On Applying Experience Sampling Method to A/B Testing of Mobile Applications: A Case Study

Myunghee Lee and Gerard J. Kim(✉)

Digital Experience Laboratory, Korea University, Seoul, Korea
{revise1,gjkim}@korea.ac.kr

Abstract. With the advent of mobile devices, the experience sampling method (ESM) is increasingly used as a convenient and effective way to capture user behaviors of, and evaluate mobile and environment-context dependent applications. Like any field based in situ testing methods, ESM is prone to biases from unreliable and unbalanced data, especially for A/B testing situations. Mitigating such effects can in turn incur significant costs in terms of the number of participants and sessions, and prolonged experimental time. In fact, ESM has rarely been applied to A/B testing nor do existing literatures reveal its operational details and difficulties. In this paper, as a step toward establishing concrete guidelines, we describe a case study of applying ESM to evaluating two competing interfaces for a mobile application. Based on the gathered data and direct interviews with the participants, we highlight the difficulties experienced and lessons learned. In addition, we make a proposal for a new ESM in which the experimental parameters are dynamically reconfigured based on the intermediate experimental results to overcome the aforementioned difficulties.

Keywords: Experience sampling method (ESM) · A/B testing · Usability

1 Introduction

Experience (or environment) sampling method (ESM) is a system evaluation and behavior capture method by which user evaluative responses are made and recorded at the exact time and place of the system usage. Compared to the old paper-and-pencil method, with the advent of mobile devices, ESM can be carried out more conveniently e.g. as or through functionalities embedded in smart phone sensors and applications. Like any field-based in situ testing methods, ESM can suffer from biases that might otherwise be controllable in a laboratory setting, but at the same time, they can be mitigated through a high number of repetitions, extended length of experimentation, a large number of participants, and thus at higher cost. However, this can also ironincally bring about even more unreliable and unbalanced data. This is more problematic in the case of a comparative evaluation in which, for a validity and fairness, it is necessary to collect a minimum and "balanced" amount of reliable data. In fact, ESM has rarely been applied to A/B testing nor do existing literatures reveal its operational details and difficulties. In

J. Abascal et al. (Eds.): INTERACT 2015, Part II, LNCS 9297, pp. 203–210, 2015.
DOI: 10.1007/978-3-319-22668-2_17

this paper, as a step toward establishing concrete guidelines for A/B testing with ESM, we describe a case study of applying ESM to evaluating two competing interfaces for a mobile application. Based on the gathered data and direct interviews with the participants, we highlight the difficulties experienced and lessons learned. In addition we make a proposal for a new ESM in which the experimental parameters are dynamically reconfigured based on the intermediate experimental results to overcome the aforementioned difficulties and run the experiment more economically.

2 Related Work

The Experience Sampling Method (ESM) was first introduced by Larson and Csikszentmihalyi as a research tool in social science [1], but has found great utility especially in mobile HCI research [2]. For instance, the Context Aware Experience Sampling Tool developed by Intille et al. [3], one of the first of its kind, allowed a flexible data solicitation by scripting in the survey questions and multiple choice answers, and including a functionality for users to easily capture and store multimedia data on a PDA. Consolvo et al. used a similarly designed ESM tool, called the iESP, to evaluate ubiquitous computing applications and further analyzed the possible pitfalls and lessons learned of using such a methodology for in situ HCI evaluation (e.g. the effectiveness of self-reporting using the mobile devices and the need for tailoring the data collection process for the target subjects) [4]. In 2007, Froehlich et al. also introduced a more advanced mobile based ESM tool called MyExpereince which offered an XML based specification method of how to solicit data from the user, sensor based context triggered sampling and structured storage of logged data to a data base server [5]. Momento, a tool developed by Carter et al. is another step in the evolution of the mobile based ESM tools offereing sampling control and on-line monitoring (e.g. visualization and analysis of incoming data) from a remote desktop server [6]. Maestro further extended the sampling control capability by exploiting long term user behavior and usage patterns for shaping personalized ESM questions to different types of users [7]. As ESM tools become more refined and enter into one of the main stream evaluation methods, its user interface/interaction design itself has emerged as an important issue as well with regards to the requirement and desire to encourage the participants to make faithful and reliable response [8]. While not usually employed in ESM, the data collection in crowdsourcing can involve "gold standard" questions to ensure the reliability and credibility of the contributors [9]. Answering performanes to the gold standard questions can be used to exclude certain data, e.g. those that are regarded too mechanical or even those of programmed bots.

