Using Wearable Sensors and Real Time Inference to Understand Human Recall of Routine Activities

Predrag Klasnja[1], Beverly L. Harrison[1], Louis LeGrand[1], Anthony LaMarca[1], Jon Froehlich[2], Scott E. Hudson[3]

[1]Intel Research Seattle
Seattle, WA 98105, USA
klasnja@u.washington.edu,
[beverly.harrison, louis.l.lagrand,
anthony.lamarca]@intel.com

[2]Computer Science & Engineering
DUB Group
University of Washington
Seattle, WA 98195, USA
froehli@cs.washington.edu

[3]HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
scott.hudson@cs.cmu.edu

ABSTRACT

Users' ability to accurately recall frequent, habitual activities is fundamental to a number of disciplines, from health sciences to machine learning. However, few, if any, studies exist that have assessed optimal sampling strategies for *in situ* self-reports. In addition, few technologies exist that facilitate benchmarking self-report accuracy for routine activities. We report on a study investigating the effect of sampling frequency of self-reports of two routine activities (sitting and walking) on recall accuracy and annoyance. We used a novel wearable sensor platform that runs a real time activity inference engine to collect *in situ* ground truth. Our results suggest that a sampling frequency of five to eight times per day may yield an optimal balance of recall and annoyance. Additionally, requesting self-reports at regular, predetermined times increases accuracy while minimizing perceived annoyance since it allows participants to anticipate these requests. We discuss our results and their implications for future studies.

Author Keywords

User study, Empirical evaluation, ESM, experience sampling method, self-reports, recall accuracy, survey frequency.

ACM Classification Keywords

H5.2 [Information interfaces and presentation]: User Interfaces

INTRODUCTION

Participants' ability to accurately recall their daily activities is central to a number of disciplines. In health sciences, health benefits from physical activity have primarily been established on the basis of self-report data, typically surveys asking people to recall physical activity that they performed in the last week or two weeks [19]. Information from such studies has established a strong connection between activity and health [22]. We now know that physically active people have lower levels of mortality from all causes [15], as well as significantly lower prevalence of a number of specific health conditions including diabetes [13], hypertension [18], stroke [3], osteoporosis [4], depression, and anxiety [20]. Some recent studies suggest that even low intensity activity such as walking or housework can have health benefits [6,11]. However, the data from self reports (or sometimes from observer reports based on labor intensive human shadowing throughout the day) suffer from accuracy problems making the reliability of such data questionable. For instance, researchers suggest that for frequent activities people tend to over-report physical activity and underreport sedentary behaviors [21]. There are few, if any, methods of determining the extent of these inaccuracies for real world behaviors since this relies upon having high quality "ground truth" against which the self report data can be assessed. (Note that this domain is unlike studies of technology-related activities which can leverage highly accurate objective data such as computer logs or cell phone usage logs,) Developing better self-report measures that can be deployed to large populations (and the ability to potentially quantify accuracy) is crucial to answering researchers' call for better assessment of low intensity activities as well as a more direct assessment of sedentary behaviors [17,21].

The need for accurate recall of frequent, habitual activities is not limited to health sciences. In machine learning, statistical models are usually trained using accurately annotated training data sets [5]. These data, typically referred to as "ground truth," are created by either manually labeling video post hoc or by having participants label activities in real time (or after a short time delay). As

UbiComp'08, September 21-24, 2008, Seoul, Korea.

automatic inference moves into the domain of everyday behaviors, researchers increasingly have to rely on user labeling for the collection of training data. Collection of accurate data sets therefore depends on people's ability to correctly remember their behavior.

Our ability to recall events accurately depends on a number of factors, including salience of the event, social pressures, the time that has elapsed between an event and when we try to recall it, biases in human perception of time (i.e., 'internal clock'), etc. [1,9,21]. Habitual activities are particularly problematic due to both their high frequency and low salience (i.e., there is nothing unusual or atypical to make them stand out). While rare or particularly novel events are easily remembered, things that occur frequently and are an integral part of our routines are much more difficult to accurately recall. Trying to remember how much time one spent sitting during the previous day, or using email or the web illustrates the difficulty. Studies of physical activity have found that self-reports of such frequent, low intensity data correlates only modestly with objective measures of activity [19]. Yet it is precisely such activities that are often of central interest to researchers.

This study investigates how to help people remember frequent, habitual activities more accurately without imposing excessive burden on study participants. Our strategy is to ask about target activities using short phone-based *in situ* surveys which are repeated several times a day. We examine optimal rates of survey presentation by varying survey frequency and measuring recall by asking participants to report how much they have performed target activities in the period since the last survey. Participants' responses are compared to ground truth data obtained from wearable sensors. We examine recall accuracy, subjective perception of annoyance and recall difficulty as the number of surveys and their timing changes. We suggest a way to use this technique to maximize recall accuracy while minimizing intrusiveness and annoyance. While this technique cannot control for every factor influencing human estimates of time, it does provide fundamental benchmarking for self-report and survey strategies under conditions of everyday memory biases.

Addressing the need in health sciences, in this study we focus on low intensity physical activity and sedentary behaviors—namely, walking and sitting. We suggest, however, that the same method could be useful for assessing other habitual activities as well. In particular, studies of other hard-to-track frequent activities such as multi-tasking, eating and snacking, and dysfunctional habits such as interrupting or aggressive verbal behavior could benefit from this method.

RELATED WORK

With their emphasis on large populations, epidemiological studies have traditionally relied on surveys for physical activity assessment, as have the majority of studies in other health sciences [19]. And while the use of surveys has a number of advantages, the validity of survey measures varies significantly for various types of physical activity. Ainsworth and colleagues [2] write that "activities that are easier to recall, such as vigorous activities or special planned activities, correlate well with direct measures of such activities. However, activities performed at light and moderate levels of intensity, or activities performed daily, correlate less well with direct measures of physical activity" (see p. 613 for details).

