

Exploring Micro-Incentive Strategies for Participant Compensation in High-Burden Studies

Mohamed Musthag, Andrew Raij[†], Deepak Ganesan, Santosh Kumar[‡], Saul Shiffman^{*}

Univ. of Massachusetts Amherst, Univ. of South Florida[†], Univ. of Memphis[‡], Univ. of Pittsburgh^{*}
{musthag,dganesan}@cs.umass.edu, rajj@usf.edu, skumar4@memphis.edu, shiffman@pinneyassociates.com

ABSTRACT

Micro-incentives represent a new but little-studied trend in participant compensation for user studies. In this paper, we use a combination of statistical analysis and models from labor economics to evaluate three canonical micro-payment schemes in the context of high-burden user studies, where participants wear sensors for extended durations. We look at how these strategies affect compliance, data quality, and retention, and show that when used carefully, micro-payments can be highly beneficial. We find that data quality is different across the micro-incentive schemes we experimented with, and therefore the incentive strategy should be chosen with care. We think that adaptive micro-payment based incentives can be used to successfully incentivize future studies at much lower cost to the study designer, while ensuring high compliance, good data quality, and lower retention issues.

Author Keywords

micro-payments, micro-incentives, user study.

ACM Classification Keywords

H.5.m Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Economics, Experimentation, Human Factors, Performance

INTRODUCTION

A growing segment of the health and science community is turning to ubiquitous computing technologies, such as wearable sensors and smartphones, to conduct studies in natural environments [4, 8, 15, 16, 20, 21]. These studies are providing new understanding of the human condition in the real world rather than the lab, and leading to more ecologically valid mechanisms to measure, evaluate, and improve health.

Often, such health and science field studies are characterized by high burden on the participant. They may ask participants to wear sensors on their body for days at a time or provide frequent self-reports (e.g., 10-20 times per day) that interrupt the normal flow of daily life. This burden has consequences. It reduces participant compliance with the study protocol

(not wearing the sensors one day or ignoring questionnaires) and can even lead to participant dropouts. By reducing compliance and retention, burden effectively reduces the amount and quality of the data, and by extension, its validity and statistical power.

Incentives are a powerful and common mechanism to counteract the effects of burden [6, 28, 29]. Initially, they are used to entice people to participate in a study. During the study, they encourage participants to follow the study protocol (compliance), remain in the study (retention), and provide high quality data [5]. Incentives come in many forms, including money or gift cards, free health care services, and the knowledge that one's participation can advance science and/or improve the human condition [10, 13, 19].

Monetary incentives are typically distributed in fixed bulk amounts for completing study milestones. In some studies, there is only one milestone, reaching the end of the study. Other studies might have several milestones, spaced days or weeks apart (e.g., completion of the first lab study, completion of 1 day of monitoring, etc.). Each incentive is then used to encourage the participant to reach the next milestone and reward the participant for continued participation. How often to award incentives (number and schedule of milestones), how much each award should be per milestone, and how much to offer in total for the entire study is decided based on the goals of the study, importance of each milestone, the overall burden to the participant, and the budget available to the study designers.

While traditional high-burden user studies have largely relied on such coarse-grained, bulk payments, the use of mobile phones and other devices offer significant new flexibility in terms of when, what, and how much incentives are offered. For example, amounts offered could be varied based on the number of hours a participant wore a sensor, upon completion of individual self-report questionnaire, or even upon completion of a single question within a questionnaire. They could also be varied by the day of the study, time of day, type of question asked, consistency or quality of answers, extent of burden, and so on. Perhaps the amounts offered could be tailored to the individual, and use personalized pricing while simultaneously reducing the overall cost of running the study. In addition, the frequency with which incentives are provided could be varied, or the presentation of incentives could be manipulated. For example, incentive amounts could only be revealed after a question is answered, making the act of compliance more like playing a lottery.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

UbiComp '11, September 17–21, 2011, Beijing, China.

Copyright 2011 ACM 978-1-4503-0630-0/11/09...\$10.00.

The increased flexibility offered by micro-incentive strategies poses a significant challenge to the study designer. What micro-incentive strategy should be chosen to improve compliance? At what granularity should incentives be provided? What effect does varying incentives have on compliance, data quality, and retention? Does one micro-incentive strategy work for all participants, or should they be personalized to the individual? What are common pitfalls of micro-incentive strategies? To what extent can micro-incentives lead to unintended adverse consequences? While the possibilities are tantalizing, the dearth of studies on micro-incentive strategies complicates the design of user studies.

This paper is an initial exploration of the spectrum of micro-payment design choices facing study designers as more field studies become mobile-based and put larger burdens on participants. While it is infeasible to cover the space of micro-incentives strategies, we use three canonical mechanisms to explore the benefits and pitfalls of these approaches. We look at a high-burden study that involves wearing a chest-band with several on-board sensors for three days. Throughout, participants experience roughly 20 interruptions from experience sampling questionnaires per day. The three micro-incentive mechanisms that we compare are simple, intuitive choices that a study designer might make: the *Uniform* strategy pays the same amount per questionnaire, the *Variable* strategy pays different amounts for each questionnaire based on a prior distribution of rewards, and the *Hidden* strategy also has different amounts per questionnaire but includes a lottery component by hiding the amount earned until the user completes the questionnaire.