In summary, it can be seen that ESM tools are continuing to evolve and being added with more functionalities (for both the participants and experiment administrators) and the methodology extended for a more reliable and credible results. In most previous related work we have reviewed, ESM was still used for capturing context dependent user behaviors for a "single" application. According to the recent survey of ESM tools by Conner [2], the data collection schedule and design are fixed

throughout the experiment, e.g. choice of participants, number of participants, sampling time, experiment duration, etc. ESM tool capabilities and methodological process need to be further extended for scalability and efficiency to handle larger subject pools, longitudinal studies and A/B testing with several factors.

3 Case Study: ESM for A/B Testing

3.1 Test Application and Evaluated Interfaces

In this case study, we evaluate the usability of two competing interfaces for a simple mobile map-logging application (see Fig. 1) in which a user can record and tag short information about the current user location (indicated on the map and sensed by the GPS sensor). The two competing interfaces compared were for entering information through (1) voice and (2) touch typing of text. The ESM is used because the application is mobile and possibly context dependent (e.g. location, time, social setting, etc.). Either by voluntary initiation or by a scheduled prompt, the participant is asked to try out one of the interfaces (chosen in a balanced order) to enter short information (i.e. record in voice, or type in). In addition, several supplemental contextual information (that cannot be easily inferred automatically with the sensors) is solicited using a menu driven interface, asking whether the participant is indoor or outdoor, the location (e.g. restaurant, classroom, streets), on-going activity (e.g. resting, eating, in a meeting), social situation (e.g. alone, with a friend), etc.

3.2 ESM Based A/B Testing Process and the Support Tool

Typically, a comparative UI experiment is conducted in a laboratory setting as a cross sectional study with repeated measures taken in batch. On the other hand, ESM is used to attain more relevant data considering the environment and context of usage, but batch collection of repeated usage is often not feasible. Rather the data is collected over some extended period of time. Nevertheless, we still regard our experiment to be "cross-sectional" since we are not (for now) interested in longitudinal change in the user response. Note ESM has primarily been used for capturing user behavior for a single application rather than for making comparisons of usability or UX. In our case, the ESM A/B testing proceeds as a within-subject type (i.e. the user tries out both interfaces and make comparisons) for a given period of time with a predetermined number of participants to first collect some minimum amount of data deemed sufficient for the power of the experiment. While there are several methods to decide on the least amount of required data or number of participants, for now the experiment duration, amount of data solicited (equally number of repetition) and the number of participants were determined arbitrarily but in a conservative fashion (e.g. long enough to gather good amount of data).

The participants were scheduled in a balanced order to fulfill a task, either using the interface A or B, and make answers to usability and gold standard questions. Due to the difficulty in inferring particular usage contexts automatically (e.g. whether a person is moving, in a meeting, at a bus station, sleeping, etc.), the data solicitation

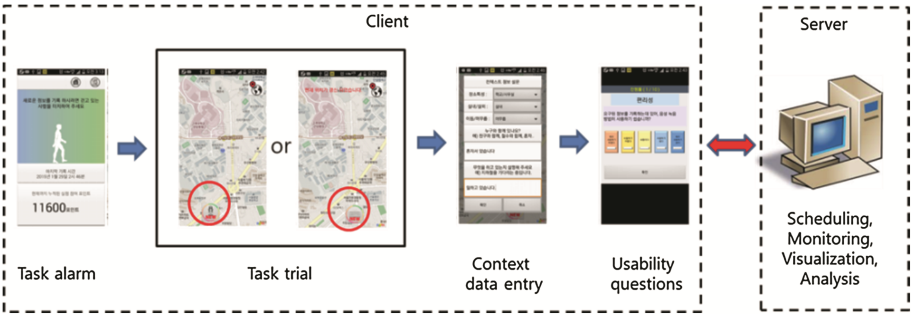

Fig. 1. The overall process of ESM based A/B testing of two interfaces (using the voice or text input) for a simple mobile map logging application.

was done according to a regular time schedule rather than invoked by automatic context detection. Figure 1 shows the overall flow of the ESM A/B testing process.

