University of Memphis

University of Memphis Digital Commons

Electronic Theses and Dissertations

12-2-2014

Can Wearable Sensors Help Assess the Reliability of Self-Reports in Mobile Health Studies?

Hillol Sarker

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

Sarker, Hillol, "Can Wearable Sensors Help Assess the Reliability of Self-Reports in Mobile Health Studies?" (2014). *Electronic Theses and Dissertations*. 1086.
https://digitalcommons.memphis.edu/etd/1086

This Thesis is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.

CAN WEARABLE SENSORS HELP ASSESS THE RELIABILITY OF SELF-REPORTS IN MOBILE HEALTH STUDIES?

by

Hillol Sarker

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Computer Science

The University of Memphis

December 2014

©Hillol Sarker, 2014

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Santosh Kumar for mentoring me to grow as a research scientist. His consistent guidance, encouragement, and inspiration helped me to pursue my research. I would also like to thank my committee members, Dr. Vasile Rus and Dr. Ebenezer Olusegun George, for their constructive comments and suggestions, and serving as my committee members.

I would like to thank Dr. Emre Ertin from the Ohio State University, Dr. Mustafa al'Absi from the University of Minnesota Medical School, Dr. Karen Hovsepian from Troy University, and Dr. J. Gayle Beck, Dr. Satish Kedia, Dr. Kenneth D. Ward, Amin Ahsan Ali, Md. Mahbubur Rahman, Syed Monowar Hossain, Rummana Bari, and Sudip Vhaduri from the University of Memphis for their contributions. I would also like to thank Jeremy Luno and Lucas Salazar regarding their help in conducting the high burden user study in the natural environment which is an integral part for my thesis.

A special thanks to my family including my supportive parents, my beloved wife Gitaly Das, and my wonderful kid Amio Sarker for their sacrifices during the journey and being a constant source of inspiration.

This work was supported in part by NSF grants CNS-0910878 (funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5)), CNS-1212901, IIS-1231754, and by NIH Grants U01DA023812 and R01DA035502, from NIDA.

Finally, I would like to thank the University of Memphis for providing me with an excellent research platform as well as the constant support throughout my academic career.

ABSTRACT

Sarker, Hillol. MS. The University of Memphis. December, 2014. Can Wearable Sensors Help Assess the Reliability of Self-Reports in Mobile Health Studies? Major Professor: Dr. Santosh Kumar.

Self-report in the form of Ecological Momentary Assessment (EMA) has been the primary instrument to collect measurements from participants in their natural environment. Given numerous sources of biases and inaccuracies in self-report, assessing and improving the reliability of self-report has been the subject of continuing research. However, to date, there exist only limited lab based methods to check the veracity of collected self-report data. Increasing adoption of sensors in field studies that sometimes can passively measure the same phenomena that have been traditionally included in EMA self-report has opened up a new opportunity to assess the reliability of self-reports.

In this paper, we use data collected in a week-long field study with wearable sensors to first investigate whether lack of agreement between self-reported location and GPS-inferred location can be used to predict the reliability of self-reports. We find this not be the case, primarily because lack of agreement on location results from sensitivity of some participants to reporting locations and it does not indicate lack of care in completing self-reports. We then investigate whether contexts of the participants, such as place (from GPS), activity level (inferred from accelerometers), or stress (inferred from physiological sensors) are associated with low reliability. We find that not being at home or work does not predict reliability of self-report, nor does the context when participants are engaged in physical activity at the time of receiving the self-report prompts. However, we do find that if the participants are stressed at the time of receiving a self-report prompt, then reliability of self-reported data is low. This implies that unless demanded by the study protocol, self-report prompts should be avoided when participants are under stress.

TABLE OF CONTENTS

Content Page

List of Figures vii

List of Tables ix

1 Introduction 1
1.1 Background 1
1.2 Problems 2
1.3 Research Objectives 3

2 Related Works 6
2.1 Behavioral Science 6
2.2 Crowdsourcing 7
2.3 Technological Assessment of Reliability 7

3 Study Description 10
3.1 Wearable Sensor Suite 10
3.2 Mobile Phone 10
3.3 Self-report Measures 11
3.4 EMA Scheduling Algorithm 12
3.5 Field Study Procedure 13
3.6 Incentives 14
3.7 Participants 14
3.8 Data Collected 15

4 Computational Procedure 16
4.1 Metric for Assessing Reliability 16
4.2 Inference of Location & Kappa Score 19
4.3 Agreement between self-reported location and GPS-inferred location 20
4.4 Context Inference 22
4.4.1 Activity Inference 22
4.4.2 Stress Inference 23

5 Results 25
5.1 Participant's alpha score 25
5.2 Participant's kappa score 25
5.3 Relation between kappa score and alpha score 27
5.4 Role of Context in Assessing Reliability of Self-reports 28
5.4.1 Role of Location Context 28
5.4.2 Role of Activity Context 30
5.4.3 Role of Psychological Context 31

6 Discussion, Implications, and Limitations 34

7 Conclusion **36**

References **37**

LIST OF FIGURES

Figure **Page**

3.1 Participant User Interface. From left, the first window shows the participant how much money he/she has earned thus far for study compliance. They can also report smoking or drinking events from this window. The second window shows a confirmation of smoking or drinking self-report. The third window shows one example of an EMA question about commuting. The fourth window shows the end of an EMA interview. At the top of each window, there is a connection status bar to help the user monitor the status of the sensor network connection and correct it if necessary. 11

4.1 Procedure to compute the reliability score and location, stress and activity states of the participants. 17

4.2 One example of GPS trace for one day of one participant. The participant drives to a work-place and a restaurant. We record the location where EMA prompt was triggered. The red line shows the path commuted by the participant. The pinned locations are the location at the time of EMA prompt. 18

4.3 Location mentioned in the top 7 boxes (blue) are possible responses of this item. There is a match if GPS log indicates that participant is in a location mentioned in green colored box and he reported a location mentioned in corresponding root blue box (ground truth location). Similarly, being at a location mentioned in yellow colored box and reporting a location mentioned in the corresponding root blue box was acceptable, and we mark their ground truth location as the one reported. Otherwise, we keep the report as it is. 21

4.4 Standard deviations < 0.21384 are labeled as stationary and others are labeled as non-stationary (i.e., walking or running). 22

4.5 The Lab Study Procedure for stress inference model. After 30 minute rest period participants receive 3 stressors, public speaking, mental arithmetic, and cold pressor. 23

5.1 Participant's Cronbach's alpha. Alpha is sub-sectioned into six groups [1], unacceptable(α <0.5), poor [0.5, 0.6), questionable [0.6, 0.7), acceptable [0.7, 0.8), good [0.8, 0.9), and excellent ($\alpha \geq$ 0.9). We observe that only two participants have unacceptable alpha. 26

5.2 Agreement score kappa (κ) measured from location inferred from GPS and participant's response to the location item. Overall κ for all participants is 0.78. Acceptable line is placed at 0.7. 26

5.3 Participant alpha in case of being at home or work, or other places response is plotted in primary axis. Difference of them for each participant is plotted in secondary axis. CI lower limit for the difference of samples of Hypothesis 2 is also plotted in secondary axis. 29

5.4 Participant alpha in case of activity or being stationary while EMA response is plotted in primary axis. Difference of stationary and activity for each participant is plotted in secondary axis. CI for difference of samples of Hypothesis 3 is also plotted in secondary axis. 31

5.5 Participant alpha in case of being stressed or not stressed response is plotted in primary axis. Difference of alpha of not stressed and stressed for each participant is plotted in secondary axis. CI lower limit for difference of samples of Hypothesis 4 is also plotted in secondary axis. 32

LIST OF TABLES

Table **Page**

3.1 Summary of data collected in the user study. 15

4.1 Confusion Matrix for the Semantic Labeling model [2]. Restaurant is sometimes confused with store. 18

4.2 Accuracy using semantic labeler model [2]. TP = true positive rate, FP = false positive rate, P = precision, R = recall, F = F-Measure, and AUC = area under the curve. 19

5.1 Context score for participants with unacceptable acceptance score of alpha or kappa. P# for participant id, α for Cronbach's alpha, κ for kappa agreement score, H+W for Participant being at Home or Work, and Other for any place other than Home or Work. 28

5.2 Hypothesis testing summary. H for Hypothesis number and n for sample count (pair). For each hypothesis both parametric paired t-test and non-parametric Wilcoxon Signed Rank test is performed. 33

Chapter 1

Introduction

Mobile technology has a potential to provide unprecedented visibility into the health status of individuals in their natural environment [3]. Sensors embedded in smart phones such as GPS, accelerometers, and microphone, and those worn on the body with wireless connectivity to smart phones to assess electrocardiography (ECG), respiration, galvanic skin response, etc. can continuously monitor an individual's health, behavior, and the surrounding environment. Machine learning algorithms have been developed to obtain measures of behavior and the exposure to the environment such as activity from accelerometers, geoexposure from GPS, stress from physiology, and social context from microphone. These automated measures of behavioral and environmental contexts complement and enrich the self-reporting methods traditionally used in health research studies conducted in the field environment.

1.1 Background

Self-report in the form of Ecological Momentary Assessment (EMA) [4] or Experience Sampling Method (ESM) has been the primary instrument to collect measurements from participants in their natural environment. These approaches involve repeated assessment of a participant's behavior, emotion, and the associated contexts in real time. EMA (or ESM) is widely used by behavioral psychologists and social scientists, as EMA provides more reliable assessment of the participant's behavior, states, and environment as compared to retrospective self-report and is capable of measuring the subtle person-environment interplays [5, 6]. Although momentary assessment reduces the limitations of recall bias and assures that self-report data are provided when requested, reliability of EMA has been questioned for a variety of reasons [7, 8]. First, the reliability of self-reports may be low due to subjective biases. Such subjective biases arise due to a number of factors, such as the participant's reporting style, lack of motivation, lack of attention [9, 10], or urgency of completing the task . Also the physical condition of the participant (e.g., fatigue, alcohol consumption, drug withdrawal) may lead to unreliable self-reports. Third, excessive study burden such as frequent prompts, long questionnaires, and long study duration may impact the reliability of self-reports. Low reliability may also occur due to concerns about the social desirability of one's response or the sensitivity of the behavior that is being assessed (e.g., illicit drug use, risky sexual behavior). Finally, EMAs have the potential to actually change the phenomena measured [11, 12], in particular reduce the occurrence of undesired behavior. This sort of measurement reactivity can greatly reduce reliability of measurement.