Our study enables us to answer several key questions including: 1) which micro-incentive schemes perform well or poorly?, 2) how does the amount of micro-incentive impact compliance and data quality?, 3) do micro-incentive strategies encourage adverse behavior from participants?, 4) how do micro-incentive schemes impact retention? and 5) do users perceive micro-incentives favorably?

We use a combination of statistical analysis and models from labor economics to answer these questions, and draw several implications from the data. We find evidence that micro-incentives are a powerful tool for the study designer, and have a significant impact on participant behavior. We think that adaptive micro-incentives can be used to reduce the cost of future studies, while providing high-quality results. However, we also find evidence that poorly chosen micro-incentive amounts can lead to adverse behavior from participants. For example, higher micro-incentive amounts may motivate lying on questionnaires to earn more money. The results from our study lead to several valuable insights that can be useful for study designers in the future.

RELATED WORK

The availability of micro-incentive based task markets such as the Amazon Mechanical Turk (AMT [1]) has led to several incentive-based studies. Mason and Watts found that higher micro-payments led to higher task completion rates on AMT [17]. However, the quality of the data provided

by participants was not affected by the size of the incentive. Similarly, Heer et al [11] used AMT to conduct a visualization study and found that larger incentives led to quicker completion times, but again, no difference in the quality of data provided across different incentive amounts. Micro-payments have also been used to monetize downloads from peer-to-peer networks [31] and websites [27], as well as motivate environmental stewardship [30].

Micro-payments are being increasingly used in mobile phone-based user studies. For example, Consolvo et al awarded participants \$1 for each questionnaire completed in a seven-day field study [8]. Similar to our study, Reddy et al studied micro-payment incentives in the context of participatory sensing [25]. Their study tested multiple micro-payment amounts as well examined how game-like characteristics affect the compliance and data quality. Participants were divided into five groups, and were assigned to take pictures of the contents of waste bins around a University campus. The first group was paid a fixed amount at the end of the study, three groups were paid a uniform micro-payment for each valid picture, and the last group was awarded a variable amount for valid images, depending on their ranking with respect to other participants.

Our study differs from [25] in three important ways: a) we examine micro-payments for *carefully controlled scientific field studies* where motivation is more difficult because of significant burden (wearable sensors, and several interruptions from questionnaires per day), rather than for more voluntary, unstructured participatory sensing tasks, b) we analyze results using labor economic models of reservation wages, which gives greater insight into the population, and c) motivated by concepts from behaviorism and “gamification” [18, 2, 9, 22, 12, 23, 7, 26], we include *Variable* and *Hidden* incentives schemes to understand the effect of different incentive strategies on compliance, data quality, and retention (discussed in the next section).

STUDY DESIGN

We designed a study to examine the effect of micro-payment incentives on high-burden experience sampling studies. Specifically, the study is designed to measure how compliance, data quality, and the perceived burden of participation are affected by three different schemes for distributing micro-payments, *Uniform*, *Variable*, and *Hidden*. In the *Uniform* scheme, participants are awarded a fixed amount for completing a micro-task. In the *Variable* scheme, awards vary randomly based on a prior distribution. In the *Hidden* scheme, awards again vary randomly, but the amount is not revealed until after the micro-task is completed. Both of these schemes are inspired by the variability and unknowns typical in game dynamics, and are intended to elicit more engagement, and ultimately, more compliance and better data quality.

To enable an analysis of these micro-incentive schemes within a realistic context, we integrate them into an existing field study. The study’s primary goal was to collect physiological, psychological, and behavioral measures of stress from people in natural environments, specifically when exposed to

interruptions in daily life (results reported separately). The study had several characteristics typical of recent ubiquitous computing and health science field studies and applications [4, 8, 15, 16, 20, 21], and was thus ideally suited to meeting the goals of this paper. These characteristics include:

- ▶ **Experience Sampling Questionnaires:** Participants provided frequent self-reports. They were prompted to complete 20 questionnaires per day, with each questionnaire having between 21 and 25 questions. Each question in a questionnaire was an opportunity to encourage compliance, data quality, and retention using micro-incentives.
- ▶ **Three Days:** The study required three consecutive days of participation, allowing analysis of the effect of waning interest in the study as the study progresses, and whether micro-incentive structures can offset this waning interest.
- ▶ **Wearable Sensors:** Participants in the study wore sensors around their chest for 10-12 hours each day of the study, not unlike other UbiComp and behavioral science applications and studies.
- ▶ **High Burden:** The above characteristics - frequent questionnaires, monitoring with wearable sensors, and for several days at a time - represent a significant burden to participants. Adding to this burden, participants had to return to the lab each day to exchange used batteries for new ones. We believe this level of burden represents the upper end of what participants in field studies are usually exposed to. Thus, our results will provide insight into appropriate incentive schemes for high-burden field studies.