Adams et al [1] found that over-reporting occurs both for light and for moderate activities, partly as a result of social desirability and social approval bias. One of the reasons for this, Adams et al believe, is that those people who are prone to over-reporting are more likely to do so with activities that they do fairly often than with activities they perform only rarely. Durante and Ainsworth's [9] findings are consistent with these results. In their review of the literature on survey accuracy, they found that while the recall of hard and very hard activities is accurate, the recall of moderate activities is poor, across all tested domains (occupation, leisure, and home). Durante and Ainsworth note that in addition to social desirability, factors influencing recall accuracy of habitual activities include the lack of salience of such activities, their frequency, and the lapse of time between the activity and the recall attempt.

To overcome issues with post hoc self-reports (e.g., surveys, journals), the experience sampling method (ESM) has been developed for *in situ* recoding [8,12]. ESM aims to reduce recall errors by asking people questions *in situ*, either by triggering questionnaires by an occurrence of an event of interest (context-triggering), or by prompting people to answer questionnaires throughout the day at random or predetermined times (interval- and signal-contingent triggering) [23].

While ESM originally relied on paper surveys, it is now possible to conduct ESM using PDAs and mobile phones [10,14]. In HCI, the ability of contemporary ESM applications to trigger surveys based on sensor readings and a variety of other contextual factors has made this method particularly useful for evaluating ubicomp technology [7]. We suggest that phone-based ESM is also well suited as a method for collecting data about routine activities such as those of interest to health sciences. As commodity mobile phones become more and more powerful, the use of phone-based ESM in even large epidemiological studies becomes a real possibility.

METHODS AND PROCEDURES

Participants

Twenty participants took part in our study, chosen to represent a variety of professional profiles. (The study was conducted as a within subject, repeated measures design.) Our sample included five homemakers, two retail workers, one dancer, one waiter, four students (we later determined that one student was a part-time tight rope walk instructor), and seven office workers. The latter group included researchers, engineers, a public health professional, and an environmental scientist. We hypothesized that office workers would tend to be mostly sedentary and might often have very predictable schedules, while homemakers, retail workers, and people working in restaurants would be characterized by much more walking and higher variability from day to day. We wanted to include and analyze the activity recall differences represented by people that have dramatically different activity profiles and professions.

Of the twenty participants that completed the study, two—one homemaker and the dancer—experienced technical failures that resulted in a loss of substantial portions of their data. Consequently, their data was excluded from the analysis. All reported results are for the remaining eighteen participants. All 20 participants were compensated for their participation.

Equipment

Ground truth activity data was collected using the Mobile Sensor Platform (MSP) [16], a pager-size device worn clipped at the waist that uses 10 sensors to detect a wide range of physical activities, including walking, sitting, bicycling, running, and the use of exercise equipment like stair machines and elliptical machines. In previous work, the MSP has been shown to detect such physical activities with about 85% accuracy [16]. For this study, we re-trained and tuned the activity inference models to increase the accuracy for detecting walking and sitting using labeled data from 12 individuals of varying ages, heights, and gender (none were study participants). There were approximately 11,000 labeled samples per activity of featurized sensor data produced at quarter second intervals. Testing of the model accuracy was done using a per-person holdout approach: 12 models were created, and each was trained on the data from 11 persons and tested on the 12th. This holdout strategy was done for all holdout combinations (i.e., each of 12 people used in turn to test against the model from other 11 people) and the resulting accuracy is reported here. Cascading Naïve Bayes Models were used (the results were comparable or better than Hidden Markov Models tested in previous versions of the system [e.g., 16]). The resulting accuracy for detecting sitting/sedentary activity was 96%, and for walking was 93%. Adding more samples or more participants did not further improve the already high accuracy. These models performed at this level in a participant-independent manner (proviso "normal walking", i.e., no crutches, significant visible limp, etc. which we did not test for) and thus we used the MSP to automatically log what we then considered the "actual" walking and sitting data for all our participants. While this may introduce some error into walking and sitting "scores", large differences in subjective recall accuracy can still be easily identified. To our knowledge, this technique is still one of the most accurate methods available for tracking real *in situ* physical activity (alternative best practices are shadowing, self-reports, or pedometers –all of which are more error-prone).

Participants wore the MSP and carried a cell phone throughout the day for 8 typical workdays. The MSP device performed real time embedded activity inference, computed four times per second. These were transmitted over Bluetooth to the mobile phone and stored on the phone's storage card for later post hoc processing. At the end of the study, the quarter-second inferences were processed to smooth them into human scale activity "episodes" that precisely corresponded to the activity definitions/rules given to the participants at the beginning of the study (e.g., how many times did you sit for more than 10 minutes without getting up? How many times did you walk for more than 2 minutes?).

Self-report data was collected using Cingular 2125 mobile phones running Windows Mobile 5 operating system. Surveys were created using the MyExperience toolkit [10], a toolkit for authoring ESM surveys. MyExperience was embedded in a custom application that randomly assigned a different survey schedule for each day of the study.

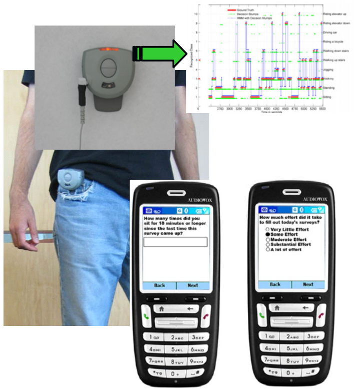
AUDIOVOX
How many times did you sit for 10 minutes or longer since the last time this survey came up?
Back
Next
AUDIOVOX
How much effort did it take to fill out today's surveys?
○ Very Little Effort
● Some Effort
○ Moderate Effort
○ Substantial Effort
○ A lot of effort
Back
Next

Figure 1. The Mobile Sensing Platform (pager sized wearable sensor platform) and ESM cell phone-based survey tool.