3.3 Detailed Experiment Procedure

The first phase of the ESM based A/B testing was conducted for six days, and solicited for data entry 8 times a day. For now, the duration of one week for the first phase was decided rather arbitrarily. A total of 30 subjects (20 males/10 females) mostly in their 20's with various occupational backgrounds participated in the study. The participants were recruited, interviewed and selected through an on-line social messaging system. The participants were given instructions as how to download and install the smartphone application, and how and when to try out the tasks and make proper data entries. Prior to the actual experiment, the participants were given instructions for a short training session for getting oneself familiarized with the application, two interfaces and data entry method.

The participants were compensated upon the completion of the whole session at rate of $0.23 per answered questions (which totaled to about $11 dollars maximum).

Considering the recommendation by [10], the task trials and data solicitations were scheduled every two hours only between 7:30 am to 10:30 pm (total of eight times a day). At the scheduled times, the application was invoked on the smartphone device automatically (4 times each for the respective interfaces in an alternative order), and an alarm was used to remind and notify the participant. Despite the reminders, it was up to the participants to actually respond. It was also possible that the alarm or smartphone itself was switched off. Thus, a few simple user behavioral checking measures were implemented. For example, 30 s of no response was regarded as a refusal of a data entry. It was also checked whether the phone was actually in active use before and after the scheduled time to guess whether the "refusal" was deliberate or not. Such piece of behavioral information was to be used collectively to assess the credibility or reliability of the participant. Upon a scheduled invocation of the task trial, the user was to enter information as attached to wherever the user was at, enter additional contextual information (as described in Sect. 3.1) and answer a series of usability questions in a 5 Likert scale (on convenience and ease of use,

general satisfaction, annoyance, fatigue). Finally, gold standard questions were given to assess and explicitly confirm the credibility and reliability of the participant. The gold standard questions were designed to be fairly easy with the least cognitive burden, yet not answerable by random guesses, such as solving a simple arithmetic (e.g. "what is (5 + 4)*(1 + 1)?") or asking of common sense knowledge in multiple choices (e.g. "who is the president of Korea?"). After completing the task of logging the map either by voice or text, there were a total of 10 questions to answer including the two gold standard quiz. After six days, the collected data were analyzed for sufficiency and (as explained more in detail in the next section) it was determined that another round of data collection was deemed necessary. A second phase of data collection for A/B testing was continued for another week. After the whole two week sessions, we administered one final survey, asking the participants about the ESM procedure itself (participants were separately compensated for it). The answers were used, in addition to the operational problems discovered during the case study, to base our proposal for an improved and extended ESM. We omit the presentation of the detailed survey questions and only brief the results in Sect. 4.4.

4 Results, Observations and Proposals

4.1 First Phase: Data Sufficiency and Balance

During the first phase (first six days), with 30 participants and 8 data collection sessions per day, we ideally would have collected 1440 sets of balanced and reliable task trial and session response data. However, only 463 session data (37.2 %) were collected due to reasons such as participant not noticing the incoming alarm (41.8 %) and deliberate refusals (20.3 %). At a closer look, the collected data were even more insufficient with respect to different contexts and number of participants. For instance, the data collected for @home context comprised more than half of the total data, while the rest scattered in little proportions to other usage location contexts and thus lacking the power for any meaningful analysis. The situation was worse for conjunctive contexts such as for a particular location and time, location and activity, etc. This was not only attributed to the fact that the data were collected based on a simple time based schedule (rather than based on intelligent, but technically difficult, context detection), certain context based usages just do not happen as often as others (e.g. staying home vs. riding on a subway). It was also possible that by the nature of the application, the users just was not up to using the application as often as necessary to gather sufficient interaction data in a short amount of time.