1.2 Problems

There has been considerable amount of work on the feasibility of EMA in psychophysiological studies of addictive behavior (e.g., smoking, drug usage, process of relapse, and coping with withdrawal) [11, 13, 14], pain [5], and psychological disorders (e.g., mood disorders, anxiety disorders, and psychosis) [15, 16, 17, 6]. These studies assess the validity of EMA based methods relative to clinical standards or standard questionnaires. To date, there exist only limited lab based methods to check the 'veracity' of collected self-report data. For example, in case of substance use, biochemical markers (e.g., urine tests for drug usage, measurement of carbon monoxide levels in breath for smoking, and in breath analyzers in case of alcohol use) have been used for this purpose [11].

Separately, several methods have been proposed to determine the reliability of self-reports via careful design of the questionnaire. One such method is to add reliability check questions into the questionnaire. These questions are intended to help the investigator determine the accuracy of self-report responses, as well as potentially assess response biases such as social desirability.

1.3 Research Objectives

With increasing adoption of sensors in health research studies, there is a new opportunity to obtain some measures of interest passively (without having to ask the participant), such as location. These sensors may thus reduce self-report burden. Since these sensors can sometimes measure the same phenomena that have traditionally been included in self-reports, they may also be used to assess the reliability of self-reports by checking agreement between self-reported answers with that obtained from the sensors. It is not known whether such a method can accurately measure the reliability of self-report since the source of unreliability in self-report may be affected by lack of attention, urgency of completion, selective sensitivity to specific items in self-report, and social desirability, among several other factors. Another way sensors could be used is to infer the context of an individual at the time of receiving the self-report prompt. Can automated detection of such contexts predict the reliability of self-reports? If the former approach is successful, unreliable self-reports may be discarded from analysis. If the latter approach works out, self-report prompts can be scheduled so as to not occur during contexts that are associated with low reliability, unless dictated by the study protocol.

In this study, we investigate whether lack of agreement between self-reported location and that inferred from GPS can be used to predict the reliability of self-reports. We also investigate whether contexts of the person such as stress (inferred from physiological sensors) and activity level (inferred from accelerometers) are associated with low reliability. We use data collected in a week-long field study with a student population who self-reported themselves to be daily smokers and social drinkers. The goal of this study was to re-examine the relationship between stress, smoking, and alcohol consumption, when these behaviors are measured by wearable sensors, as compared with self-reports. The participants in the study wore a wireless physiological sensor suit that collected ECG, respiration, and accelerometry and carried a smart phone that included GPS and accelerometers. Self-reports were obtained on the same phone that collected data from all the sensors.

To measure reliability of self-reports, we compute Cronbach's alpha (α) [18] over six items on the questionnaire that were intended to measure the same psychological construct, namely, affective state of the participants. We first check whether these six items indeed measure the same construct. The value of α for these six items for all subjects for entire self-reports was 0.88, which indicates that these six items are indeed consistent. Second, we compute the degree of agreement (κ) [19] between self-reported location with that inferred from the GPS sensor. We find that unacceptable agreement on the location item does not predict low reliability of self-report. This may be due to the sensitivity of participants in reporting their location. In other words, they may not be comfortable in reporting their location, under certain situations, but answer the affect questions carefully (indicated by acceptable α), which we take to imply that their self-reports are reliable. Third, we infer the physical activity episodes from accelerometers worn on the body and stress from physiological measurements [20]. We hypothesize that subjects may be pressed for time when they are in motion. However, we find that physical activity is not associated with low reliability. Interestingly, we find that if the participants are stressed at the time of receiving the self-report prompt, their reliability across the affect items are indeed unacceptable with statistical significance. Stress has traditionally been measured using the six affect items that was included in the self-report measure and hence it may not have been possible to infer from the answers whether the self-report was reliable under high-stress situations. Development of stress measure that is inferred passively from physiological arousal has opened up a new opportunity for assessing the reliability of self-report when under stress, which we do find to be associated with low reliability. In addition to addictive behavior such as alcohol and illicit drug use that has usually been associated with low reliability of self-report, our work indicates that stress may be another factor of interest when assessing reliability of self-report. We also examined whether a self-report completed while driving is associated with low reliability[1], but we did not have sufficient data points to test for statistical significance for this context.

The rest of this paper is organized as follows. Chapter 2 discusses related works. Chapter 3 describes the study and the data collected. Chapter 4 provides computational methods involved in making inferences from the sensors and the reliability metrics used. Chapter 5 presents the results of testing the reliability of self-reports utilizing inferences made from sensors. We discuss the implications and limitations of this work in Chapter 6 and conclude the paper in Chapter 7.

[1] Given the random nature of EMA[4], some did occur during driving. Participants were instructed to park the car to a safe place and answer the EMA.

Chapter 2

Related Works

In this chapter, we discuss related works on validity of EMA and self-report in behavioral research and in crowdsourcing works. We then discuss technological methods that have been used to assess validity of self-reports.

2.1 Behavioral Science

As mentioned in the Introduction, adoption of EMA assessment is quite extensive in behavioral science research and assessment of validity of self-reports has continued to be of interest to the research community [11, 15, 16, 6]. One approach to assess the validity of EMA is to check its agreement with traditional or clinical instruments. This approach, however, has produced mixed results [21]. Although a good correlation between the measurements obtained by EMA methods and recall based methods confirm the validity of EMA, a moderate or even low correlation does not necessarily indicate that EMA methods suffer from low validity. A low correlation could result from inaccuracy in recall based methods, due to bias. Although, clinical standards are considered to be 'gold standards' meet agreement with EMA is not assured, as clinical standards are derived from the lab while EMA is derived from the natural environment. Therefore, assessing the validity of EMAs in field studies is more challenging.

A second approach is to compare EMA with an objective measure. For example, in studies on substance use, researchers propose the use of biochemical markers. For example, urine tests for drug usage, measurement of carbon monoxide levels in breath for smoking, and in breathanalyzers for alcohol use, are reported to be used in different studies [11]. However, most of these tests require participant's compliance and lab equipment, which may not scale well.

2.2 Crowdsourcing

Crowdsourced human judgments using services such as Amazon's Mechanical Turk [22, 23] struggle with problems like validity or reliability. Similar to the concerns associated with the reliability of EMA responses, there are concerns of biases in responses, error in responses, and spamming. To increase measurement reliability, [24] proposed to add verifiable questions in a data collection study. In this particular work, the users were asked to rate the quality of Wikipedia pages. Authors noted that the percentage of invalid summaries reduced from 38% to 7% in the case where the user had to input the number of references, images, and sections the page had as well as 4-6 keywors that provide a good summery of the topic of the page. Also, the correlation between the ratings provided by the user and that of Wikipedia administrators were significantly higher than when users were asked only to rate the quality of a specific Wikipedia page. These verifiable questions helped the researchers to identify unreliable responses as well as signaled to users that their responses might be scrutinized. Zhu [25] showed that verifiable questions alone may not be enough to identify unreliable users. They suggest using the time spent on individual items and the pattern of responses provided (e.g., lack of variance in responses) to create a metric of reliability. As another approach, [23] suggested using multiple indicators per construct in order to improve the reliability of responses. None of these works make use of sensors.

2.3 Technological Assessment of Reliability

Prince [26] present an extensive literature review about direct versus self-report measures for assessing physical activity. Direct measures provide precise estimation of energy expenditure using sensors like accelerometers, pedometers, heart rate monitors, calorimetry (i.e., using doubly labeled water), and physiological markers (i.e., cardiorespiratory fitness). Correlations between self-report and direct measure ranged from -0.71 to 0.96. Given such wide variability, it was not clear in that case whether the two instruments measured the same phenomena.

The reliability of self-reported behavior was examined in a long-term study on smartphone use in [27]. In this work, a background service logged the usage of Gmail and Facebook by participants. Using self-report, participants provided their recall about usage of these two applications. Three conditions were included: voluntary reporting, prompted reporting on a set interval, and prompted event reporting. It was shown that self-report was not able to provide a reliable estimation of application usage duration in any of these conditions while using Gmail. However, in case of Facebook it is likely to overestimate the usage duration.

Elgethun [28] proposes the use of GPS to collect information about participant location and showed that parents tend to under-report time spent by their children at home, and overestimate when in other locations such being outdoors or in transit. Diaries of mothers doing unskilled labor jobs or staying at home also have low concordance with GPS. Stopher [29] shown that participants tend to under-report about travel distances made but over-report the total duration of travel time. These studies indicate that use of GPS can be a reliable tool to measure the reliability of self-report about location. But, it is not clear if lack of agreement between GPS inferences and self-report translate over to lack of reliability in rest of the items in self-report.

In summary, the problem of assessing reliability of self-report collected in the natural field environment is still an open problem, which can be revisited due to the increasing adoption of sensors in field studies. It is of great interest to determine if agreement between measures obtained from sensors and that from self-report can predict the reliability of self-report. And, if not, what role do contexts inferred from sensor data play in assessing the reliability of self-reports?

Chapter 3

Study Description

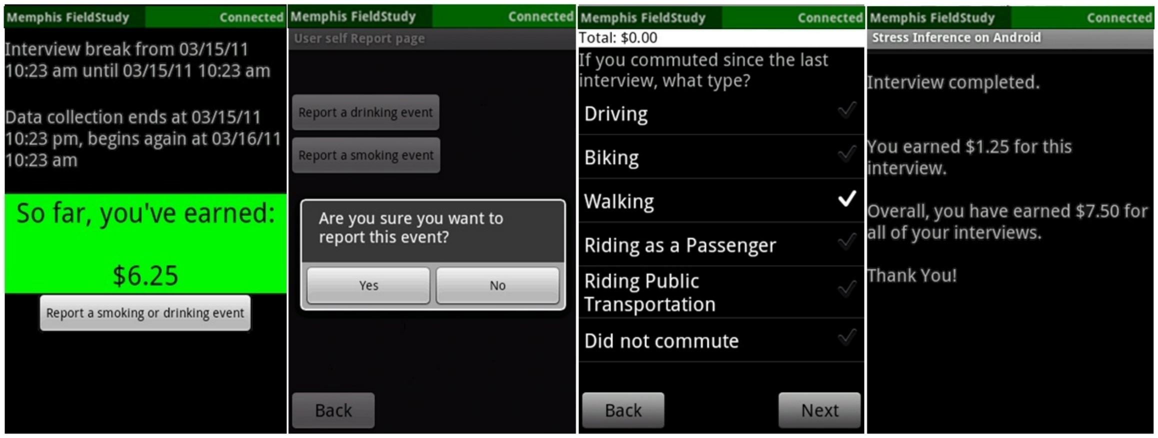
We use data collected in a scientific user study that was designed to investigate the relationship among stress, smoking and alcohol use in the natural environment over seven consecutive days. The study was approved by the Institutional Review Board (IRB), and all participants provided written informed consents. The novelty of the study was the use of sensors to assess stress, physiology, and alcohol use.

