %	213	4.5 %	223

Figure 6. Logged application usage (number of Facebook, in the table abbreviated as Fb, and Mail sessions) for each week in the conditions *Voluntary*, *Interval* and *Event*. Decreases between weeks are marked red, increases are marked green.

tween week 4 and 6). In the last week, reporting rates range between 21.4% (Facebook, *Voluntary*) and 67.2% (Mail, *Interval*) for direct self-reports, and 74.5% (Mail, *Interval*) when indirect self-reports are considered as well.

As already outlined in Figure 3, reporting rates start and remain lowest in the *Voluntary* condition. They begin at a comparable level in *Interval* and *Event*, but rates decrease more in the *Event* condition that in *Interval*.

The tendencies in the results coincide with Hypothesis **H3**, where we hypothesized that self-reports decrease in the course of the study.

App Usage Over Time

Figure 6 summarizes the actual number of Facebook and Mail sessions determined by logging. Between week 1 and 2, usages decreased in all conditions, partly by more than 50%. This trend partly continues until week 4; however, towards the end of the study, app usages rise again in all conditions.

In fact, subjects stated that their behavior was influenced by self-reporting. This effect was stronger in *Interval* than in *Voluntary*, and stronger in *Event* than in *Interval*. At the end of the study, participants were asked for statements on the self-reporting mode they used (*Voluntary*, *Interval*, *Event*), whether and how it had changed their application usage, and they were invited to give general feedback. Users' statements were translated to English.

Voluntary

Subjects mostly liked the way of voluntary self-reporting. P3 found it *"fast and playful"* and liked that it is *"low effort and can be filled out any time"*. P6 said that *"it's simple and is actually uncomplicated. There's not much interface necessary."* Some subjects would have preferred some kind of automation, e.g. P5: *"I'd have preferred to be asked automatically, once or several times per day, to fill out a survey. Cause I was in a hurry or I forgot it fairly often, I didn't fill out the questionnaire each time. A daily notification would have been sufficient for me."* For five subjects, usage habits did not change. However, five reported to use apps shorter or less frequently. P1 stated to *"use apps more consciously, only when I really wanted to use them and not just started them because there was nothing to do"*. According to P10, the effect was *"no endless surfing any more and reduced usage"*.

Interval

Similar to *Voluntary*, most participants in this group stated to like this way of reporting. One subject (P15) even would have favored *Event*, stating that *"questionnaires should pop up by default after the app"*. Some participants found the effect of self-reporting interesting. *"[I was] more aware in estimating my usage time"* (P19). P13 said that it was *"interesting to yourself how often you open the apps"* and found the effort *"acceptable"*. P17 admitted to have become sloppy with the time, but did not perceive questionnaires annoying. However, six of ten participants in this group stated to have changed their behavior, e.g. checked emails less frequently (P13) and used the apps generally less so that they didn't have to answer questionnaires (P14, P20). P12 stated to *"not have looked at every single mail, and moved Facebook usage to the PC"*.

Event

Comments from participants in this condition were rather critical. P23 said: *"Too time-intensive and complicated. Periodically answering the same questions over and over again is annoying. Sometimes, you even avoid using those apps."* Another user suggested to use less clicks in the questionnaire to make self-reporting more convenient. P26 did not like event-based reporting because it was *"...too annoying"* and he would *"prefer background logging"*. Nine out of ten subjects also indicated that they used Facebook and Mail differently or at least thought about it. P21 said: *"I partly looked up e-mails at my PC when I was too lazy to fill out the questionnaire on the mobile"*. Other participants reported to have used the apps significantly less, especially towards the end of the study (P22, P24, P25). P23 stated to *"often have read only the notification but not started the app any more"*. Unlike most other participants in the *Event* condition, P28 was happy about the increased awareness of app usage: *"I now know how much time I've wasted with that! I should waste my time with other things."*

Commitment

Subjects were also asked to judge their commitment to self-report (see Figure 7). In the first two weeks, 0% and 3% agreed that they have not regularly filled out questionnaires.

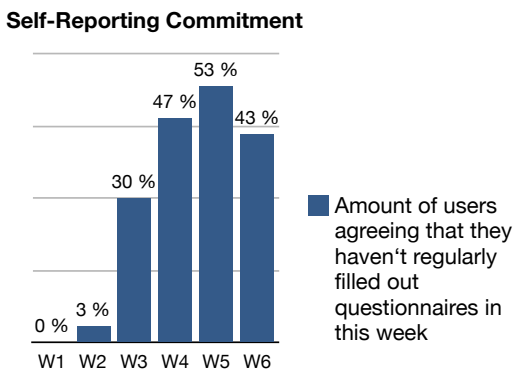

Figure 7. User commitment to answer questionnaires after each usage of Facebook or Mail on the smartphone.