Thus, in order to reduce the experiment duration, save cost, relieve the burden on the participants and ultimately make the study more focused rather than open-ended and exploratory, we propose that intermittent data analysis would be necessary (as part of an extended ESM A/B testing study methodology) to check data sufficiency and analysis power, carry out the mid-evaluation if possible, and eliminate certain dependent variable measurements if the analysis results are clear (e.g. very low p-value, high R^2, high χ^2 etc.). In this study, the experiment continued on for another week (second phase) and about the same amount of data were additionally collected (See Fig. 2). The comparative

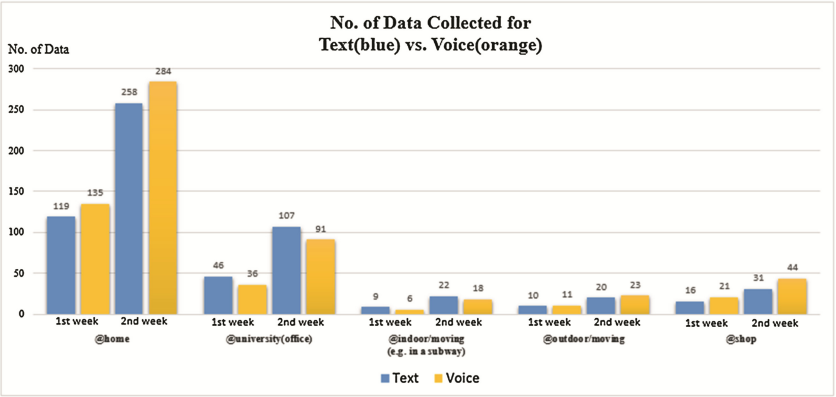

Fig. 2. Comparative usability data collected during the first six days (first two bars among the four) and after the second week (cumulative, the second two bars) for different location contexts (@home, @university, etc.). The dark blue bar represents data for text, and the light yellow for voice based interfaces (Color figure online).

usability between the text and voice based interfaces for the context of @home usage (which had sufficient data for analysis after the first week) did not change. Note that comparative qualitative assessments of the interfaces can be gathered as well but would require a subjective analysis.

In addition, the data collection process and scheduling must be tailored toward the particular context of interest. If automatic context detection is technically difficult, then a pilot study should be conducted ahead of time e.g. to recruit a participant who is likely to make a particular context-based usage of interest, or personalize the data entry session schedule at a right timing by an analysis of one's daily activities. Furthermore, the data were unbalanced in many ways, e.g. between the treatments (e.g. data for voice based interface vs. text based), among the participants, and among the contexts. The missed data entry sessions, deliberate or not, did not originate uniformly among different participants. While the unbalanced amount of data between treatments is a serious problem to making comparative analysis, it can be viewed as an indication of usability or preference. Thus, through continuous monitoring, the ESM must be administered in such a way to solicit data and closely balance the competing data e.g. by scheduling for treatments that lack data, encouraging the non-responding participants, and analyzing whether the unbalanced response is in fact due to preference or certain operational constraint or contexts. Note that such provisions can contaminate the balanced presentation of treatments (for mitigating the learning effect), thus must be applied carefully in an incremental manner.

4.2 Participant Reliability

The gold standard quiz is only a partial and indirect indicator of participant/data reliability. At any rate, we judged that the data themselves were reasonably reliable

and credible because only less than 5 % of the gold standard quizzes were incorrect overall. In addition, the response behavior did not change much over the second week. Thus, it seemed more important to single out the participants who tended to "refuse" (especially deliberately so) the data entry in the first place too often. Future ESM tools should have the support capability for monitoring for these participants and replacing them if necessary.