3.1 Wearable Sensor Suite

Participants wore a wireless physiological sensor suite underneath their clothes. The wearable sensor suite consists of two unobtrusive, flexible bands worn about the chest and upper arm, respectively. The chest band provided respiration data by measuring the expansion and contraction of the chest via inductive plethysmography (called RIP), two-lead electrocardiograph (ECG), 3-axis accelerometer, temperature sensors (ambient and skin), and galvanic skin response (GSR). The band worn about the upper arm contained a WrisTAS transdermal alcohol sensor, allowing measurement of the participant's alcohol consumption, GSR, and skin temperature.

3.2 Mobile Phone

Participants carried a smart phone where software was installed to communicate with the sensor suite. The mobile phone had four roles. First, it robustly and reliably received and stored data transmitted by the sensor suite. Second, it stored data from sensors local to the phone, including GPS and accelerometers. These measurements were synchronized to the measurements transmitted from wearable sensors. Third, participants used the phone to complete system-initiated self-reports in the field. Fourth, participants self-reported the beginning of drinking and smoking episodes by pushing a button.



Interview break from 03/15/11 10:23 am until 03/15/11 10:23 am

Data collection ends at 03/15/11 10:23 pm, begins again at 03/16/11 10:23 am

So far, you've earned:

$6.25

Are you sure you want to report this event?

Total: $0.00

If you commuted since the last interview, what type?

Interview completed.

You earned $1.25 for this interview.

Overall, you have earned $7.50 for all of your interviews.

Thank You!

Fig. 3.1: Participant User Interface. From left, the first window shows the participant how much money he/she has earned thus far for study compliance. They can also report smoking or drinking events from this window. The second window shows a confirmation of smoking or drinking self-report. The third window shows one example of an EMA question about commuting. The fourth window shows the end of an EMA interview. At the top of each window, there is a connection status bar to help the user monitor the status of the sensor network connection and correct it if necessary.

The phone software has user interfaces (UI) for both the study coordinator and study participants. Participant UI (see figure 3.1) is used by the participants to provide self-report. The participants have the option to move backward and forward through the questions using the buttons at the bottom of the interface. However, the interface does not allow viewing the next question unless the current question is answered. At the end of the EMA questionnaire, the interface also shows the incentives earned for responding to the EMA and the total earned so far.

3.3 Self-report Measures

The mobile phone initiated field questionnaires based on a scheduling algorithm (described below), which is a hybrid of time-based and event-based EMA triggering mechanism. The 42-item EMA asked participants to rate their subjective stress level on a 6-point scale as well as provide additional contextual data on stress, smoking, and drinking episodes, such as whether the user is in conversation, whether the user is smoking alone or with others, and the number of drinks consumed.

Several measures were adopted to reduce the burden of the study on the participant. First, the smart phone software was programmed to deliver no more than 20 questionnaire prompts in a day. Second, at least 18 minutes had to pass between prompts. Third, the questionnaire was designed to limit the time required to complete it to between 1 and 3 minutes. Fourth, participants had the option of delaying a questionnaire for up to 10 minutes. If the participant did not respond to the prompt at the second opportunity, the prompt would disappear. Fifth, participants were also allowed to specify time periods when they did not wish to receive prompts (e.g., during exams).

3.4 EMA Scheduling Algorithm

The scheduling algorithm is designed to balance two competing goals: (1) minimize the burden on participants and (2) maximize the collection of fine-grained ecologically valid self-reports. To keep the burden of the study low, a minimum duration between two successive EMAs is always maintained (currently 18 min). Further, the scheduling algorithm uses budgets to guarantee some data is collected for events of interest (assuming the event occurs) without exceeding limits on participant burden.

The first budget is a global daily budget. When the global daily budget is exceeded, no more EMAs can be triggered by the system until the following day. The global daily budget is split into two subcategories (1) Event-based triggering and (2) Random triggering. Randomly triggered EMAs capture baseline behavior over the course of the day. Event-based triggering captures information associated with a specific event (behavior) of interest.

Each type of event (e.g., smoking, drinking, speaking, etc.) also has its own daily budget. As with the global daily budget, when an event's daily budget is exceeded, no more EMAs can be triggered in response to this event until the following day (because we collected a sufficient amount of information about that event). When an event of interest is detected, the decision to trigger an EMA is made with a probability proportional to the ratio of expected number of remaining events of that type on the current day to the number of remaining budget for events of that type. Initially, empirical estimates of the expected number of each event per day are used. In long-term studies, these estimates can be personalized to the participant.

Whenever an EMA is completed, the next time to trigger an EMA is computed by dividing the remaining time of the day by the remaining total budget and adding a random amount of time to it (±5min). Finally, the triggering of EMA is tied to physiological data collection. If in the last 30 minutes more than 40% of the sensor data is of bad quality [30], no EMA is triggered. As participants incentives are proportional to the number of EMAs completed, participants must wear the physiological sensors if they are to earn any incentives.

3.5 Field Study Procedure

A training session was conducted to instruct participants on the proper use of the field study devices. Participants were instructed on the proper procedures to remove the sensors before going to bed and put them back on correctly the next morning. In addition, participants received an overview of the smart phone software's user interface, including the EMA questionnaires and the self-report interface. Once the study coordinator felt the participant understood the technology, the participant left the lab and went about their normal life for seven days. For all seven days, the participant was asked to wear the sensors during waking hours, complete EMA questionnaires when prompted, and self-report smoking and drinking episodes.

3.6 Incentives

Participants could earn up to $300 for participating in the study. Participants earned $15 for completing a 90-minute lab session on affect; $35 for attending daily lab appointment during the field study ($5 per day); and $75 for completing all end-of-study procedures and returning the sensing equipment. The remaining $175 was awarded based on compliance with the field study protocol. Completing a self-report questionnaire was worth $1. An additional $0.25 bonus was awarded if the questionnaire was completed within five minutes. A maximum of 20 requests for self-reports occurred each day. Thus, the participant could earn up to $25 per day ($1.25 x 20 self-report requests), adding up to $175 over seven days of field study ($25 x 7). Since wearing physiological sensors and answering 42-items questionnaire upto 20 times daily are highly burdensome, level of compensation was derived from the prevailing wage in similar behavioral science studies [31] that involve wearable sensors. Most user-studies provide fixed incentive to participants for completing the study [32, 33, 34, 35], while some studies were purely voluntary [36]. We believe that micro-incentive [31] associated with each EMA helps obtain a stronger measure of unreliability in self-report.

3.7 Participants

Participants in the study were recruited from the student population at a large university (approximately 23,000 students) in the United States. Thirty participants (15 male, 15 female) with a mean age of 24.25 years (range 18-37) were selected who self-reported about being “daily smoker" and “social drinker". Two participants dropped out from our study. One of them indicated that the length of the chest-band and sensor connector was not large enough for the participant’s size, and another fell sick during her scheduled participation time.

Table 3.1: Summary of data collected in the user study.

Category	Description
Participant	30 (15 male, 15 female)
Age	24.25 ± 6.25
Duration	1 week
Data Collected	2,064 hours of good quality sensors data 9.83 hours per day
Self-report prompts (EMA)	2717 in total 13.3 EMA per day (upper limit 20) compliance 94%

3.8 Data Collected

Average number of EMA prompts delivered per day was 13.33, well below the upper limit of 20 per day. EMA compliance rate is 94%. Participants delayed 2.28% of EMAs due to being busy or not being available at that specific time of EMA prompt. An average of 9.83 hours per day of good quality sensors data was collected from physiological sensors across all participants. Table 3.1 summarize data collected in the user study.

Chapter 4

Computational Procedure

In this section, we describe the metric for assessing the reliability of self-reports and our procedure for computing various contexts such as location, stress, and activity from sensor data. Figure 4.1 describes the components and the outcomes of the computational procedure described in this section.

4.1 Metric for Assessing Reliability

We use Cronbach's alpha [18] to assess the reliability of EMA responses. Cronbach's alpha measures the internal consistency of items that measures the same psychological construct. Let k be the number of items, where σ_i^2 is the variance of the i-th item, and σ_T^2 is the variance of the total scores formed by summing up all the items. Cronbach's alpha score is given by

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum \sigma_i^2}{\sigma_T^2})$$

We observe that if all the items have equal variance and thus were perfectly correlated, we obtained $\alpha = 1$. On the other hand, if all the items were independent, $\alpha = 0$. In most studies, an alpha score of 0.7 or higher is regarded as acceptable [18].

Our study questionnaire contains several affect items, *Cheerful?*, *Happy?*, *Energetic?*, *Frustrated/Angry?*, *Nervous/Stressed?*, and *Sad?*, where participants responded on a Likert scale of 1–6. To compute alpha, items that assessed positive affect (*Cheerful*, *Happy*, and *Energetic*) were retained as scored and items that assessed negative affect (*Frustrated/Angry*. *Nervous/Stressed*, and *Sad*) were reverse coded (e.g., 1 becomes 6). The overall alpha score for our questionnaire was found to be 0.88, which falls in the good region [1]. This implies that there is good internal consistency among the selected items.