By week 3, the amount rose to 30%, by week 4 to 47%, and in week 5, more than half of participants (53%) admitted to not regularly fill out questionnaires any more. In the last week, the estimation was slightly lower again; 43% of subjects stated to have missed questionnaires. While the rise of self-reports in week 6 can not be confirmed by the measured data, the estimation matches the decrease in the measured self-report ratio at the beginning of the study.

It is normal that app usages vary over time, depending on diverse factors (e.g., people being on holiday). However, in light of subjects' statements to actually have used applications less because of the study, there is evidence that self-reports can actually have influenced subjects' behavior. Hypothesis **H4** is therefore likely to be rejected.

User Satisfaction

At the end of the study, subjects answered a final questionnaire in which they indicated how satisfied they were with self-reporting during the study. Questions were answered on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The results are summarized in Figure 8.

Self-reporting was perceived less burdening in *Voluntary* than in *Interval* and *Event*. Subjects responded to the statement "Answering the questionnaire was low effort" in *Voluntary* with averagely 4.1 (both for Facebook and Mail), in *Interval* with 3.8 (both for Facebook and Mail), and in *Event* with 3.7 (Facebook) and 3.4 (Mail).

Subjects responded above average that they always filled out the questionnaire after using Facebook or Mail. In *Voluntary*, their average agreement to this statement was 3.6 (Facebook) and 4.1 (Mail); in *Interval*, it was 3.7 (Facebook) and 3.5 (Mail); in *Event*, it was 4.2 (Facebook and Mail).

The effect on application usage habits was rated differently depending on conditions. In *Voluntary*, subjects agreed that answering questionnaires changed usage habits below average with 2.3 (Facebook and Mail). In *Interval*, the agreement level was 2.2 (Facebook) and 2.1 (Mail). In *Event*, the estimation that reporting changed usage habits was highest: average agreement was 3.5 for Facebook and 2.9 for Mail. Even though those differences are not significant, participants' statements and logged app usage times indicate that

Figure 8. Qualitative user feedback at the end of the study in the different conditions. Answers were given on a Likert scale, ranging from 1 = strongly disagree, to 5 = strongly agree.

there is actually an influence of self-reporting to actual behavior.

Regarding Hypothesis **H5**, we can state that the effort to answer questionnaires was actually perceived higher in *Event* than in *Interval*, and in *Interval* higher than in *Voluntary*, as we expected. However, the differences are smaller as anticipated, which matches also subjects' feedback that they felt reports burdening already in *Interval* or even *Voluntary*.

DISCUSSION AND LESSONS LEARNED

Our study showed that self-reports on application usage can generally not be considered as accurate. Depending on the condition, only approximately 40% to 70% of actual app usages were reported by subjects within six weeks. Subjects responded to more than 90% of all questionnaires that automatically popped up in the *Event* condition, but this frequency of automated questionnaires would most likely not be feasible in a long-term study, since it would annoy users too much. Already in our setup, where only in 49% of app usages in the *Event* condition a questionnaire appeared, subjects found reporting too time-intensive and annoying. While some participants in *Interval* wished for automated surveys to prevent that they forget to report (which is exactly what we provided in the *Event* condition), subjects in *Event* felt overly burdened and wished for a larger interval between reports, or for logging. Thus, there are probably two explanations for the rather low reporting rates in all conditions: in *Voluntary*, subjects forgot about reporting; in *Event*, they deliberately did not answer questionnaires because it was too much effort to them.

Based on our findings, we summarize some lessons learned, which can guide further usage of self-reporting in user studies and outline directions for future work.

Awareness for Inaccuracy

Researchers must not blindly trust self-reported data, but take into account that this data can be unreliable. Self-reports are a valuable data collection method in long-term studies and should be employed, but researchers should take into account that a corrective factor might be necessary when analyzing and interpreting the results. For example, estimated usage times in our study have shown that subjects overestimate the duration of app sessions, and these results stand in line with earlier research [8, 13].

Multiple Data Sources

Often, studies strive for both quantitative and qualitative data. For smartphone usage, logging can capture a variety of usage information in an unobtrusive way. This is, however, not possible in all scenarios, for which self-reporting can then be an option to obtain data. In order to assess the accuracy of self-reports as a qualitative method, we chose app usage as a criterion that can be compared to quantitative logs as ground truth. Diary reports might not reflect entirely reliable usage frequencies, but this does not make them less reliable for the assessment of the experiences recorded. Self-reports have their unique advantage for gaining additional insights which cannot be obtained in an automated way.

We showed that self-reporting can even influence actual behavior, in our case the usage frequency of the observed apps (and it is critical when the research method influences the observation!). If possible, researchers should therefore consider a combination of self-reports and logged data to achieve additional certainty. As of today, where self-reports are often recorded with smartphones, automated collection and logging of information is in many cases a small extra effort.