Incentive Structures

We ran a between-subjects study, where participants were assigned to one of three incentive schemes, *Uniform*, *Variable*, and *Hidden*. Each scheme paid the participant a micro-payment for completing a single question in a questionnaire. We incentivize each question in a questionnaire, because a question is the finest-grain opportunity to measure compliance and data quality in experience sampling studies. *Uniform* participants were paid a fixed amount of four cents per question completed. Thus, if participants answer all 25 questions in a questionnaire, they earned \$1. This level of compensation is similar to that given in [8], where \$1 was awarded for each questionnaire completed.

Variable and *Hidden* participants were paid a variable amount, between 2 and 12 cents per question. This amount changed from questionnaire to questionnaire according to a random distribution. The distribution was biased such that smaller micro-payments were offered more frequently than larger ones. For example, 2-cent awards were offered 30% of the time, 4-cent awards were offered 10% of the time, and 12 cent awards were offered 1% of the time. If a *Variable* participant was offered 12 cents per questions, they stood to earn as much as \$3 if they completed all questions - three times as much as *Uniform*.

The distribution was also chosen such that the expected amount earned for completing all 25 questions in a questionnaire was approximately \$1, the same amount earned by *Uniform*. Thus, even though *Variable* and *Hidden* could occasionally

earn far more than *Uniform* on a questionnaire, on average all three groups could earn the same amount of money (assuming they completed all questionnaires presented). Guaranteeing similar potential earnings across all three groups ensured all participants were compensated similarly and fairly. The only difference between *Variable* and *Hidden* incentives is that *Hidden* was not told the amount each question was worth until they completed an entire questionnaire.

Justification of Incentive Structures: The design space of micro-incentive schemes is large, with many options available to the study designer. From this space, we chose to study the *Uniform*, *Variable*, and *Hidden* micro-incentive structures described above for several reasons.

The *Uniform* scheme represents a likely choice for a study designer interested in incentivizing micro-tasks in a study. Uniform micro-payments are simple to implement (they do not require calculation of individual micro-incentives) and are a natural evolution of bulk payment schemes that are already familiar to study designers. Indeed, this evolution is apparent in two recent studies [25, 8] that adopted or examined uniform micro-payments. Furthermore, since the *Uniform* scheme is a likely choice for study designers aiming to incentivize micro-tasks (e.g., question completion), we use it as a pseudo-control condition for comparison to the *Variable* and *Hidden* incentive structures.

The *Variable* incentive scheme was chosen because of the effectiveness of variable rewards in sustaining behavior in games [2, 18, 12] and broader behavioral contexts [9, 22]. In addition, variable incentives allow examination of the influence of reward sizes on desired outcomes (high compliance, data quality, etc.).

Lastly, the *Hidden* incentive scheme was chosen because it is akin to playing the lottery or gambling where the exact reward is not always known [2, 23]. Lotteries usually offer higher rewards (jackpots) at far lower frequencies than lower rewards. The contrast between low and high rewards and their frequencies create a sense of scarcity that artificially increases the perceived value of the reward [7], and by extension, the motivation to play. Furthermore, the occasional higher reward or near-miss of a higher reward creates the impression that high rewards are possible, thus encouraging the player to continue playing [26]. We note that participating in our study is not the same as gambling. Unlike gambling, our participants always earn money and cannot lose money for answering a question.

Procedure

36 participants were recruited from the student population at the University of Memphis in the United States using flyers and word-of-mouth. Participants received compensation according to one of the three incentive structures to which they were randomly assigned. 12 participants were assigned to each incentive structure. Earned incentives were paid in full when participants completed the study.

Day 1 Start: After providing informed consent, participants were introduced to the wearable sensor suite they wore about their chest underneath their clothes throughout the study. They were also provided smartphones, which they used to complete the experience sampling questionnaires in the field. They were asked to wear the sensors and carry the phone with them during waking hours (10-12 hours per day). The participant was then randomly assigned to one of the three incentive schemes. The study coordinator explained the assigned incentive scheme and how participants could keep track of their earned incentives (see below). Participants were not told about the other incentive schemes. All participants were told they could earn approximately \$60 if they complete all questionnaires requested of them (Average of \$1 per questionnaire \times 20 questionnaires per day \times 3 days). Participants returned to the lab each day at a scheduled time to replace sensor batteries, provide backups of data collected, and mark down the incentives earned each day.