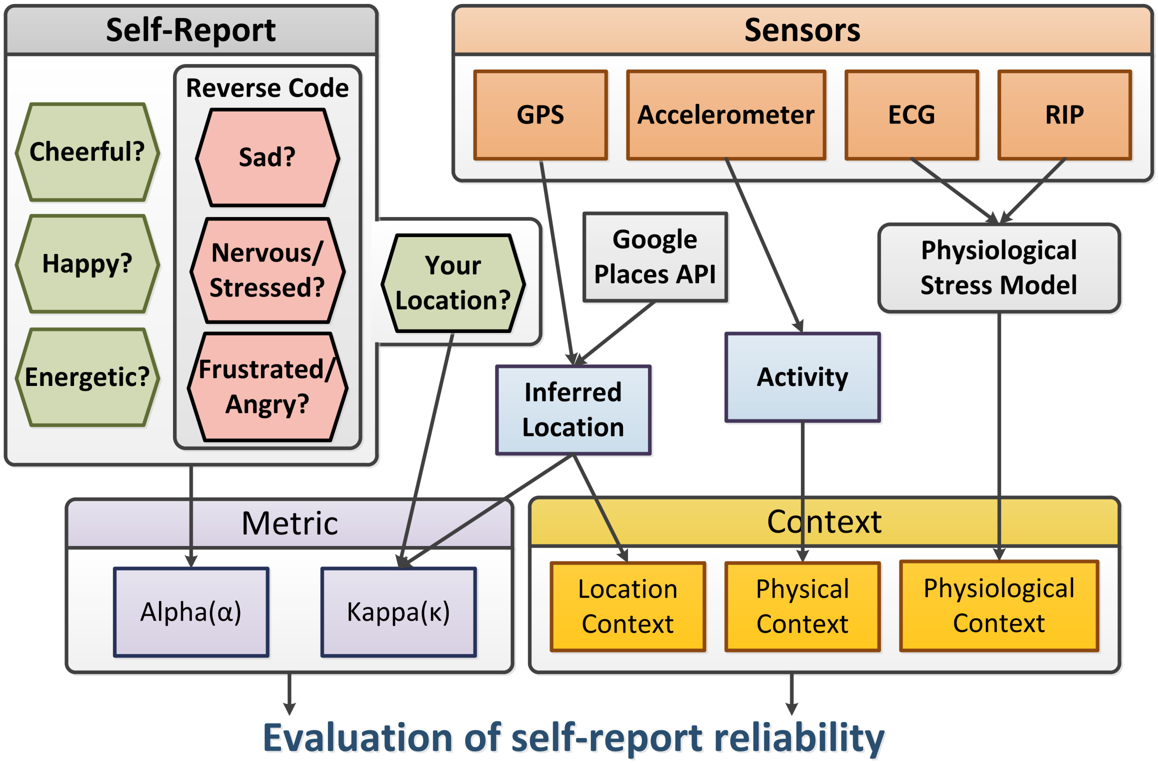


Fig. 4.1: Procedure to compute the reliability score and location, stress and activity states of the participants.

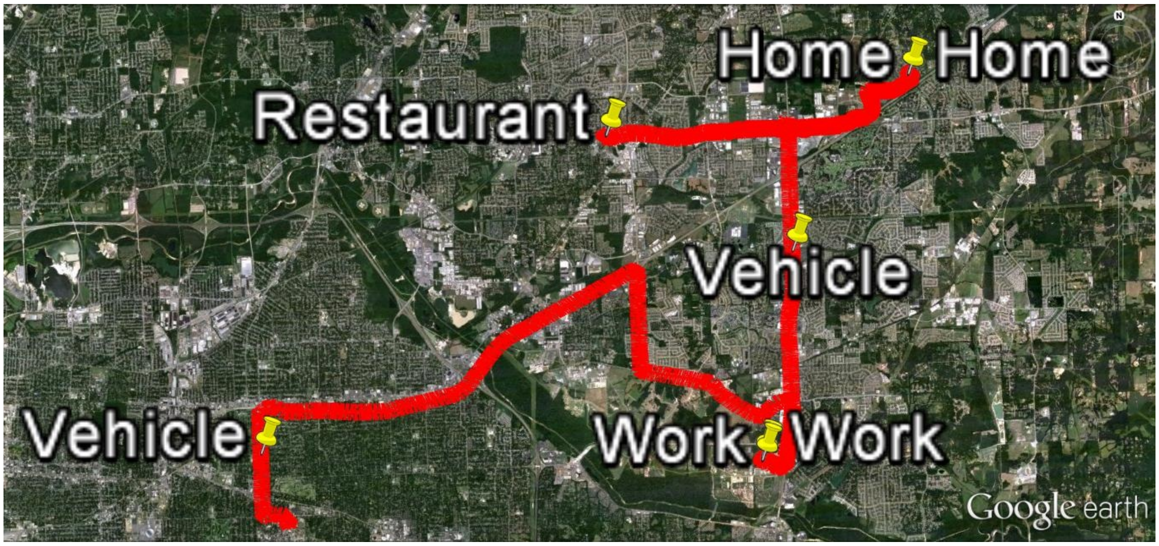


Fig. 4.2: One example of GPS trace for one day of one participant. The participant drives to a work-place and a restaurant. We record the location where EMA prompt was triggered. The red line shows the path commuted by the participant. The pinned locations are the location at the time of EMA prompt.

Table 4.1: Confusion Matrix for the Semantic Labeling model [2]. Restaurant is sometimes confused with store.

	Classified as				
	Home	Work	Store	Restaurant	Other
Home	617	11	10	0	4
Work	12	708	1	0	1
Store	8	7	203	6	9
Restaurant	4	1	43	27	3
Other	62	14	40	1	96

To measure reliability of EMA responses, we compute Cronbach's alpha across all EMA's that satisfy a specific condition (e.g., all EMAs that were triggered when at the "work" location). Given that we observed overall Cronbach's alpha 0.88 for selected 6 items, we compute alpha for a specific condition only if there exist at least 10 EMAs so that we get a consistency score with power 0.95 at level of significance $\alpha = 0.05$ [37, 38].

Table 4.2: Accuracy using semantic labeler model [2]. TP = true positive rate, FP = false positive rate, P = precision, R = recall, F = F-Measure, and AUC = area under the curve.

	TP	FP	P	R	F	AUC
Home	0.95	0.08	0.86	0.95	0.90	0.98
Work	0.97	0.03	0.96	0.97	0.97	0.99
Store	0.79	0.06	0.67	0.79	0.73	0.96
Restaurant	0.37	0.01	0.60	0.37	0.46	0.92
Other	0.43	0.02	0.77	0.43	0.55	0.89
	0.86	0.05	0.85	0.86	0.85	0.97

4.2 Inference of Location & Kappa Score

Locations of interest and their semantic labels were determined from the GPS traces that were collected on the phone. Figure 4.2 shows a typical GPS trace of a participant for one day. Places of interest for a participant were places where the participant spent a significant amount of time. We first applied a clustering algorithm to the GPS data using the method proposed in [39]. Distance threshold of 100 meters and temporal threshold of 5 minutes were used to find the spatio-temporal clusters throughout the day for each participant. These clusters represented the locations of interest. Next, we assigned semantic labels to these locations using the method proposed in [2]. This method used demographic, temporal and business features. Demographic features include the age and gender of the participants which were collected from the participant recruitment forms. The temporal features included the arrival time, visit midpoint time, departure time, season, holiday, and the duration of stay at that location. These features were also computed from the GPS traces and clusters. Lastly, the business features include the count of different types of business entities such as Arts/Entertainment, Food/Dining, Government/community, Education etc. within the different distance thresholds of the current location (see [2] for details). To compute the business features, we used Google Places API. For this model developed in Weka, we got an accuracy of 85.75% and a kappa of 0.80. Table 4.2 presents detailed accuracy across different locations and Table 4.1 presents the confusion matrix of this semantic context labeler model. It can be noticed that *Home*, *Work*, and *Store* are detected quite well. However, *restaurant* is confused with *store*, because a store and a restaurant can be located close to one another. We, therefore, corrected the labels (if necessary) by plotting the GPS traces in Google earth and visually inspecting the locations. These location labels were considered as ground truth. However, for some locations it was still not possible to have a single label, for example, for some locations we could not reliably distinguish between a store and a restaurant. We discarded these data points.

4.3 Agreement between self-reported location and GPS-inferred location

In our study questionnaire, the possible responses to the question, *What is your Location?*, are *Home*, *Work*, *Store*, *Restaurant*, *Vehicle*, *Outside*, and *Other*. Therefore, we mapped the location labels obtained from the procedure described above to these seven categories. We used the tree in Figure 4.3 to resolve ambiguities in the mapping process. Location mentioned in the top 7 boxes (in blue) are possible responses. For each of these locations, Figure 4.3 shows possible locations that are mapped to it. For those locations for which we have more than one label, we consider the self-reported location to be a match if it matches with any of the possible labels.

We used Cohen's Kappa (κ) [19] to measure agreement between the self-reported location and the GPS-inferred location. Kappa varies from -1 to 1 where, κ scores of 1 and -1 indicates absolute agreement and disagreement respectively, while $\kappa = 0$ indicates that agreement is due to random chance. In most behavioral studies, $\kappa > 0.7$ is considered to be satisfactory. Similar to the

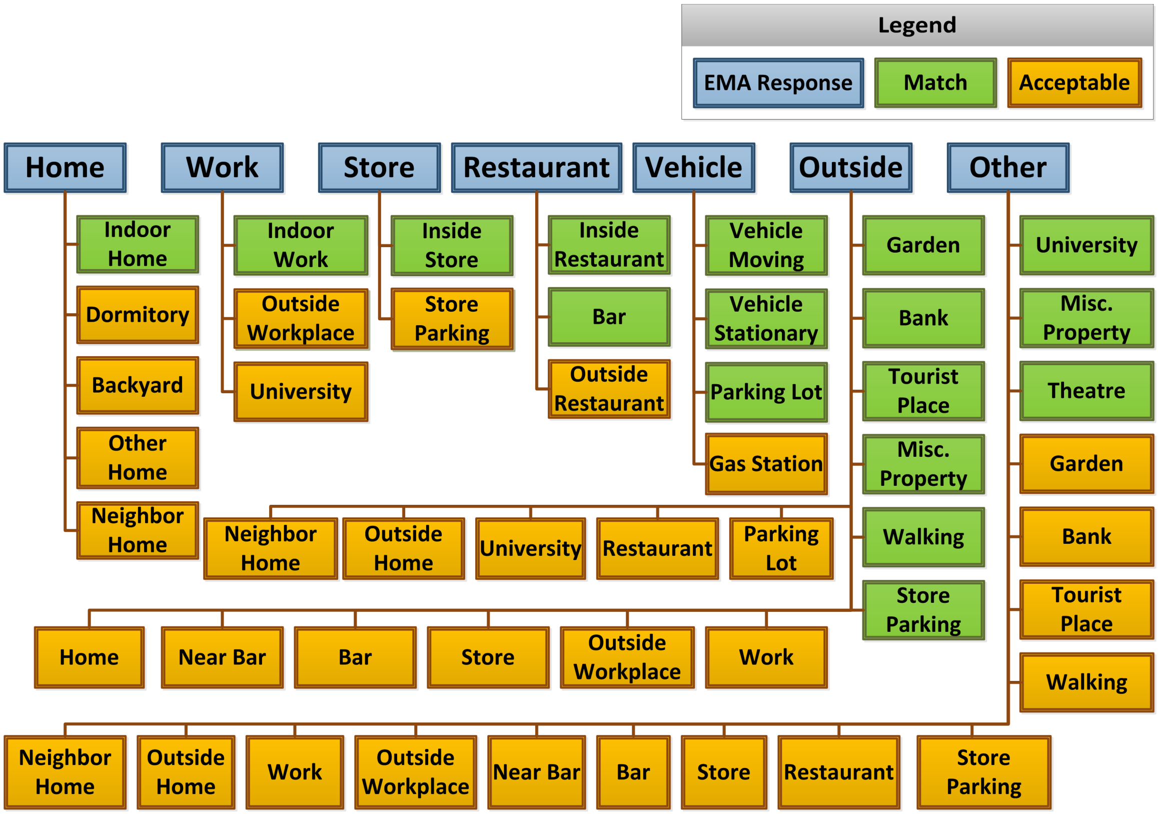


Fig. 4.3: Location mentioned in the top 7 boxes (blue) are possible responses of this item. There is a match if GPS log indicates that participant is in a location mentioned in green colored box and he reported a location mentioned in corresponding root blue box (ground truth location). Similarly, being at a location mentioned in yellow colored box and reporting a location mentioned in the corresponding root blue box was acceptable, and we mark their ground truth location as the one reported. Otherwise, we keep the report as it is.