Study procedures

The study protocol lasted eight typical workdays. Study days did not have to be consecutive, and participants were instructed not to run the study on days when they knew that their schedules would be atypical (e.g., if they were going to spend much of the day in an airplane, for example), or on days where they could not be interrupted. By focusing on typical workdays (e.g., Monday to Friday for office workers) this helped to ensure that surveys would not be skewed by dramatic changes in activity patterns that often occur on weekends. This gave us consistent patterns of data across scheduling conditions (see analysis later).

On study days, participants performed three main tasks: (1) they wore the MSP from the time they started their day in the morning until seven o'clock in the evening; (2) they answered an eight question activity survey that came up throughout the day a varying number of times according to each day's survey scheduling condition; and (3) they answered an evening survey that asked them about their experiences with that day's survey schedule.

An activity survey consisted of eight questions that were repeated in the same order every time the survey came up. The survey asked (1) *how many times* the participant performed the target activity, what were the (2) *longest* and (3) *shortest* episodes of the activity, and (4) the *total time* spent performing the target activity. The questions were first asked for walking and then for sitting. Each survey asked for this information for the period of time since the previous survey or, for the first survey of the day, since the participant started wearing the MSP device that morning. This meant that the length of the time for which participants needed to remember their activity was inversely proportional to the frequency of surveys for a given day. The more frequent the surveys, the shorter the time for which participants had to remember their activity.

Our phone application randomized the order of survey schedules for each participants. We did this to mitigate ordering and learning effects on recall accuracy.

At seven o'clock every evening an end-of-the-day survey came up on the participant's mobile phone. The survey asked participants to rate on a five point Likert scale how accurately they thought they remembered their activity that day, how difficult they found it to do so, and how annoying they found that day's schedule of surveys to be. The survey also reminded participants to fill out a short paper form that asked them what strategies they used that day to remember their activity, and whether they thought that the day's survey schedule helped them in this task. The participants were also asked to draw a rough diagram on the form, illustrating what they thought their activity pattern for that day looked like.

Defining Activity Episodes

Participants were not asked to remember every single moment of sitting and walking during their day. At the beginning of the study they were given definitions of what constituted an episode of sitting and walking, and were asked to only track episodes of these activities that matched the definitions. Our activity definitions were as follows: An episode of walking is two minutes or longer, with a possible break of up to 1 minute (which would account for the street light changes we timed, for instance). An episode of sitting is defined as being ten minutes or longer, with a possible interruption of up to 90 seconds. Both walking and sitting could have more than one brief interruption, but any interruption longer than the critical value would break the activity into two episodes. Participants were given examples of these prior to the experiment.

The definitions themselves were designed with two criteria in mind. First, they were supposed to make it easier to remember one's activity by making it unnecessary for participants to recall very brief instances of walking and sitting. Second, definitions tried to mimic how we normally think of activities in everyday life, allowing for short breaks that do not interrupt an episode of an activity. For example, we often think of walking to the grocery store as one walk even though we might have to stop several times at traffic lights. Similarly, if we are sitting at our desk we can briefly get up to get a stapler or retrieve a printout and still think that we sat for, say, two hours. Our activity definitions try to accommodate this way of thinking about continuous activity by allowing brief interruptions to be seen as a part of a longer episode of walking or sitting. Finally, our definitions made a minimum length of sitting longer than the minimum length of walking since sitting episodes generally seemed to be longer and less "bursty" than walking (i.e., we rarely sit for 1 minute only, then get up for a minute, then sit for one more minute, get up again, etc.). We did this in order to make it easier for our participants to track how much they sat, but also to try to capture relatively short episodes of walking that pilot testing suggests might be the only "walks" that an office worker gets during a day.

Survey Scheduling Conditions

The study compared six different survey schedules: once per day at the end of the day (mimicking traditional end of day journaling), 3 times per day, 5 times per day, 8 times per day, 12 times per day (once every hour), and 20 times per day. All participants performed all survey schedules. The order of survey schedules was randomized for each participant.

We used two types of schedules: pseudo-random and fixed. For the three pseudo-random schedules—5 times a day, 8 times a day, and 20 times a day—there were minimum and maximum thresholds on survey frequency. Surveys were guaranteed to be at least 10 minutes apart and we ensured a "reasonable" distribution over the full day by dividing the day into <n> roughly equal length periods (where <n> is the total number of surveys scheduled), and triggering one survey at a pseudo-random time during each one of these time periods. The remaining three schedules—one time per day, three time per day, and the hourly schedule—were fixed schedules, based upon our interest in comparing to standard self-report practices. On these schedules, surveys came at predetermined times. The once per day survey came up at 6:50 pm, three times per day surveys came up at 10 am, 2:30 pm, and 6:50 pm, and the hourly surveys came up every hour on the hour between 9 am and 7 pm. The first two of these fixed schedules represent traditional daily journaling instruments used in a large body of literature. The hourly schedule, commonly used in journaling practice, was chosen both for its frequency and its regularity. We hypothesized that this regularity might potentially make it easier to remember activity patterns.

Given these characteristics, pseudo-random schedules allowed us to test intermediate values for sampling frequencies, while introducing a small element of unpredictability that would make it more difficult for participants to remember exactly when the last survey came up and to predict exactly when the next one will occur. We hypothesized that this unpredictability would potentially influence the perceived difficulty of recall and the perceived annoyance with the survey schedule.

Finally, to increase the number of data points for low frequency intervals, both the one-time per day and the three-time per day surveys were repeated twice during the study for each participant. In this way, if a participant accidentally missed hearing or responding to the survey for those days we would have some redundancy in the data rather than missing a day's worth of data entirely.

Hypotheses

We anticipated six possible outcomes from our study (stated as assertions rather than null hypotheses):

H1: Recall accuracy would improve with an increase in survey frequency since participants would need to recall their activity for shorter periods of time.