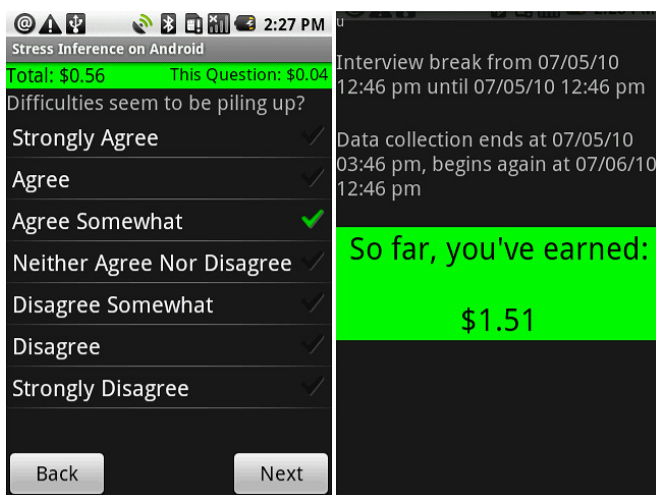


Figure 1. Left - The user interface for the experience sampling questionnaires (left) and the display shown on the phone when participants were not completing questionnaires (right). The questionnaire interface updates incentives in real-time as questions are completed.

Field Study: In the field, participants were prompted to complete questionnaires 20 times per day. Ten appeared when the participant’s stress or speaking state changed (context-aware experience-sampling [14]), as detected on the smartphone using data collected by the sensor suite [20, 24]. The remaining ten questionnaires were scheduled to appear 55 minutes after the last questionnaire (whether triggered by inferences or scheduled). To keep participant burden within reason, participants did not receive questionnaire prompts within 30 minutes of a previous prompt. The mobile phone vibrated and played an audible tone to alert the participant to a new questionnaire. To encourage timely completion, time limits were imposed on questionnaires. If the participant could not attend to the questionnaire immediately, he/she could choose a one-time delay of ten minutes.

To help participants keep track of awarded micro-payments, the smartphone software included several reminders of the incentives available and already earned. Before starting the

questionnaire, a message appears informing the participant how much money they could earn per question answered (“Each question you answer in this interview will earn you _____.” In the case of the *Hidden* group, the blank was filled with “???” followed by the statement: “We’ll tell you what you earned after the interview.” Next, the first question in the questionnaire would appear and participants could answer by selecting choices from a list (Figure 1 - Left). In the case of the *Uniform* and *Variable* groups, the top of the questionnaire screen also showed the amount the participant could earn for the current question and the amount earned on the current questionnaire thus far. The amount earned thus far changed as soon as the participant answered the question, providing instant positive feedback of compliance. After completing all questions, a summary screen would appear displaying the amount earned per question, the total earned for the just-completed questionnaire, and the total amount earned since the beginning of the study (Figure 1 - Right). For the *Hidden* group, this was the first and only time they were made aware of the amount earned per question and the total earned for the just-completed questionnaire. The main application screen also provided a persistent display of the total amount earned since the beginning of the study.

Field Questionnaire: The field questionnaire contained questions that assess stress levels and situational information about the user at the time of the administration. Sample questions include “Angry/Frustrated?,” “Happy?,” and “Talking?” Meta-data useful for assessing compliance and data quality were also collected, including the start and completion time of each question and questionnaire, the number of questions answered, and whether the questionnaire was delayed.

The questionnaire also includes gateway questions, which if answered in a particular way leads to one or more optional questions. For example, if the user responds “yes” to “Talking?,” three additional questions appear: “If talking, on the phone?,” “If talking, with whom?” (men and/or women), and “If talking, with whom?” (significant other, children, relatives, etc.). Micro-payments are also earned for these optional questions, providing an opportunity to assess how incentives affect compliance and data quality for optional questions. Micro-incentives could encourage participants to answer these questions more but they could also lead to inaccurate responses to earn more money.

End-of-Study: Participants returned to the lab and returned equipment at the end of three days. They then completed an end-of-study questionnaire to gather subjective feedback about the incentive scheme and their perception of the burden they experienced during the study.

RESULTS

We analyzed participant data with respect to the goals of the study as discussed in the preceding section. We now discuss the results and their interpretation. We use mixed-effects regressions analyses, which account for correlation of observations within individuals, to test for significant differences. Where space allows, p-values are reported with the means and confidence bounds of the corresponding distributions.

Compliance

Compliance is the extent to which participants execute the study protocol. The compliance for an experience sampling questionnaire is measured using two metrics: the completion rate, and the number of questions answered. Completion rate refers to the percentage of questionnaires completed of the total prompted to participants. The number of questions answered refers to the total number of questions answered on completed questionnaires. This can vary across participants due to the presence of optional questions.

Completion Rate

We look at completion rates along two axes: the overall completion rate across different micro-payment sizes, and across different incentive mechanisms.

First, we look at the *Variable* scheme and ask if the likelihood of questionnaires being completed varied with the size of the micro-payment offered on questionnaires. Note that only participants in *Variable* observed a changing micro-payment incentive from survey to survey. Surprisingly, we do not find any evidence to indicate that the likelihood of questionnaires getting completed was affected by the offered micro-payment size for the *Variable* group. In fact, the EMAs with lowest micro-payment were just as likely to be completed as the EMAs with highest micro-payment.