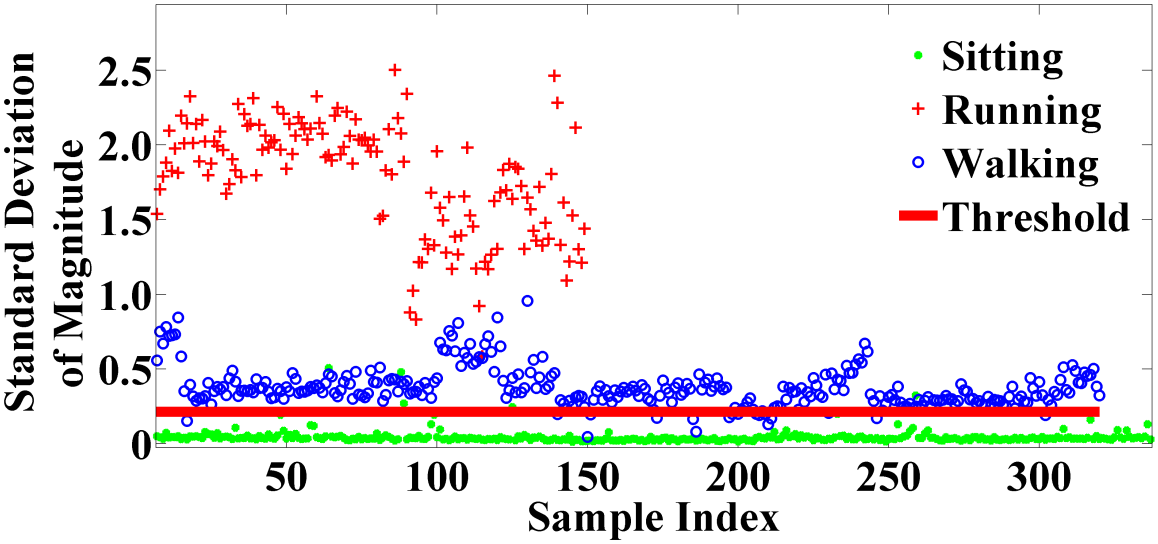


Fig. 4.4: Standard deviations < 0.21384 are labeled as stationary and others are labeled as non-stationary (i.e., walking or running).

case of computing alpha under specific conditions, we computed kappa when there were at least 10 EMAs that satisfied a chosen condition.

4.4 Context Inference

Participant's context can indicate whether a factor such as lack of attention may be responsible for unreliable responses. In this paper, we considered two such contexts, physical activity and physiological stress. These are computed from the accelerometer and physiological sensors (i.e., ECG and RIP) present in the wearable sensor suite. In both cases, we adapted existing inference algorithms.

4.4.1 Activity Inference

To infer whether a subject is in motion or not, we used a simple threshold based activity detector using the 3-axis on-body accelerometer (placed on chest). Phone accelerometer data was not used because the phone may not be on the person and thus may not indicate actual physical activity. We utilized the existing physical movement detection approach [40, 41] and adapted it to fit our study. As the placement of the accelerometer and the participant population is different from

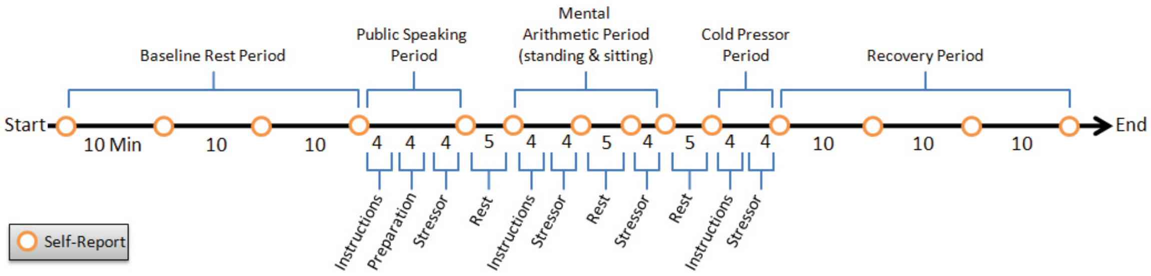


Fig. 4.5: The Lab Study Procedure for stress inference model. After 30 minute rest period participants receive 3 stressors, public speaking, mental arithmetic, and cold pressor.

that presented in the prior works, we collected training data to determine an appropriate threshold for detecting activity. We have collected labeled data under walking and running (354.16 minutes), and stationary (1426.50 minutes) states from seven pilot participants who wore the same sensor suite. Figure 4.4 shows the training data from seven pilot participants. We filtered the raw signal, removed the drift, and extracted the standard deviation of magnitude, which is independent of the orientation of the accelerometers and suggested by literature [40, 41]. We find the distinguishing threshold for our accelerometer to be 0.21384, which is able to distinguish stationary from non-stationary states with an accuracy of 97% in 10-fold cross-validation [30].

4.4.2 Stress Inference

Measurements from the ECG and RIP sensors were used to extract features for physiological stress model as proposed in [20]. To develop stress model a lab study was conducted. There are several factors that can influence physiological signals besides stress. Reasoning this during lab study participants were instructed to avoid caffeine, tobacco, alcohol, drug (e.g., pain killer), and excessive physical activity like exercise prior to lab session. Figure 4.5 summarize the lab session. At the beginning participants had a 30 minute rest period so that his physiology becomes at baseline. After that for ground truth of stress 21 participants were exposed to three well known stressors like, public speaking, mental arithmetic, and cold pressor [42, 43, 44, 45] with 5 minutes rest period in-between. For building model ECG features like RR interval, ratio between low and high frequency components of heart beat, and heart beat frequency in 3 bands (low, medium, and high) is used. As proposed in [20] we also considered respiration features like respiration duration, inhalation duration, exhalation duration, IE ratio, stretch, ratio of minute ventilation and minute volume, and breath rate. Support Vector Machine (SVM) [46] based model is able to classify stress at an accuracy of 89.17%. The model produces binary outputs, on 30 second segments of measurements. A correlation of 0.71 is observed between the stress model and the self-reported rating of stress. As proposed in [20], stress inference was discarded when the participant was detected to be not stationary from the activity inference.

Chapter 5

Results

In this section, we first present the reliability scores (alpha) and agreement scores (kappa) of the participants. Next, we take a closer look at the data obtained from the participants for whom we observe low agreement between self-reported location and the GPS-inferred location. We investigate whether this low agreement is due to one of the following reasons: (i) They are sensitive to the location question, (ii) They do not wish to report certain locations due to privacy concerns, or (iii) They are not fully available to respond accurately. We then examine whether participants with low kappa have low reliability of self-reports as measured using Cronbach's alpha. Finally, we examine whether contexts such as "stressed", "physical activity" or "away from home or at work" at the time of receiving an EMA prompt are associated with low reliability (i.e., $\alpha < 0.7$), as we speculate that during these contexts participants may not be fully available to respond carefully to EMA.

5.1 Participant's alpha score

Figure 5.1 presents the alpha score for each participant with an unacceptable alpha or kappa score. We observe that two participants (P#20 and P#44) had questionable alpha score. Upon closer examination of these participants' alpha scores in different location contexts, we observe that their alpha scores are low in some contexts (e.g., 0.16 when outside of home or work for P#20), but acceptable in other locations (e.g., 0.72 for P#20 when in home or work location). Therefore, no participant was always inconsistent and every participant reported reliably in at least at one context.

5.2 Participant's kappa score

Figure 5.2 shows κ scores for each participant, where five participants have absolute agreement with $\kappa = 1$ and eight participants have $\kappa < 0.7$. We

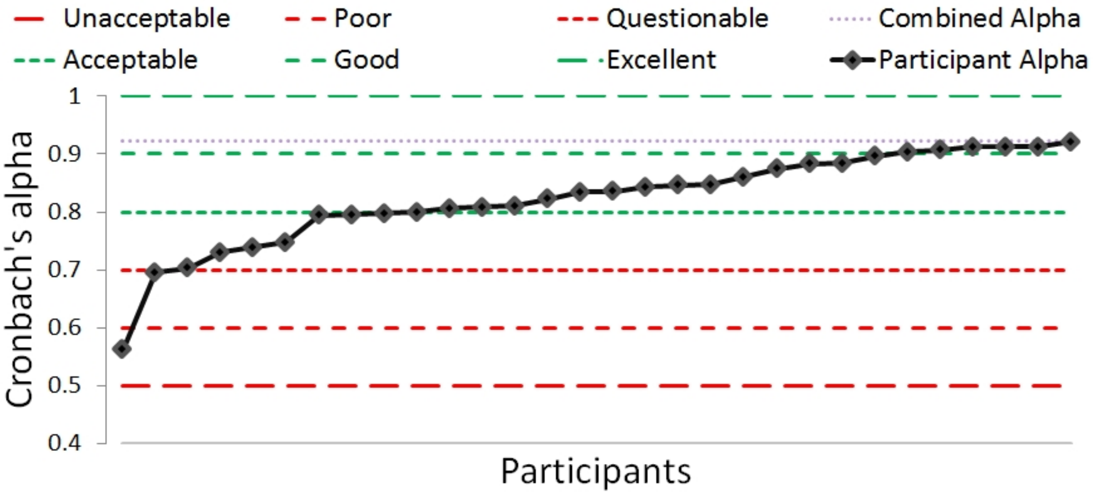


Fig. 5.1: Participant's Cronbach's alpha. Alpha is sub-sectioned into six groups [1], unacceptable(α <0.5), poor [0.5, 0.6), questionable [0.6, 0.7), acceptable [0.7, 0.8), good [0.8, 0.9), and excellent ($\alpha \geq 0.9$). We observe that only two participants have unacceptable alpha.