H2: Participant annoyance would increase with the increase in survey frequency. We expected the annoyance levels to sharply increase at some critical point. Anecdotal results from one of our past ESM studies, unrelated to this work, suggest that 13 times per day was the maximum for the number of interruptions a participant would tolerate [10] .

H3: Less frequent survey schedules would be rated as more difficult than frequent surveys since participants must work harder to remember or reconstruct their day or time interval.

H4: Random schedules would be perceived as being more difficult than fixed schedules due to their unpredictability.

H5: Based on the health sciences literature (e.g., [1]), we expected participants would: (a) overestimate how much they walk; (b) underestimate how much they sit.

H6: Due to more predictable schedules, office workers would likely remember their activity more accurately than participants whose daily activity is more variable.

Analysis Method

The MSP device passively logged data throughout the day for the entire time it was worn (roughly 7 am to 7 pm). However, participant ESM survey responses about their perceived activity times referred to only the time segment since the last survey. We therefore processed the MSP activity data to likewise partition it into exactly matching time-segments to facilitate comparison with participant responses. For example, for 5 times per day schedule, this gave us 5 time-segments with matched starting and ending points for the MSP data that exactly reflected when the 5 surveys were sent to participants. For each time-segment, we used the difference between total activity duration as reported by the MSP and the total duration reported by participants as the main measure of recall accuracy (*Recall Error*). This measure was chosen for two reasons: First, the total activity duration is the measure that is most independent of the particular definitions of an episode of walking and sitting used in this study. While the count of episodes of walking and sitting depended on the participant's ability to accurately gauge whether a particular break—e.g., stopping at a traffic light—was sufficiently long to break the activity into two episodes, this was far less an issue for total duration. Second, total activity duration is the measure most relevant in domains like health sciences. This measure allowed us to compare our results to health science literature, and examine the effects of the ESM schedule on how much people overestimate their walking and underestimate their sitting.

All statistical analyses, both of accuracy data and of subjective measures, were done within subject. For the analyses of recall accuracy, we used a Mixed Model that used absolute Recall Error as the dependent variable, Survey Schedule, Occupation and their two-way interaction as fixed effects, and the Participant ID as a random effect. For the analysis of subjective measures, we used the Friedman test. To make the Friedman test work, each participant's two ratings for Fixed 1 and Fixed 3 schedules were averaged to get one rating needed to create a complete block design. In addition to the statistical analyses, qualitative interview data were analyzed for main themes to supplement quantitative results.

RESULTS

Effect of Survey Schedules on Recall Accuracy

The number of daily surveys had a significant effect on the accuracy of recall of both walking and sitting. Figure 2 shows the magnitude of the recall error as a function of the daily survey schedule. As hypothesized, as the number of surveys increased, the recall accuracy became substantially better (i.e., recall error was significantly lower). The effect of schedule was highly statistically significant. Mixed Model analysis revealed a strong main effect of Survey Schedule on the magnitude of the Recall Error for both sitting and walking (for sitting, $F[5, 663]=96.25$, $p < .001$; for walking, $F[5, 658]=36.76$, $p < .001$).

Note that in Figure 2 (top), the Fixed 12 condition appears to have an unanticipated increase in recall error (contrary to expectation). A closer investigation of the data explained this unexpected result. The increased error/accuracy drop is caused by outlier data from two participants: a waiter who found the hourly schedule to be highly disruptive to his work and therefore was unable to respond to a number of the surveys, and the student/tight rope walking instructor who reported significantly more walking than the MSP detected on the 12-survey day. The bulk of the walking on the day in question for this participant was done on the wire which, not surprisingly, our device did not correctly detect. (We initially recruited this participant without realizing that he was a part-time tight rope walk instructor.) When these two participants were excluded from the data set, the relationship between walking recall accuracy and daily survey schedules closely approximated what we saw for sitting (Figure 3, adjusted). However, our analysis was done including these outlier data and was still statistically significant as reported above.

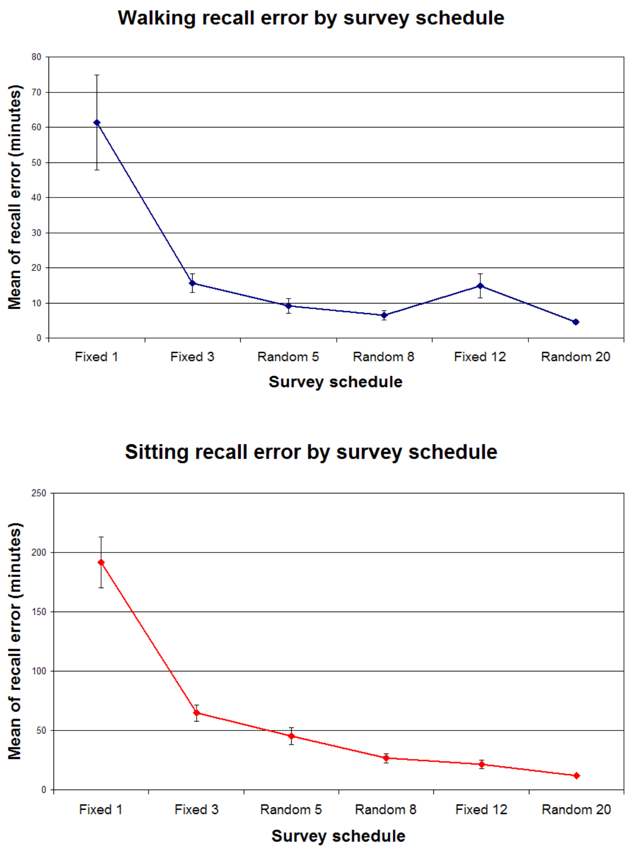

Figure 2: Magnitude of recall error. Absolute difference between sensed and reported total activity duration in minutes for walking (top) and sitting (bottom) as a function of daily survey schedule. Standard error bars indicated.