The influence of incentives on completion rate is, perhaps, better articulated by using the notion of reservation wage of an individual [3]. In Labor Economics, the *reservation wage* refers to the minimum amount an individual has to be paid for the individual to complete some task. The reservation wage of an individual is the product of the individual’s assessment of the benefit and cost of participation in the particular task. Depending on the individual’s preferences, this assessment process could take in to consideration things like the private value to him of contributing to the advancement of Science. Based on his budget, the study designer chooses how much to pay the participant, i.e. he has to choose w_j , which is the per question incentive offered to participants for each questionnaire j . The participant then decides whether to complete the questionnaire or not, based on w_j and r_i , where r_i is the *reservation wage* of individual i . Note that r_i can be modeled as a small range for practical purposes. Thus, the individual will choose to complete the questionnaire only if $w_j \geq r_i$.

One conclusion from this result is that we overpaid individuals for their EMAs i.e. that the reservation wage for most individuals was below even the lower threshold that we had chosen. This implies that, for our sample, we could have reduced the incentives by significant amounts without reducing the completion rate. In fact, the results also suggest that we may have offered too much for the *Uniform* strategy ($4c$ per question); if we had instead offered an amount of $2c$ (low end of the *Variable* strategy), we would have reduced overall costs by roughly 50% with similar completion rates.

Second, we look at overall completion rates across different incentive mechanisms. Table 1 shows the completion rates

Groups	Mean (%)
<i>Overall</i>	85.03 \pm 1.61
<i>Uniform</i>	86.28 \pm 2.58
<i>Variable</i>	86.92 \pm 2.70
<i>Hidden</i>	81.68 \pm 3.12

Table 1. Completion rate across groups. (Participant Avg % \pm 95% CI)

we observe in our study, over all participants, and across different incentive mechanisms. We find high completion rates, over all participants, with the mean completion rate standing at 85%. We find no statistically significant difference in completion rates across the groups, although notable is the fact that *Hidden* performs worst amongst the three schemes – a trend that we see across several analyses.

Number of questions answered

In order to see if particular incentive schemes encouraged participants to answer more questions, we pose several optional questions in each questionnaire. A questionnaire is considered completed if all required questions have been answered. Thus, for completed questionnaires, the number of questions answered varies only to the extent the number of optional questions answered varies. We analyze the optional questions data in three ways: across different micro-payment sizes, over time, and across different incentive schemes.

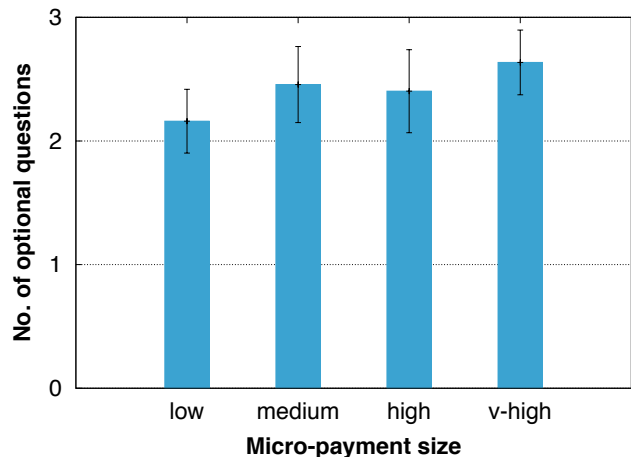


Figure 2. Number of optional questions answered vs. micro-payment size for *Variable*. The bins are quartiles — low $2c$; medium $3c$; high $4c$; v-high $> 4c$

First, we ask if the micro-payment offered on questionnaires had an effect on the number of optional questions answered by participants in *Variable*. We find a statistically significant positive effect of the size of the micro-payment incentive on the number of optional questions answered, i.e. larger micro-payment incentives encouraged participants in *Variable* to answer more optional questions ($p < 0.013$). Figure 2 shows the average number of optional questions answered broken down by the size of the micro-payment incentive offered on questionnaires, for the *Variable* incentive scheme.

On this surface, incentivizing participants to answer more optional questions appears useful, but this depends on the truthfulness of the answers provided. To understand this behavior, we turn to a model from Labor Economics. The decision making process can be explained as a simplistic two player game between the study designer and the participant. The study designer has to decide how much to pay the participant, i.e. he chooses w_j , which is the per question incentive offered to participants for each questionnaire j . The participant then decides whether to answer honestly or dishonestly on the gateway question, that would lead to one or more subsequent optional questions, based on w_j and r_i , where r_i is the *reservation wage* of individual i .

	$w_j < r_i$	$w_j = r_i$	$w_j > r_i$
Talking	NT	T	T
Not Talking	NT	NT	T

Table 2. Choice matrix facing an individual when considering answering the gateway question “Are you Talking?”