4.3 Experiment Extension and the Utility of ESM

Because sufficient data were not collected for a meaningful comparative analysis for different contexts of usage (except for the @home usage), the experiment was continued on for the second week (see Fig. 2). Future ESM tools should administer such an extended experiment in a systematic fashion through data analysis. Note that during the second week, the missed data entry session and data unbalance were still at the similar level. Future ESM tools must take measures to minimize these types of data insufficiency. On the other hand, sufficient data were then collected for the @university usage case, and showed different usability results from the @home usage (e.g. participants preferred text based input more for @university than @home). Thus, this at least confirms the very utility of the ESM in that it can capture different usability and user behavior depending on the usage context. After the additional experimentation, nothing much has changed except only the data for @university usage context became sufficient (for power of analysis). Data for other contexts were still lacking and unbalanced, and the additional data for @home usage did not change the initial analysis results.

4.4 Participant Responses About the ESM Process Itself

Participants mostly acknowledged the experimenter's sentiment of the difficulty in collecting reliable data. They suggested for a system of incentive based compensation, better alarm mechanisms, and pre-notification of the upcoming data entry sessions. Two main reasons for the missed data entry were not being able to notice the alarms and scheduled events overlapping with uninterruptable on-going activities. They also expressed that the gold standard questioning was effective not only as an indicator for credibility but also encouraged the participants to be more thoughtful and reliable.

5 Conclusion and Future Work

In this paper, we described a detailed case study of applying ESM to evaluating two competing interfaces for a mobile application. Based on the gathered data and direct interviews with the participants, we highlighted the difficulties experienced and lessons learned. In addition, we made several proposals for a new ESM (also our future work) with the capabilities to flexibly revise the parameters of the experiment on-line so that the ESM can be run economically, efficiently but with the same reliability.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2011-0030079) and funded by the Forensic Research Program of the National Forensic Service (NFS), Ministry of Government Administration and Home Affairs, Korea. (NFS-2015-DIGITAL-04).

References

1. Larson, R., Csikszentmihalyi, M.: The experience sampling method. New Dir. Methodol. Soc. Behav. Sci. **15**, 41–56 (1983)
2. Conner, T.S.: Experience sampling and ecological momentary assessment with mobile phones (2013). http://www.otago.ac.nz/psychology/otago047475.pdf
3. Intille, S.S., Rondoni, J., Kukla, C., Ancona, I., Bao, L.: A context-aware experience sampling tool. In: Proceedings of SIGCHI Conference Extended Abstracts, pp. 972–973. ACM (2003)
4. Consolvo, S., Walker, M.: Using the experience sampling method to evaluate ubicomp applications. IEEE Pervasive Comput. **2**(2), 24–31 (2003)
5. Froehlich, J., Chen, M.Y., Consolvo, S., Harrison, B., Landay, J.A.: MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In: Proceedings of the International Conference on Mobile Systems, Applications and Services, pp. 57–70. ACM (2007)
6. Carter, S., Mankoff, J., Heer, J.: Momento: support for situated ubicomp experimentation. In: Proceedings of the SIGCHI Conference, pp. 125–134. ACM (2007)
7. Meschtscherjakov, A., Reitberger, W., Tscheligi, M.: MAESTRO: orchestrating user behavior driven and context triggered experience sampling. In: Proceedings of International Conference on Methods and Techniques in Behavioral Research, p. 29. ACM (2010)
8. Consolvo, S., Harrison, B., Smith, I., Chen, M.Y., Everitt, K., Froehlich, J., Landay, J.A.: Conducting in situ evaluations for and with ubiquitous computing technologies. Int. J. Hum. Comput. Interact. **22**(1), 107–122 (2007)
9. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system?: screening mechanical turk workers. In: Proceedings of the SIGCHI Conference, pp. 2399–2402. ACM (2010)
10. Abdesslem, F.B., Parris, I., Henderson, T.N.H.: Mobile experience sampling: reaching the parts of Facebook other methods cannot reach. In: Proceedings of Privacy and Usability Methods Powwow (2010)