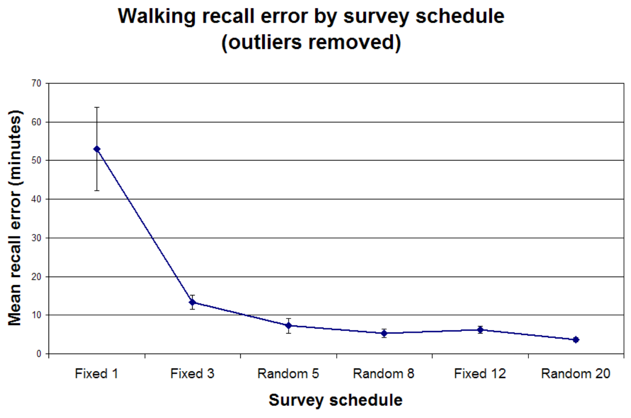

Figure 3: Walking recall errors (in minutes) by schedule with outliers excluded. Absolute difference between sensed and reported total activity duration for walking as a function of daily survey schedule. Standard error bars indicated.

Post hoc pair wise comparisons using least square means Tukey-Kramer HSD tests, revealed that for both walking and sitting activities, the one survey per day schedule led to recall errors that were significantly greater than those from all other schedules. Differences between other schedules were more subtle. Figure 4 shows significant groupings at alpha < 0.05 level. Schedules that do not share a letter are significantly different; groupings with the same letter do not have statistically significant differences. (These are standard format tables for presenting pairwise comparison data.)

a) Pairwise comparisons of survey schedule effects for **walking**

Survey Schedule				**Least Sq. Mean**
Fixed 1	A			70.78
Fixed 12		B		23.69
Fixed 3		B		20.58
Random 8		B	C	14.79
Random 5		B	C	12.94
Random 20			C	7.50

b) Pairwise comparisons of survey schedule effects for **sitting**

Survey Schedule					**Least Sq. Mean**
Fixed 1	A				194.54
Fixed 3		B			66.82
Random 5		B			55.10
Random 8			C		30.87
Fixed 12			C		27.71
Random 20				D	13.75

Figure 4: Pairwise comparison of survey schedule effects on activity recall errors. Alpha <0.05. Significantly different schedule groupings are assigned different letters.

Recall Errors for Walking versus Sitting

The overall magnitude of recall error for sitting was significantly higher than for walking (F[1, 1357]=83.267, p<.001). The mean absolute recall error for walking across all survey schedules was 10.93 minutes (SD=28.48), while for sitting it was 32.14 minutes (SD=56.01). This result is not surprising. Most of us spend far more time sitting than we do walking. Insofar as we incorrectly remember our activity, the magnitude of error is likely to be higher for a more frequently occurring activity.

Are Activity Times Under- or Overestimated?

Health literature has reported that people tend to over-report their amount of physical activity (e.g., walking) and they under-report their sedentary behavior [1,21]. One explanation offered for this finding is that activity recall is in part shaped by social pressure and desirability bias–study participants tend to answer questions in socially desirable ways. For physical activity, the result is that they over-report how physically active they are, and they particularly do so for frequent activities such as housework and walking. Over-reporting is far lower for more rarely occurring activities.

Our results provide only partial support for over-reporting of walking and under-reporting of sedentary behavior. As health research suggested, our subjects consistently underestimated how much they sat (Table 1, right column). Underestimation was greatest for the one time per day schedule, but was found consistently for all schedules for sitting. However, contrary to health research findings, the estimates for walking actually indicate a surprising accuracy rather than consistent over-estimates (Table 1, left). Similar to sitting, walking was significantly *underestimated* for the one time per day schedule (contrary to our hypothesis). However, all other schedules showed only slight over- or under-estimates for walking (over-estimates are shaded in Table 1) suggesting far more accuracy than we hypothesized and no systematic over-reporting.

	Walking (in minutes)	**Sitting (in minutes)**
Fixed 1	M=-23.0, SD=90.25	M=-170.46,SD=140.66
Fixed 3	M=-4.49, SD=28.41	M=-48.5, SD=75.0
Rand. 5	M=2.44, SD=18.76	M=-30.94, SD=67.02
Rand. 8	M=0.15, SD=15.36	M=-12.65, SD=44.95
Fixed 12	M=6.46, SD=43.59	M=-0.33, SD=48.0
Rand. 20	M=-0.85, SD=10.23	M=-5.30, SD=17.55

Table 1: Magnitude and direction of recall over/under-estimates by survey schedule (mean and standard deviation in minutes are shown). Shaded cells mark recall overestimation.

Effects of Occupation

To test if occupation had any influence on recall accuracy, we divided our sample into two groups: office workers (n=10) and non-office (i.e., less-sedentary) occupations (n=8). All office workers indicated in interviews that they spent most of their day at their desk. Both homemakers and the three participants from other occupations (retail worker, student/tight rope walker, and waiter) reported that their schedules were quite variable and they experienced different amounts of activity from day to day. In the light of this, and due to a smaller number of participants in the two non-sedentary occupations, we grouped them together for this analysis.

Occupation effects were found both for walking and sitting (for walking, F[1, 13]=9.58, p<.01; for sitting, F[1, 19]=16.655, p<.001). For both activities, office workers made significantly lower recall errors than participants from other occupations. These results can be at least partly explained by the regularity of most office workers' schedules and their overall lower levels of walking. The relative rarity of walking events on the one hand, and the predictability of the periods when they would be sitting on the other, might have made it easier for office workers to remember their activity accurately.

The significant interaction (F[5, 662]=7.69, p<.001) of occupation and prompting schedules found for walking appears to support this conclusion. On days when they only had one survey—the most error-prone schedule in the study—office workers made significantly lower recall errors for walking than other participants, suggesting that their walking events were quite limited.

Subjective Measures

At the end of each day, our participants rated how they experienced that day's survey schedule on three dimensions: annoyance, difficulty of recall, and subjective perception of accuracy. All three dimensions were rated on five point Likert scales. The annoyance scale ranged from "not at all annoying" to "very annoying," accuracy scale ranged from "very inaccurate" to "very accurate," and difficulty scale ranged from "very easy" to "very difficult."