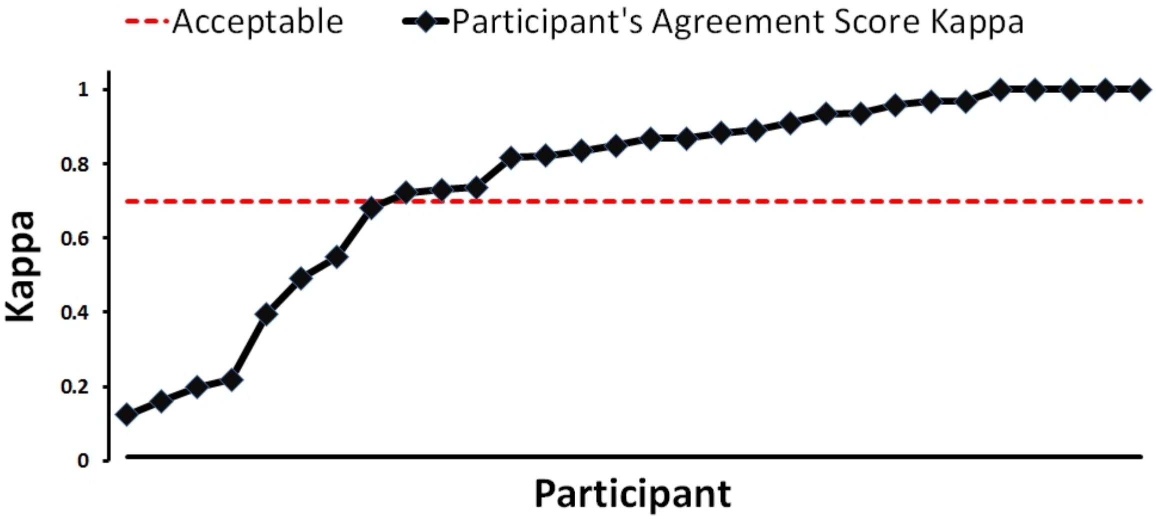


Fig. 5.2: Agreement score kappa (κ) measured from location inferred from GPS and participant's response to the location item. Overall κ for all participants is 0.78. Acceptable line is placed at 0.7.

observed wide variability among the participants. We find some participants for whom kappa score is 1, indicating that GPS-inference matched perfectly with self-reported location for these participants.

5.3 Relation between kappa score and alpha score

We first test the following alternate hypothesis.

Hypothesis 1. *Participants with unacceptable agreement between inferred location and participant's response to the location item, will have less consistency in other self-reported items than those who have agreement on location.*

H$_0$: $\alpha_{AcceptableKappa}$ = $\alpha_{UnacceptableKappa}$

H$_a$: $\alpha_{AcceptableKappa}$ >$\alpha_{UnacceptableKappa}$

We computed α for each of the 30 participants and divided them into two groups based on whether agreement of location item is acceptable ($\kappa \geq 0.7$) or not. Acceptable group had 22 participants with a mean of α=0.826 and σ=0.065, whereas the unacceptable group had eight participants with a mean of α=0.813 and σ=0.120. A two sample two-tail F-test showed that these two groups did not have equal variance (p-value 0.028). We therefore conducted a two sample one-tail Welch's t-test and estimated the p-value 0.384. Non-parametric Wilcoxon rank sum test show that there is no significant difference between these two groups with p-value 0.6549. Permutation test with statistics like mean and median we get p-values 0.323 and 0.523, respectively for 1000 random permutations. As a result, based on consensus in both parametric test and non-parametric test, we cannot reject the null hypothesis (H$_0$) and conclude that unacceptable agreement between participants' self-reported location and the one inferred from GPS does not imply less consistency in self-report.

Table 5.1: Context score for participants with unacceptable acceptance score of alpha or kappa. P# for participant id, α for Cronbach's alpha, κ for kappa agreement score, H+W for Participant being at Home or Work, and Other for any place other than Home or Work.

P#	α	κ	α		κ	
			H+W	Other	H+W	Other
16	0.85	0.55	0.85	0.68	0.67	0.20
17	0.91	0.68	0.88	0.83	0.49	1.00
20	0.56	0.16	0.72	-	0.16	-
24	0.73	0.20	0.75	-	0.20	-
25	0.91	0.49	0.91	0.86	0.54	0.21
28	0.82	0.22	0.80	-	0.20	-
30	0.81	0.12	0.83	0.46	0.08	0.35
31	0.91	0.40	0.92	0.82	0.32	0.59
44	0.69	0.87	0.73	-	1.00	-

To understand lack of strong association between low kappa and low alpha, we further analyzed the reliability scores for the eight participants who have unacceptable kappa (see Table 5.1). We observe that the alpha scores for all these participants, except one (P#20), are greater than 0.7. We take this as further corroboration that lack of agreement in location does not imply lack of care in completing self-reports. We also observe that kappa is unacceptable for these participants whether they were at home or at work (which we would expect not be a sensitive location to reveal) or outside of these standard places. We conclude that several of these participants may be sensitive to the location item itself, irrespective of where they may be located.

5.4 Role of Context in Assessing Reliability of Self-reports

We now examine the role of three contextual components in predicting the reliability of self-reports - location, activity, and stress.

5.4.1 Role of Location Context

Table 5.1 indicates that for two participants, the alpha scores when outside of home or work, is lower than 0.7. This indicates that for at least some

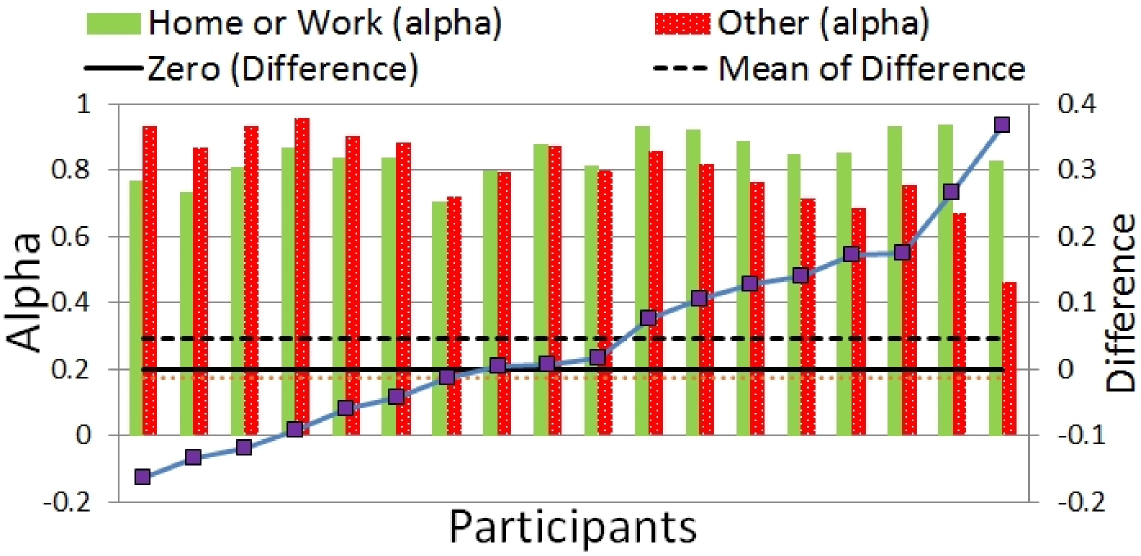


Fig. 5.3: Participant alpha in case of being at home or work, or other places response is plotted in primary axis. Difference of them for each participant is plotted in secondary axis. CI lower limit for the difference of samples of Hypothesis 2 is also plotted in secondary axis.

participants, when they are not at home or work, they may not be fully available to complete self-reports and hence their reports may have lower reliability.

To examine the role of location context, we formulate the following alternate hypothesis.

Hypothesis 2. *Participants will have lower reliability scores when they are not at home or work.*

H$_0$: $\alpha_{HomeWork}$ = α_{Other}

H$_a$: $\alpha_{HomeWork}$ >α_{Other}

Responses from all participants were again categorized - one category contained EMAs when the participant was at home or at work and the other category contained EMAs when the participant was in other locations (e.g., a store, a restaurant, in a vehicle, or outside). We calculated alpha for both of the categories for each participant and conducted a one-tail paired t-test over 18 sample pairs. Mean alpha of home or work, and other locations were 0.844 and 0.798, respectively, with a mean difference of 0.046. One tail paired t-test estimated p-value 0.095. For Non-parametric Wilcoxon Signed Rank Test we get p-value 0.123. As a result, we were not able to reject the null hypothesis (H_0) and conclude that participant's consistency does not get affected due to being in less frequent places, i.e., not at home or work, where participants spend 91.4% of their time. Figure 5.3 presents the samples used in this test.

5.4.2 Role of Activity Context

To examine the role of physical activity on reporting reliability, we formulate the following alternate hypothesis.