Not Overcharging Participants

Do not use an overly high self-reporting intensity or interval. In our study, report rates already started below 70% and decreased from the second week on, why dense self-reports from participants are hard to justify. If the burden is too high, participants will get annoyed so that they refrain from reporting, or more severely, the alter their actual behavior. Subjects stated to have reduced the usage of applications in order to reduce the logging effort. Further analyses are necessary how self-reporting can be designed to be convenient from the beginning in order to keep users engaged. Our results show that reporting rates in *Interval* were in average higher than in the more demanding *Event* condition, suggesting that less 'pressure' can lead to even more satisfying results. Data collected by SERENA could help here in future research.

Make Use of Reminders

The reporting rate over six weeks tended to decrease in general. However, it is also notable that a relative increase in the second half of the study could be observed, which might correspond to the reminder emails sent after week 2 and 4, and with a guilty conscience of participants who had neglected reporting and now had a stronger sense of duty towards the end. Subsequent work could systematically investigate how reminders can be adapted (regarding time and frequency) to cause significant effects.

Adapt Method to Scenario

Reliability also differs according to the scenario. In our study, in particular many Facebook usages were missed by subjects' self-reports, which might be due to the more unconscious nature of mobile Facebook usage as of today. Some subjects confirmed that they were not aware of their usage frequency before, and that only through self-reporting they initially became conscious of how often they log in to Facebook.

The data collection method should thus carefully be adapted to the scenario and the actual data to be gathered. If, e.g., just random experiences or impressions should be collected, it can even be preferential when users do not report too often, because researchers can then learn which moments are salient to subjects. However, when quantitative data or "instances" should be captured, self-logging can provide unreliable and incomplete data.

CONCLUSION

We have compared the reliability of self-reporting methods regarding smartphone app usage with different reporting intervals (*Voluntary*, *Interval*, *Event*), using log data as ground truth. To our knowledge, we provide the first quantitative analysis of the accuracy of self-reporting with mobile phones in long-term studies. Our six-week experiment showed that self-reports do not provide a complete image of actual application usage and that subjects misestimate durations of application sessions. The self-reporting interval thereby plays a less important role than hypothesized. A lower required report rate can, on the contrary, even lead to better results (e.g., subjects reported more frequently with interval-based reminders than with event-based questionnaires). Results also strongly differ on the task; Facebook usage was e.g. underestimated more than Mail usage.

Nonetheless, self-reporting is an important data collection technique, in particular in scenarios when no automated logging can be employed. It is a substantial method for recording subjective and qualitative experiences with an application. The selective nature of self-reports also helps to identify what is important to users. Researchers should, however, be aware of the potential inaccuracy, dependent on the scenario, when using self-reports. Future work thus should deeper investigate dependencies between reporting conditions and task types, in order to maximize reporting reliability. Since inconvenient self-reporting modalities can influence the behavior that is to be logged, participants' satisfaction is crucial to successful self-reporting.

REFERENCES

1. Barkhuus, L., and Polichar, V. E. Empowerment through Seamfulness: Smart Phones in Everyday Life. *Personal and Ubiquitous Computing 15*, 6 (2010), 629–639.
2. Barrett, L. F., and Barrett, D. J. An Introduction to Computerized Experience Sampling in Psychology. *Social Science Computer Review 19*, 2 (2001), 175–185.
3. Böhmer, M., Hecht, B., Schöning, J., Krüger, A., and Bauer, G. Falling asleep with Angry Birds, Facebook and Kindle: A Large Scale Study on Mobile Application

Usage. In *Proc. of the 13th Intl. Conference on Human Computer Interaction with Mobile Devices and Services*, ACM (Stockholm, Sweden, 2011), 47–56.