As hypothesized, Friedman test revealed that survey schedule had a significant effect on participants' annoyance level (chi-square=34.758, df=5, p<.001). In post hoc pairwise comparisons using Bonferroni corrected Wilcoxon test, the 20 times per day schedule was found to be significantly more annoying than all other schedules. No other significant differences in annoyance were found in pairwise comparisons. Interestingly, the 12 times per day schedule was not significantly more annoying than any of the less frequent survey schedules (Figure 5). The reason for this seems to be its hourly regularity. In exit interviews, five participants said that this was their favorite schedule since they always knew when the next survey was coming up and when the previous one had occurred. This made it easier to remember what time interval they needed to account for and thus how much they walked and sat. In fact, one participant, a Ph.D. student studying for her general exam, said in the interview that the hourly schedule was really helpful as a way to structure her day. Whenever her survey would go off, she would know that "another hour has passed." After the first few surveys she started using this rhythm to set reading and writing goals for herself. Aside from this, on exit interviews participants did not report that they modified their daily routines.

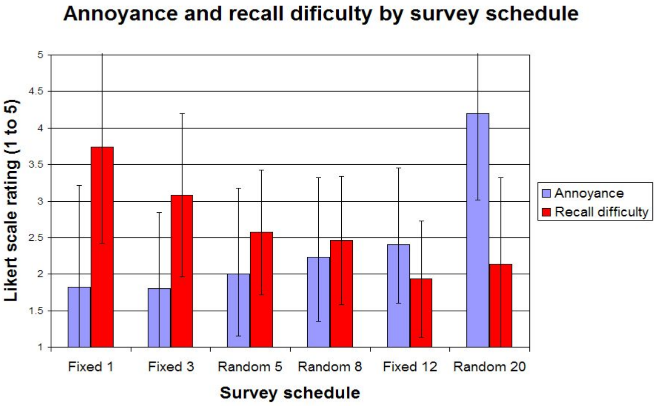

Figure 5: Mean annoyance and difficulty ratings by survey schedule (annoyance –blue bar on left, difficulty – red bar on right), Rated on 5 point Likert scale; Mean and std dev shown. Survey schedule main effect is significant for both perceived difficult and annoyance.

Perception of recall difficulty also supported our hypothesis. Friedman test found a significant effect for survey schedule (chi-square=29.259, df=5, p<.001), and post hoc Bonferroni corrected pairwise testing indicated that participants perceived recall difficulty to be significantly higher when they received only one survey at the end of the day than when they received twelve or more surveys. Similarly, the three times per day schedule was seen as being significantly more difficult than the hourly schedule (Figure 5), which was deemed easiest of all schedules. No significant differences were found for other schedule pairs.

Finally, participants' *perceptions* of their own accuracy in remembering their activity corresponded closely to the frequency of daily surveys. The more surveys they received, the more accurate our participants thought they were at remembering how much they walked and sat (Figure 6). Friedman test revealed that these differences were statistically significant (chi-square=38.865, df=5, p<.001), and post hoc Wilcoxon pairwise testing found that when they had only one survey a day our participants judged their accuracy as significantly worse than when they received five or more surveys a day. Perceived accuracy on the 3 surveys per day schedule was also significantly lower than on the 20 times per day schedule.

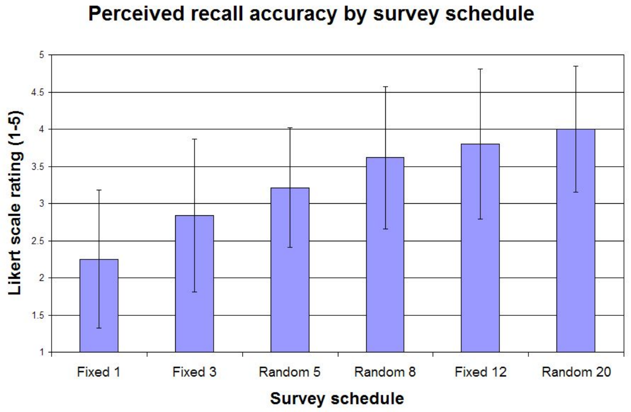

Figure 6: Subjective (perceived) recall accuracy ratings by survey schedule (one score per day for both walking and sitting combined). Fixed 1 is statistically significantly different than schedules of 5 or more surveys. Mean and Std Dev shown.

Random vs. Fixed Schedules

Finally, post study interviews revealed a very clear preference for regular, predictable schedules. Virtually all of our participants noted how much easier the hourly schedule was due to its predictability. Once they realized the pattern after the second or third survey, participants could rely on knowing both when the last survey came up and when the next one was going to be. This made remembering their activity much easier. Interestingly, a number of participants thought that the hourly schedule was the only regular schedule besides the once per day schedule. Although survey times on 3 times per day schedule were also fixed, apparently the frequency of surveys on these days was too low to allow these participants to discover the regularity of the pattern. Also, two participants thought that 20 times per day schedule was a fixed schedule that occurred once every 30 minutes. (In fact, since we enforced a minimum of 10 minutes between surveys, in order to complete 20 surveys in 12 hours this was almost the case). This supports the notion that at high frequencies the two types of schedules appear to be fixed schedules to participants.

There were three main issues that were brought up in relation to our pseudo-random prompting times during final participant interviews. For surveys that came up at random intervals, it was very difficult to remember when exactly the last survey came up. Since all questions asked about activity performed since the last survey, participants needed to figure out how long this time interval was in order to accurately recall their activities. Having a variable time interval made accurate recollection of activity times difficult. The other issue raised was that since they did not know when the next survey would come up, participants kept constantly checking the phone to make sure that they had not missed a survey (since it was unpredictable). Nearly all participants mentioned that regular, predictable schedules of surveys would have made both the study and their recall much easier.