Hypothesis 3. *Participants will have lower consistency scores when they are engaged in physical activity.*

H_0: $\alpha_{Stationary} = \alpha_{Activity}$

H_a: $\alpha_{Stationary} > \alpha_{Activity}$

Physical activity is defined as participant performing an activity that has the intensity at-least equal to that of during taking a walk, while stationary refers to participant activity intensity level is less than walking. We calculate a pair of alpha for participant being stationary or being in some physical activity for each participant (following the same process described earlier). Our total sample pair was 24, mean alpha of stationary and activity was 0.792 and 0.743, respectively, and the mean of difference was 0.049. Two-tail paired t-test estimated p-value 0.313. Here, we are unable to reject the null hypothesis and report that physical activity has no effect on consistency of self-reports. However, one-tail paired t-test estimated p-value 0.156. For Non-parametric Wilcoxon Signed Rank Test we get p-value 0.265. We are unable to reject the null hypothesis of (H_0) and conclude that physical activity does not result in inconsistent self-reports. Figure 5.4

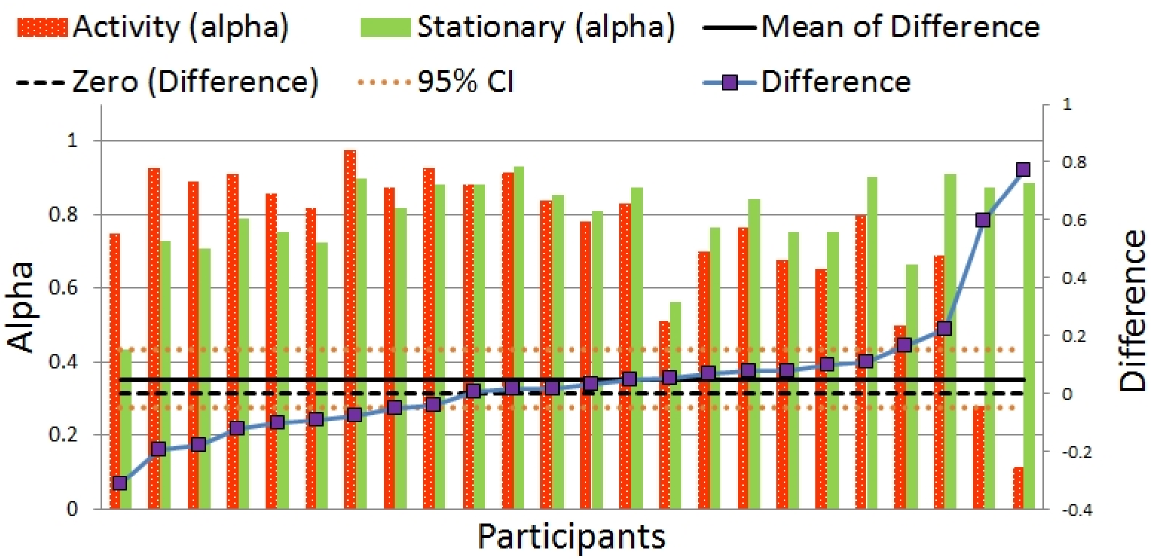


Fig. 5.4: Participant alpha in case of activity or being stationary while EMA response is plotted in primary axis. Difference of stationary and activity for each participant is plotted in secondary axis. CI for difference of samples of Hypothesis 3 is also plotted in secondary axis.

presents the samples used in this test. We hypothesize that during daily physical activity like walking, people are more likely in transitioning state in their daily life [47] and people have enough time to complete the self-report carefully and consistently.

5.4.3 Role of Psychological Context

Next, we investigated whether participants' psychological state can cause them to provide less consistent response. This may be the case when they are stressed and as a result may be too burdened psychologically to be fully available to focus on completing the self-report. We test the following alternate hypothesis to investigate the role of stress.

Hypothesis 4. *Participants will have lower consistency scores when they are stressed.*

H$_0$: $\alpha_{NotStressed}$ = $\alpha_{Stressed}$

H$_a$: $\alpha_{NotStressed}$ >$\alpha_{Stressed}$

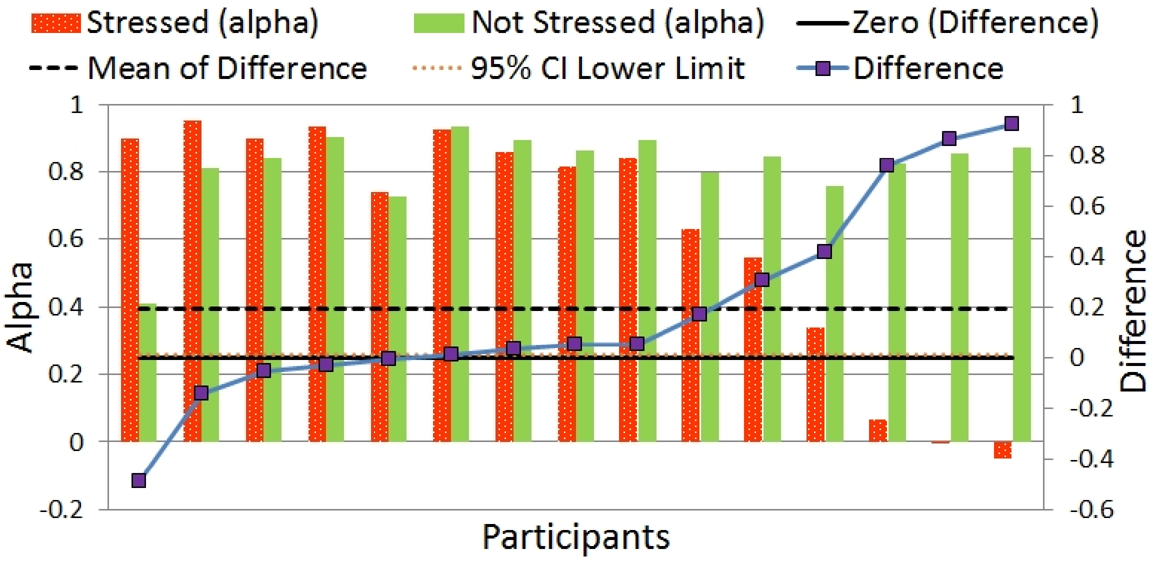


Fig. 5.5: Participant alpha in case of being stressed or not stressed response is plotted in primary axis. Difference of alpha of not stressed and stressed for each participant is plotted in secondary axis. CI lower limit for difference of samples of Hypothesis 4 is also plotted in secondary axis.

It is possible that prompting EMA can cause physiological stress to the participants. In order to select an unbiased binary stress indicator, we used five consecutive 30-second intervals preceding the delivery of a self-report prompt and considered the state of the individual to be stressed if they were found to be stressed in 3 out of the 5 preceding 30-second intervals. Since the stress model is a machine learning model, using 5 windows provides robustness of inference. We selected only those instances where the 30-second intervals preceding the EMA prompt did not have an activity episode that could interfere with stress model. Responses collected from each participant were categorized into two groups, “stressed” and “not stressed”. If a group contained less than 10 EMA, we excluded that participant from this computation. We calculated alpha for both stressed and not stressed groups for each participant and conducted a one-tail paired t-test over 15 sample pairs. Mean alpha of stressed and not stressed was 0.626 and 0.816, respectively, with a mean difference of 0.189. One-tail paired

Table 5.2: Hypothesis testing summary. H for Hypothesis number and n for sample count (pair). For each hypothesis both parametric paired t-test and non-parametric Wilcoxon Signed Rank test is performed.

H	n	Group 1 (α)		Group 2 (α)		Mean of Difference	p-value	
		Name	Mean	Name	Mean		t-test	Wilcoxon
2	18	At Home or Work	0.844	Other	0.798	0.046	0.095	0.123
3	24	Stationary	0.792	Activity	0.743	0.049	0.156	0.265
4	15	Not Stressed	0.816	Stressed	0.626	0.189	0.042	0.060

t-test estimated p-value 0.042. For Non-parametric Wilcoxon Signed Rank Test we get p-value 0.060. We reject null hypothesis (H_0). Figure 5.5 presents the samples used in this test. We hypothesized that due to cognitive impairment during a stressful episode [48] causes this inconsistency in self-report.

Table 5.2 summarizes the hypothesis testing results indicating the role of location, physical, and physiological contexts in self-report reliability.

Chapter 6

Discussion, Implications, and Limitations

Our results show that agreement between self-reported location and GPS-inferred location may not indicate the reliability of an entire self-report and self-report should not be discarded merely because the location report does not match the GPS-inferred location. A lack of agreement may be due to the sensitivity of some participants to the location question. This result also implies that since GPS collects data passively in the background, participants may not be consciously aware of their location being captured. Providing an option to stop the location capture based on predefined rules or retrospective erasure of location traces may be needed in future studies with GPS sensors.

Another way to use sensors to assess reliability of self-reports is by inferring the context of the participant at the time of receiving a prompt and testing if certain contexts should be avoided. Among the three contexts we examined (location, activity, and stress), we find that location (at the level of whether at home or work, or somewhere else) and activity (at the level of stationary or non-stationary) are not associated with low reliability of reporting. Hence, self-report prompts may still be delivered irrespective of the location and of physical activity status. Further research may be needed to investigate the role of these contexts at a deeper level, for example, by collecting sufficient data points at various locations where participants may be pressed for time (e.g., when driving) to investigate the role of context in predicting reliability of self-reports.

The third context we tested, namely stress, was associated with low reliability. This may be due to the participants' not being in the right state of mind to focus on completing the self-report. This may be similar to the case when participants fill out self-reports upon alcohol use or illicit drug use. These contexts are known to result in self-reports with less than desired reliability. It is interesting to note that such conditions, however, do not necessarily have effect on the compliance of the participants. We conclude, however, that unless the research protocol demands prompting the participants when they are under stress, self-report prompts should be deferred until the participant recovers from stress. Doing so may improve the reliability of self-reports.

Several other contexts, e.g., eating, using the phone, playing games, or working, where a person can be similarly mentally occupied as during stress, can have similar effect on reliability. This relates to interruptibility studies [49] whose goal has been to investigate interruptibility in work environments. Our results imply that interruptibility should also be investigated in more varied contexts, especially in the natural environment of the participants. Once such contexts are found, avoiding those contexts can improve the reliability of self-reports.

We would like to point out several limitations of our work. First, although we can detect driving episodes, a potential context that can affect reliability, we didn't have enough EMAs triggered during driving to statistically test its impact on reliability[1]. Second, inclusion of additional sensors such as microphones to identify conversation episodes, can uncover candidate contexts that can also predict reliability. Third, even though this work suggests that avoiding stressful events when prompting for self-report may improve reliability of self-reports, whether doing so indeed improves reliability needs to be investigated separately. Finally, participants for this study came from a mid-sized city in the U.S. and are students in a large university. The results obtained here may not generalize to the general population, other specialized population, or to other locations.

[1] Given the random nature of EMA[4], some did occur during driving. Participants were instructed to park the car to a safe place and answer the EMA.