Table 2 shows the choice matrix facing the individual, when he is considering answering one of the several gateway questions in the questionnaire, “Are you Talking?”. If the individual answers “Yes”, that leads to three additional questions that can be answered to earn more on the questionnaire. There are three cases to consider:

- ▶ When $w_j < r_i$, the individual does not feel that the reward is sufficient for the burden of answering the optional questions. Therefore, to avoid answering the subsequent optional questions, he might be tempted to answer in the negative (NT), even if he was in fact talking (T).
- ▶ When $w_j = r_i$, the individual is indifferent between answering, or not answering the subsequent optional questions. There is no incentive for him to be dishonest, so he will tend to answer gateway question truthfully.
- ▶ When $w_j > r_i$, the reward outweighs the burden of answering the optional questions. Therefore, in order to maximize his reward, he would be tempted to answer in the affirmative (T), even if he was, in fact, not talking (NT).

The lack of ground truth validation makes it difficult to determine the truthfulness of results provided by users in our study. However, the trend in Figure 2 does indicate that users are more incentivized to answer high-paid optional questions, suggesting that the above model may indeed have some validity. These observations have important implications for study design. In particular, we think that when using micro-payment strategies together with optional questions, it is important to have ground truth validation. For instance, sensors on a smartphone may be used to collect context about whether an individual is talking or not. This can help check if the gateway question was answered correctly, and a study designer can be confident that micro-payment incentives have played a positive role in nudging the participant to answer more questions. In our study, we were unable to make this correlation since the data from our sensors were too noisy for accurate activity classification. In cases where accurate ground truth estimation is not possible, we recommend that

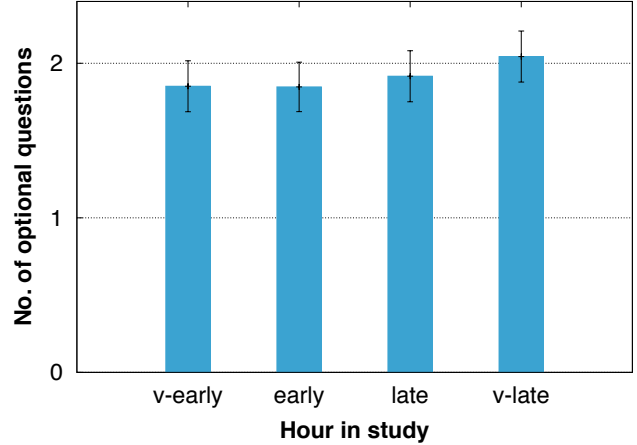


Figure 3. Number of optional questions answered over time for Variable. The bins are four quartiles of the study duration.

any optional branches pay out identical sums, so that the individual has no incentive to pursue one of the branches.

Second, we look at the trend in answering optional questions over time to see whether there was any notable change. Curiously, we find that participants do indeed tend to answer more optional questions over time ($p < 0.055$). Figure 3 shows a breakdown by quartile. This result is not easy to explain — is it due to honest reporting or are users taking advantage of optional questions as they become more familiar with the study design? Again, these results make the case for using context inference to judge the validity of the data provided by users.

Groups	Mean
Uniform	2.08 ± 0.14
Variable	2.09 ± 0.14
Hidden	1.55 ± 0.14

Table 3. Number of optional questions answered across groups. (Participant Avg \pm 95% CI)

Third, we check to see if the number of optional questions answered differed across incentive schemes. Table 3 shows the number of optional questions participants answered across different incentive schemes. We find that participants in the *Hidden* incentive scheme are likely to answer fewer optional questions than participants in the *Uniform* ($p < 0.045$) and the *Variable* ($p < 0.060$) incentive schemes.

Data Quality

Data quality is the extent to which questionnaire answers are truthfully completed. We measure data quality using a consistency score that we compute by comparing answers on questionnaires which implicitly capture the same information. Specifically, we consider two opposing question-pairs to compute consistency: 1) Happy vs. Sad, and 2) Cheerful vs. Sad. We argue that these two metrics are reasonable for measuring questionnaire consistency because they represent different extremes in the valence of emotion, i.e. a partici-

part is extremely unlikely to be both “Happy” and “Sad”, or “Cheerful” and “Sad” at the same time.

The pairs of questions in the consistency metrics above are *inversely-related*, in that a high score on one question should imply a low score on the other. Participants choose an answer amongst these choices: 1 ‘Strongly Disagree’, 2 ‘Disagree’, 3 ‘Disagree Somewhat’, 4 ‘Neither Disagree Nor Agree’, 5 ‘Agree Somewhat’, 6 ‘Agree’, 7 ‘Strongly Agree’. Let $r_{jq} \in \{1, 2, 3, 4, 5, 6, 7\}$ be the response to question q in questionnaire j . The consistency score, $cs_{jqj'}$, for a pair of *inversely-related* questions, q and q' , on questionnaire j is computed as follows:

$$cs_{jqj'} = \begin{cases} 1 & \text{if } (r_{jq} > 4 \text{ and } r_{jq'} < 4) \text{ or } (r_{jq} < 4 \text{ and } r_{jq'} > 4) \\ 0 & \text{otherwise} \end{cases}$$