DISCUSSION

These results raise a number of points. First, it seems clear that if our method is to be used for accurate capture of data on routine activities, the prompting schedule needs to be regular and predictable. Four of our participants noted that they felt that a prompt every two hours would be optimal, and even the participants who said that they really liked the hourly schedule expressed that it was probably too frequent. Given our accuracy and annoyance graphs, it appears that a bi-hourly schedule—5 to 8 times per day, depending on the study—would be optimal. In our results, 5 and 8 times per day schedules resulted in fairly accurate recall, while keeping annoyance and recall difficulty low, well under the median.

Even fixed ESM schedules have their problems, however. It is inevitable that even a regular schedule will occasionally catch a participant at a bad moment—when they are driving, in a meeting, or at some other time when they just cannot take the survey (one of our participants received three surveys while he was 45 feet in the air on a tight rope—albeit a fairly "atypical" user!). An optimal prompting strategy would be to use a regular schedule, but one that leverages the inference platform to detect if the user was occupied with things that would prevent him/her from answering the survey, and then defer the questionnaire until the user is available. The MyExperience tool already allows for some interruptibility logic as it will not prompt if it detects that a person is on a call or, if paired with an MSP, if it detects that a person is running or driving. This type of functionality needs to be a core feature of ESM systems if participant burden and intrusiveness are to be kept to the minimum.

Given prior results from health sciences (e.g., [1, 21]) our finding that people did not typically overestimate their walking behavior was surprising. A part of the explanation might be that our study was clearly presented as a study of memory and not of physical activity. It is possible that this framing reduced the social pressures that are found in other physical activity research [21]. Another issue might be that the MSP detected walks that are missed in more typical pedometer research or that (often reported) pedometer miscalibration contributed to underestimation of walking in the literature. If so, the MSP data would reflect higher levels of walking and thus might be closer to the actual activity levels than what is obtained in these other studies. This would cause participants' prior overestimates to in fact be closer to accurate estimates or be underestimates. More investigation is needed here.

A related issue concerns particularly large inaccuracies we found for recall of sitting behavior, especially for low frequency sampling schedules. While we expected people to be inaccurate, we were surprised by the extent of the inaccuracy. We suspect that people simply do not have a very good sense for just how much time they spend sitting during a typical day. A small factor in the magnitude of the observed recall inaccuracy, however, is probably due to potential errors from using the MSP for collection of "ground truth" sitting data. While the MSP detects sitting behavior very accurately (recall that the accuracy of MSP for this was about 96%)., we have observed occasional confusion between standing still and sitting. On these occasions, some activity classified as sitting might actually be that of standing still; thus "sitting" would be slightly overestimated by the MSP. From observations and collecting training data, such periods of standing still are not frequent and are of short durations. (Participants did not wear the MSP while they were lying down, eliminating another source of possible confusion.) The magnitude of the participants' underestimates cannot be fully explained by this confusion, however. Additionally, this would not mask out the differences we found between surveys. (More accurate MSP sitting data would shift the recall error curves down overall across all conditions.) Future work will determine if we can separately infer standing from sitting or instruct participants to recall *both* standing *and* sitting, rather than only sitting to mitigate for this.

Whenever one is using sensor devices for activity detection or data collection, specific applications always require explicit definitions of what constitute an episode of the activity of interest. For example, in fitness applications one might be more interested in *sustained* cardio activities and might want activity episodes to be a minimum of 30 minutes with small or no breaks. For rehabilitative medicine or physiotherapy, one might want physical activities to be *no more than* 10 consecutive minutes to avoid re-injury. To some extent, these definitions are always going to be at least partly arbitrary. In the current study, we made decisions that, we hoped, generated activity definitions that appeared to match everyday practice for non-specific exercise. We collected pilot test data about walking and the amount of pause time needed to wait for street lights, pause for dog "potty" breaks, etc. We tried to differentiate in-office walks to break rooms, washrooms and coffee machines from longer periods of activity people actually considered "walking". We timed various types of sitting episodes to approximate durations for sitting that people actually considered "real episodes" and how much time they might get up for before they considered their sitting "episode" as 2 segments rather than one. However, creating this timing information is difficult. The interruption time we used to bound episodes of walking might be potentially too short and could possibly break episodes of walking when one stops at a particularly long traffic light. It is also possibly too short to tolerate a person stopping off for a cup of coffee or exchanging a few words with an acquaintance one runs into in the street. We did collect data to base our assumptions on, however, it is unclear what better methods there might be to individually tune these episode rules to get better results (or if, in fact, this is attainable).

One consequence of conducting studies that incorporate notions of activity episodes is that at least some of the critical resulting measures need to be independent of the user interpretation of these activity episode definitions. In our case, we focused our analysis on the total time spent walking and sitting, rather than episode counts or shortest and longest episodes of an activity (though we did analyze this other data as well). The point is more general, however. While sensor technology enables us to gain access to a wide range of activity data that would otherwise be nearly impossible to capture, the use of this technology also comes with a new set of methodological requirements and limitations that need to be kept in mind as we design our experiments.

It is worth noting (though perhaps not surprising) that the majority of our participants expressed that participating in the study substantially increased their self-awareness of their activity levels. For office workers this meant that they became aware of just how sedentary they were. However, the increase in awareness was not limited to sedentary behavior. One of the homemakers in the study mentioned that while he knew that he walked quite a bit—he uses a pedometer regularly—the surveys gave him a new sense of just how much time he spent on his feet. In addition, although this was not the intent or goal of the study, several participants noted that the increased awareness of their activity levels motivated them to at least somewhat increase their physical activity. If nothing else, they would go for walks they would have not otherwise taken.

Lastly, although the present study focused on physical activity, we suggested that the same method of regularly spaced short phone-based questionnaires could be useful for studying other frequent, routine activities that are hard to monitor in an automated fashion. For this to be the case, however, our findings would need to generalize. We believe that better recall accuracy for shorter time periods holds for other activities as well, but this could be confirmed through a study parallel to ours that examined a frequent behavior from a different domain but for which ground truth could be easily established. Investigating frequent cell phone use might be a good candidate for such a follow-up study.