Chapter 7

Conclusion

Self-report is the primary tool used in behavior research and social science to collect data from individuals in their natural environment. It is well known that self-report measurements are vulnerable to a wide range of problems such as subjective bias, physical condition of the participant, and social desirability of the response provided, to name a few. Limited and often expensive lab experiments have been the only methods available thus far to assess the accuracy of self-reports. In this paper, we attempt to use sensor data to assess reliability of self-reports. We show that disagreement between location reported and the location inferred from GPS is not an indicator of low reliability of the responses to the rest of EMA. Inferring the physical activity episodes from accelerometers we have identified that a person engaged in a physical activity such as walking is very much likely to reliably answer EMAs. We inferred stressful episodes from ECG and RIP sensor measurements and found that, unlike activity, stress does have an effect on reliability. Our findings indicated that such mental states where an individual is not fully (physically, mentally or emotionally) available, can have an adverse effect on the reliability of self-reports. Knowledge of such states or contexts can be very useful in designing future studies, especially EMA prompting mechanisms.

REFERENCES

[1] D. George and M. Mallery, “Using spss for windows step by step: a simple guide and reference,” 2003.

[2] J. Krumm and D. Rouhana, “Placer: Semantic place labels from diary data,” in *ACM UbiComp*, 2013, pp. 163–172.

[3] S. Kumar, W. Nilson, M. Pavel, and M. Srivastava, “Mobile health: Revolutionizing healthcare through trans-disciplinary research,” *IEEE Computer*, vol. 46, no. 1, pp. 28–35, 2013.

[4] S. Shiffman, A. Stone, and M. Hufford, “Ecological momentary assessment,” *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[5] A. Stone and J. Broderick, “Real-time data collection for pain: Appraisal and current status,” *Pain Medicine*, vol. 8, no. s3, pp. S85–S93, 2007.

[6] M. Oorschot, T. Kwapil, P. Delespaul, and I. Myin-Germeys, “Momentary assessment research in psychosis.” *Psychological assessment*, vol. 21, no. 4, pp. 498–505, 2009.

[7] D. Moskowitz and S. Young, “Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology,” *J Psychiatry Neurosci*, vol. 31, no. 1, pp. 13–20, 2006.

[8] C. Scollon, C. Kim-Prieto, and E. Diener, “Experience sampling: Promises and pitfalls, strengths and weaknesses,” *Journal of Happiness studies*, vol. 4, no. 1, pp. 5–34, 2003.

[9] F. Boca and J. Noll, “Truth or consequences: the validity of self-report data in health services research on addictions,” *Addiction*, vol. 95, no. 11s3, pp. 347–360, 2000.

[10] K. Wilson, R. Hopkins, M. deVries, and J. Copeland, “Research alliance and the limit of compliance: experience sampling with the depressed elderly,” *The experience of psychopathology: Investigating mental disorders in their natural settings*, pp. 339–346, 1992.

[11] S. Shiffman, “Ecological momentary assessment (ema) in studies of substance use.” *Psychological Assessment*, vol. 21, no. 4, p. 486, 2009.

[12] M. Hufford, A. Shields, S. Shiffman, J. Paty, and M. Balabanis, “Reactivity to ecological momentary assessment: an example using undergraduate problem drinkers.” *Psychology of addictive behaviors*, vol. 16, no. 3, p. 205, 2002.

[13] F. Serre, M. Fatseas, R. Debrabant, J. Alexandre, M. Auriacombe, and J. Swendsen, “Ecological momentary assessment in alcohol, tobacco, cannabis and opiate dependence: A comparison of feasibility and validity,” *Drug and alcohol dependence*, vol. 126, no. 1, pp. 118–123, 2012.

[14] H. Minami, D. McCarthy, D. Jorenby, and T. Baker, “An ecological momentary assessment analysis of relations among coping, affect and smoking during a quit attempt,” *Addiction*, vol. 106, no. 3, pp. 641–650, 2011.

[15] G. Alpers, “Ambulatory assessment in panic disorder and specific phobia.” *Psychological Assessment*, vol. 21, no. 4, p. 476, 2009.

[16] S. J. Wenze and I. W. Miller, “Use of ecological momentary assessment in mood disorders research,” *Clinical psychology review*, vol. 30, no. 6, pp. 794–804, 2010.

[17] R. Marije, K. Hogenelst, and R. Schoevers, “Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies,” *Clinical Psychology Review*, vol. 32, no. 6, pp. 510–523, 2012.

[18] J. Bland and D. Altman, “Statistics : notes cronbach’s alpha,” *BMJ*, vol. 314, no. 7080, pp. 572–572, 1997.

[19] J. Cohen *et al.*, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[20] K. Plarre, A. Raij, S. Hossain, A. Ali, M. Nakajima, M. Al’absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, *et al.*, “Continuous inference of psychological stress from sensory measurements collected in the natural environment,” in *IEEE IPSN*, 2011, pp. 97–108.

[21] F. Wilhelm and P. Grossman, “Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment,” *Biological Psychology*, vol. 84, no. 3, pp. 552–569, 2010.

[22] S. Komarov, K. Reinecke, and K. Gajos, “Crowdsourcing performance evaluations of user interfaces,” in *ACM CHI*, 2013, pp. 207–216.

[23] K. Crowston, “Amazon mechanical turk: A research tool for organizations and information systems scholars,” in *Shaping the Future of ICT Research. Methods and Approaches*, 2012, pp. 210–221.

[24] A. Kittur, E. Chi, and B. Suh, “Crowdsourcing user studies with mechanical turk,” in *ACM CHI*, 2008, pp. 453–456.

[25] D. Zhu and B. Carterette, “An analysis of assessor behavior in crowdsourced preference judgments,” in *ACM SIGIR*, 2010, pp. 17–20.

[26] S. Prince, K. Adamo, M. Hamel, J. Hardt, S. Gorber, and M. Tremblay, “A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review,” *International Journal of Behavioral Nutrition and Physical Activity*, vol. 5, no. 1, p. 56, 2008.

[27] A. Möller, M. Kranz, B. Schmid, L. Roalter, and S. Diewald, “Investigating self-reporting behavior in long-term studies,” in *ACM CHI*, 2013, pp. 2931–2940.

[28] K. Elgethun, M. Yost, C. Fitzpatrick, T. Nyerges, and R. Fenske, “Comparison of global positioning system (gps) tracking and parent-report diaries to characterize children’s time–location patterns,” *Journal of Exposure Science and Environmental Epidemiology*, vol. 17, no. 2, pp. 196–206, 2006.

[29] P. Stopher, C. FitzGerald, and M. Xu, “Assessing the accuracy of the sydney household travel survey with gps,” *Transportation*, vol. 34, no. 6, pp. 723–741, 2007.

[30] M. Rahman, R. Bari, A. Ali, M. Sharmin, A. Raij, K. Hovsepian, S. Hossain, E. Ertin, A. Kennedy, D. Epstein, K. Preston, M. Jobes, G. Beck, S. Kedia, K. Ward, M. al’Absi, and S. Kumar, “Are we there yet? feasibility of continuous stress assessment via wireless physiological sensors,” 2014.

[31] M. Musthag, A. Raij, D. Ganesan, S. Kumar, and S. Shiffman, “Exploring micro-incentive strategies for participant compensation in high-burden studies,” in *ACM UbiComp*, 2011, pp. 435–444.

[32] G. Mark, D. Gudith, and U. Klocke, “The cost of interrupted work: more speed and stress,” in *ACM CHI*, 2008, pp. 107–110.

[33] D. McFarlane, “Comparison of four primary methods for coordinating the interruption of people in human-computer interaction,” *Human-Computer Interaction*, vol. 17, no. 1, pp. 63–139, 2002.

[34] J. Ho and S. Intille, “Using context-aware computing to reduce the perceived burden of interruptions from mobile devices,” in *ACM CHI*, 2005, pp. 909–918.

[35] S. Iqbal and B. Bailey, “Effects of intelligent notification management on users and their tasks,” in *ACM CHI*, 2008, pp. 93–102.

[36] B. Poppinga, W. Heuten, and S. Boll, “Sensor-based identification of opportune moments for triggering notifications,” *IEEE Pervasive Computing*, vol. 13, no. 1, pp. 22–29, 2014.

[37] D. Bonett, “Sample size requirements for testing and estimating coefficient alpha,” *Journal of educational and behavioral statistics*, vol. 27, no. 4, pp. 335–340, 2002.

[38] “Sample size for estimating a single alpha program,” https://www.statstodo.com/SSiz1Alpha_Pgm.php, Accessed: November 2014.

[39] R. Montoliu, J. Blom, and D. Gatica-Perez, “Discovering places of interest in everyday life from smartphone data,” *Multimedia Tools and Applications*, vol. 62, no. 1, pp. 179–207, 2013.

[40] P. Pandian, K. Mohanavelu, K. Safeer, T. Kotresh, D. Shakunthala, P. Gopal, and V. Padaki, “Smart vest: Wearable multi-parameter remote physiological monitoring system,” *Medical engineering & physics*, vol. 30, no. 4, pp. 466–477, 2008.

[41] L. Atallah, B. Lo, R. King, and G. Yang, “Sensor placement for activity detection using wearable accelerometers,” in *Body Sensor Networks (BSN)*, 2010, pp. 24–29.

[42] M. al’Absi, S. Bongard, T. Buchanan, G. Pincomb, J. Licinio, and W. Lovallo, “Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors,” *Psychophysiology*, vol. 34, no. 3, pp. 266–275, 1997.

[43] M. al’Absi, T. Buchanan, and W. Lovallo, “Pain perception and cardiovascular responses in men with positive parental history for hypertension,” *Psychophysiology*, vol. 33, no. 6, pp. 655–661, 1996.

[44] M. al’Absi, D. Hatsukami, and G. Davis, “Attenuated adrenocorticotropic responses to psychological stress are associated with early smoking relapse,” *Psychopharmacology*, vol. 181, no. 1, pp. 107–117, 2005.

[45] M. al’Absi, K. Petersen, and L. Wittmers, “Adrenocortical and hemodynamic predictors of pain perception in men and women,” *Pain*, vol. 96, no. 1-2, pp. 197–204, 2002.

[46] J. Platt, “Fast training of support vector machines using sequential minimal optimization,” in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.

[47] H. Sarker, M. Sharmin, A. Ali, M. Rahman, R. Bari, S. Hossain, and S. Kumar, “Assessing the availability of users to engage in just-in-time intervention in the natural environment,” in *ACM UbiComp*, 2014, pp. 909–920.

[48] C. Sandi, “Stress, cognitive impairment and cell adhesion molecules,” *Nature Reviews Neuroscience*, vol. 5, no. 12, p. 917, 2004.

[49] J. Fogarty, S. Hudson, and J. Lai, “Examining the robustness of sensor-based statistical models of human interruptibility,” in *ACM CHI*, 2004, pp. 207–214.