Groups	Happy vs. Sad (%)	Cheerful vs. Sad (%)
<i>Uniform</i>	94.74 ± 1.68	92.99 ± 1.92
<i>Variable</i>	95.70 ± 1.62	94.21 ± 1.87
<i>Hidden</i>	87.56 ± 2.66	87.06 ± 2.70

Table 4. Happy vs. Sad and Cheerful vs. Sad consistency across groups. (Participant Avg % ± 95% CI)

Table 4 shows the proportion of surveys that are consistent as measured using Happy vs. Sad and Cheerful vs. Sad consistency metrics, respectively. In terms of the Happy vs. Sad metric, we find that questionnaires completed by participants in *Hidden* are less likely to be consistent than questionnaires completed by participants in both *Uniform* ($p < 0.007$) and *Variable* ($p < 0.005$). In terms of the Cheerful vs. Sad metric, we find that questionnaires completed by participants in *Hidden* are less likely to be consistent than questionnaires completed by participants in both *Uniform* ($p < 0.040$) and *Variable* ($p < 0.018$).

If we look at the *Hidden* group more closely, we see that participants in this group do not know how much they would earn by completing the questionnaire at hand. This could perhaps frustrate this group and lead to the behavior we observe here.

Groups	Mean (hours)
<i>Uniform</i>	62.77 ± 10.48
<i>Variable</i>	62.46 ± 7.54
<i>Hidden</i>	62.84 ± 10.35

Table 5. Retention across incentive schemes. (Participant Avg (h) % ± 95% CI)

Groups	Completed (%)	Delayed (%)
v-early	87.87 ± 3.07	4.81 ± 2.01
early	85.56 ± 3.21	4.31 ± 1.36
late	84.40 ± 3.39	7.40 ± 2.13
v-late	82.61 ± 3.39	6.00 ± 2.13

Table 6. Completion rate and Delay rate over time. Each bin shows a quartile. v-early (Q1) < 10h; 10 ≤ early (Q2) < 30h; 30h ≤ late (Q3) < 52h; v-late (Q4) ≥ 52h. (Participant Avg (h) % ± 95% CI)

Next, we drill further down, and examine *how* inconsistent the inconsistent questionnaires were. We breakdown incon-

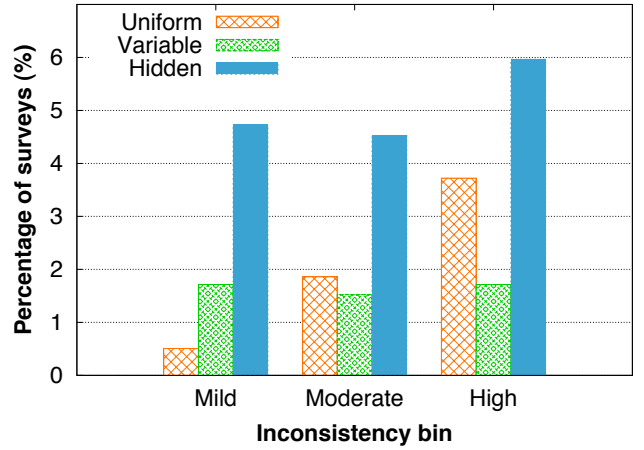


Figure 4. Happy vs. Sad - Degree of failure

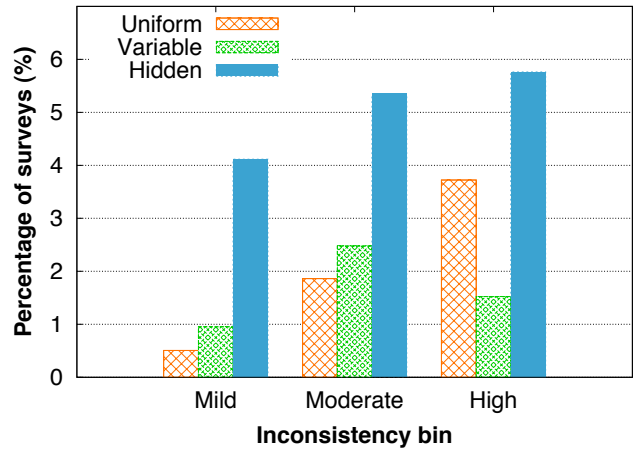


Figure 5. Cheerful vs. Sad - Degree of failure

sistent behavior into three bins: mildly, moderately, and highly inconsistent. An inconsistent question-pair is: a) *mildly* inconsistent when participants ‘Agree Somewhat’ or ‘Disagree Somewhat’ to both questions, b) *highly* inconsistent when one of the answers is ‘Strongly Disagree’ or ‘Strongly Agree’, and c) *moderately* inconsistent otherwise. Figures 4 and 5 show a breakdown for Happy vs. Sad and Cheerful vs. Sad, respectively. The figures show that *Hidden* has the highest level of inconsistency, with 6% of all questionnaires completed being highly inconsistent, and 5% being moderately inconsistent. This, perhaps, reinforces our hypothesis that participants in *Hidden* may be getting frustrated.