CONCLUSIONS

This study determined optimal tradeoff points between accuracy and participant burden to help inform the community about *in situ* survey frequency and methodology. To achieve this we used the Mobile Sensing Platform (MSP) for highly accurate automated logging of basic physical activities such as walking and sitting in conjunction with a phone-based ESM toolkit (MyExperience). Our study specifically investigated the tradeoff between human recall accuracy and survey frequency for self-reported data on routine physical activity (walking and sitting). It is unknown if learning effects and sustained increase in awareness could significantly improve recall accuracy if a system like this were used long term. It is also unclear how long term use would impact annoyance. We are not aware of any long term studies that evaluated automated prompting. User tolerance for effortful self-report is generally directly related to the perceived benefit and value that the data provide (for example, migraine pattern journals, dietary journals, weight training records are typically kept for months or even years). We suspect that this would equally apply to automated journaling. These long term issues are an area for future investigation.

ACKNOWLEDGMENTS

We thank Daniel Avrahami for his helpful comments.

REFERENCES

1. Adams, S.A., Matthews, C.E., Ebbeling, C.B., Moore, C.G., Cunningham, J.E., Fulton, J. and Herbert, J.R. The effect of social desirability and social approval on self-reports of physical activity. *American Journal of Epidemiology*, *161, 4*, (2005), 389-398.
2. Ainsworth, B.E., Sternfeld, B., Slattery, M.L., Dagulsé, V. and Zahrn, S.H. Physical activity and breast cancer: Evaluation of physical activity assessment methods. *Cancer*, *83, 3S*, (1998), 611-620.
3. Alevizos, A., Lentzas, J., Kokkoris, S., Mariolis, A. and Korantzopoulos, P. Physical activity and stroke risk. *International Journal of Clinical Practice*, *59, 8*, (2005), 922-930.
4. American College of Sports Medicine American College of Sports Medicine position stand: Osteoporosis and exercise. *Medicine & Science in Sports & Exercise*, *27, 4*, (1995), i-vii.
5. Bishop, C.M. *Pattern recognition and machine learning*. Springer, New York, 2006.
6. Blair, S.N. and Connelly, J.C. How much physical activity should we do? The case for moderate amounts and intensities of physical activity. *Research Quarterly for Exercise and Sport*, *67, 2*, (1996), 193-205.
7. Consolvo, S. and Walker, M. Using the Experience Sampling Method to Evaluate Ubicomp Applications. *IEEE Pervasive Computing Magazine: The Human Experience, 2, 2*, (2003), 24-31
8. Csikszentmihalyi, M. and Larson, R. Validity and reliability of the Experience-Sampling Method. *Journal of Nervous and Mental Disease*, *175, 9*, (1987), 526-536.
9. Durante, R. and Ainsworth, B.E. The recall of physical activity: Using a cognitive model of the question-answering process. *Medicine & Science in Sports & Exercise*, *28, 10*, (1996), 1282-1291.
10. 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 *Proc MobiSys '07*, (2007), 57-70.
11. Hakim, A.A., Curb, J.D., Petrovitch, H., Rodriguez, B.L., Yano, K., Ross, G.W., White, L.R. and Abbott, R.D. Effects of walking on coronary heart disease in elderly men: The Honolulu Heart Program. *Circulation*, *100, 1*, (1999), 9-13.
12. Hektner, J.M. and Csikszentmihalyi, M. The experience sampling method: Measuring the context and content of lives. In Bechtel, R.B. and Churchman, A. eds. *Handbook of environmental psychology*, John Wiley & Sons, Inc., New York, 2002, 233-243.
13. Helmrich, S.P., Ragland, D.R., W., L.R. and Paffenbarger, R.S., Jr. Physical activity and reduced occurrence of non-insulin-dependent diabetes mellitus. *New England Journal of Medicine*, *325, 3*, (1991), 147-152.
14. Intille, S.S., Rondoni, J., Kukla, C., Ancona, I., and Bao, L. Context-aware experience sampling tool. In *Proc CHI 2003,* (2003), 972-3.
15. Lee, I.-M. and Skerrett, P.J. Physical activity and all-cuase mortality: What is the dose-response relationship? *Medicine & Science in Sports & Exercise*, *33, 6, Suppl.*, (2001), S549-471.
16. Lester, J. Choudhury, T., & Borriello, G. A practical approach to recognizing physical activities. *Proc Pervasive '06*, (2006), 1-16
17. Livingstone, M.B.E., Robson, P.J., Wallace, J.M.W. and McKinley, M.C. How active are we? Levels of physical activity in children and adults. *Proceedings of the Nutrition Society*, *62*, (2003), 681-701.
18. Paffenbarger, R.S., Jr., Wing, A.L., Hyde, R.T. and Jung, D.L. Physical activity and incidence of hypertension in college alumni. *American Journal of Epidemiology*, *117, 3*, (1983), 245-257.
19. Sallis, J.F. and Saelens, B.E. Assessment of physical activity by self-report: Status, limitations, and future directions. *Research Quarterly for Exercise and Sport*, *71, 2*, (2000), 1-14.
20. Salmon, P. Effects of physical exercise on anxiety, depression, and sensitivity to stress: A unifying theory. *Clinical Psychology Review*, *21, 1*, (2000), 33-61.
21. Shephard, R.J. Limits to the measurement of habitual physical activity by questionnaires. *British Journal of Sports Medicine*, *37*, (2003), 197-206.
22. U. S. Department of Health and Human Services *Physical activity and health: A report of the Surgeon General*. U. S. Department of Health and Human Services, Atlanta, 1996.
23. Wheeler, L. and Reis, H.T. Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality*, *59, 3*, (1991), 339-354.