4. Bolger, N., Davis, A., and Rafaeli, E. Diary Methods: Capturing Life as it is Lived. *Annual Review of Psychology 54* (2003), 579–616.
5. Carter, S., and Mankoff, J. When Participants do the Capturing: the Role of Media in Diary Studies. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (Portland, OR, USA, 2005), 899–908.
6. Carter, S., Mankoff, J., and Heer, J. Momento: Support for Situated Ubicomp Experimentation. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (San Jose, CA, USA, 2007), 125–134.
7. Consolvo, S., Harrison, B., Smith, I., Chen, M., Everitt, K., Froehlich, J., and Landay, J. A. Conducting In Situ Evaluations for and with Ubiquitous Computing Technologies. *Intl. Journal of Human-Computer Interaction 22*, 1–2 (2007), 103–118.
8. Deane, F. P., Podd, J., and Henderson, R. D. Relationship between Self-Report and Log Data Estimates of Information System Usage. *Computers in Human Behavior 14*, 4 (1998), 621–636.
9. Demumieux, R., and Losquin, P. Gather Customer's Real Usage on Mobile Phones. In *Proc. of the 7th Intl. Conference on Human Computer Interaction with Mobile Devices and Services*, ACM (Salzburg, Austria, 2005), 267–270.
10. Eagle, N., Pentland, A. S., and Lazer, D. Inferring Friendship Network Structure by using Mobile Phone Data. *Proc. of the National Academy of Sciences 106*, 36 (2009), 15274–15278.
11. Falaki, H., Mahajan, R., Kandula, S., Lymberopoulos, D., Govindan, R., and Estrin, D. Diversity in Smartphone Usage. In *Proc. of the 8th Intl. Conference on Mobile Systems, Applications, and Services*, ACM (New York, NY, USA, 2010), 179–194.
12. Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., and Landay, J. A. MyExperience: A System For In Situ Tracing and Capturing of User Feedback on Mobile Phones. In *Proc. of the 5th Intl. Conference on Mobile Systems, Applications and Services*, ACM (San Juan, Puerto Rico, 2007), 57–70.
13. Hartley, C., Brecht, M., Pagerey, P., Weeks, G., Chapanis, A., and Hoecker, D. Subjective Time Estimates of Work Tasks by Office Workers. *Journal of Occupational Psychology 50*, 1 (2011), 23–36.
14. Henze, N., Rukzio, E., and Boll, S. Observational and Experimental Investigation of Typing Behaviour Using Virtual Keyboards for Mobile Devices. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (Austin, TX, USA, 2012), 2659–2668.
15. Hufford, M. R., and Shields, A. L. Electronic Diaries: Applications and What Works in the Field. *Applied Clinical Trials* (2002), 46–59.
16. Hulkko, S., Mattelmäki, T., Virtanen, K., and Keinonen, T. Mobile Probes. In *Proc. of the 3rd Nordic Conference on Human-Computer Interaction*, ACM (Tampere, Finland, 2004), 43–51.
17. Jansen, B. J., Taksa, I., and Spink, A. *Handbook of Research on Web Log Analysis*. IGI Global, 2009.
18. Jeon, M. H., Na, D. Y., Ahn, J. H., and Hong, J. Y. User segmentation & UI Optimization Through Mobile Phone Log Analysis. In *Proc. of the 10th Intl. Conference on Human Computer Interaction with Mobile Devices and Services*, ACM Press (Amsterdam, The Netherlands, 2008), 495–496.
19. Kärkkäinen, T., Vaittinen, T., and Väänänen-Vainio-Mattila, K. I Don't Mind Being Logged, But Want to Remain in Control: A Field Study of Mobile Activity and Context Logging. In *Proc. of the 28th Intl. Conference on Human Factors in Computing Systems*, ACM (Atlanta, GA, USA, 2010), 163–172.
20. Krumm, J. *Ubiquitous Computing Fundamentals*. Chapman & Hall / CRC, 2009.
21. Lazar, J., Feng, J. H., and Hochheiser, H. *Research Methods in Human-Computer Interaction*. John Wiley & Sons Ltd, Chichester, 2010.
22. Lester, J., Choudhury, T., and Borriello, G. A Practical Approach to Recognizing Physical Activities. *Pervasive Computing* (2006), 1–16.
23. Mankoff, J., and Carter, S. Crossing Qualitative and Quantitative Evaluation in the Domain of Ubiquitous Computing. In *CHI2005 Workshop "Usage analysis: Combining Logging and Qualitative Methods"* (2005).
24. Möller, A., Michahelles, F., Diewald, S., Roalter, L., and Kranz, M. Update Behavior in App Markets and Security Implications: A Case Study in Google Play. In *Proc. of the 3rd Intl. Workshop on Research in the Large. Held in Conjunction with Mobile HCI* (2012), 3–6.
25. Nielsen, J. *Usability Engineering*. Morgan Kaufmann, 1993.
26. Palen, L., and Salzman, M. Voice-Mail Diary Studies for Naturalistic Data Capture Under Mobile Conditions. In *Proc. of the 2002 ACM Conference on Computer Supported Cooperative Work*, ACM (New Orleans, LA, USA, 2002), 87–95.
27. Raento, M., Oulasvirta, A., Petit, R., and Toivonen, H. ContextPhone: A Prototyping Platform for Context-Aware Mobile Applications. *Pervasive Computing, IEEE 4*, 2 (2005), 51–59.
28. Robinson, J. P., and Godbey, G. *Time For Life: The Surprising Ways Americans Use Their Time*. Penn State University Press, 1997.
29. Siek, K. A., Connelly, K. H., Rogers, Y., Rohwer, P., Lambert, D., and Welch, J. L. When Do We Eat? An Evaluation of Food Items Input Into an Electronic Food Monitoring Application. In *Pervasive Health Conference and Workshops, 2006*, IEEE (2006), 1–10.