Retention

Retention is the extent to which participants remain active over the course of the study. Here, we measure retention as the number of hours participants remain in the study, *i.e.* the difference between the time of the last and first questionnaires they answered. While this metric does not account for periods of inactivity within this duration, previously discussed compliance metrics already capture this effect.

Question	<i>Uniform</i>	<i>Variable</i>	<i>Hidden</i>
1 I felt more engaged in the study because of micro-payments	3.08 ± 0.57	3.17 ± 0.53	2.92 ± 0.63
2 I'd rather receive \$50 bulk-payment, than micro-payments for each question	2.83 ± 0.60	2.42 ± 0.74	2.83 ± 0.60
3 The phone interfered with my daily activities	2.25 ± 0.39	2.17 ± 0.60	2.08 ± 0.57
4 The phone interfered with my social interactions	2.67 ± 0.49	2.25 ± 0.61	2.00 ± 0.54
5 Overall, the phone was a nuisance	1.83 ± 0.46	2.0 ± 0.66	1.92 ± 0.50
6 The chestband interfered with my daily activities	2.08 ± 0.42	2.67 ± 0.56	2.33 ± 0.56
7 The chestband interfered with my social interactions	2.17 ± 0.53	2.08 ± 0.74	1.67 ± 0.41
8 Overall, the chestband was a nuisance	2.00 ± 0.54	2.58 ± 0.57	2.58 ± 0.50

Table 7. Participant Perception. Participants responded on a scale of 1 to 4. 1 - 'Strongly Disagree'; 2 - 'Disagree'; 3 - 'Agree'; 4 - 'Strongly Agree'. (Participant Avg ± 95% CI)

#	Age group	Interruptions	Duration	Payment	Completion rate
1	23-50y	Every 45min, take measurement, and complete 58-item survey.	6 days	\$42/day	81% completed at least 75% of prompts.
2	50-70y	Every 45min, take measurement, and complete 45-item survey.	6 days	\$33/day	88% completed at least 57% of prompts.
3	17-30y	Roughly every 45min, complete 25-item survey.	3 days	\$20/day	81% completed at least 75% of prompts, and 100% completed at least 57% of prompts.

Table 8. Comparison to two other studies with similar burden that use bulk-payment incentives.

Table 5 shows that retention was similar across different incentive schemes. However, Table 6 shows a result that may be indicative of waning interest over time. We find a negative effect trending towards significance on the effect of time on completion rate, i.e. completion rate falls over time ($p < 0.096$). The mean completion rate drops from 88% in the 'v-early' period to 83% in the 'v-late' period. We do not find evidence that participants delayed more questionnaires over time. However it is interesting to note that the mean likelihood that a questionnaire is delayed is considerably lower in the first two quartiles (4.81% and 4.31%) than in the last two quartiles (7.40% and 6.0%).

Participant perception

Upon completion of the study, participants complete an End-of-Study questionnaire, which we use to gather subjective feedback about the incentive schemes. Because these responses are collected per-participant, we do not have sufficient data to make statistical inferences in many cases. Nonetheless, we present this data here because we feel it provides insights for study designers into how participants perceive micro-incentive approaches.

Rows 1 and 2 in Table 7 show responses from the participants that relate specifically to micro-payments based incentive schemes. 75% of participants felt more engaged in the study because of the micro-payments attached to each questionnaire, which is encouraging and explains the high compliance that we obtained. However, 53% of participants also would have preferred to receive \$50 at the end of the study, instead of receiving smaller amounts for each question. Intuitively, this makes sense since committing to a micro-payment based study generally implies more risk than committing to a bulk-payment based study, and thus risk-averse individuals are likely to prefer the safer alternative. Furthermore, we note that participants in *Hidden* appear to

prefer bulk payments more than the other groups, implying they perceived more burden than the other groups.

Rows 3 to 5 in Table 7 show the responses that relate specifically to the impact of the *phone* on participants' lives. Most participants did not feel that the phone adversely affected their daily activities or social life, and most participants did not find the phone a nuisance to use. Rows 6 to 8 in Table 7 show the responses that relate specifically to the impact of the *chestband* on participants' lives. Most participants did not feel that the chestband had an adverse effect on their daily activities or social life. Furthermore, *Uniform* participants felt that the chestband was not a nuisance, but participants in *Variable* ($p < 0.03$) and *Hidden* ($p < 0.03$) did see the chestband as a nuisance. This could, perhaps, be attributed to the occasional frustrations *Variable* and *Hidden* participants experienced when they were paid low wages on some questionnaires.

Comparison with bulk-payment based incentives

We now look at how the completion rates we observed in our micro-payment based study compares against the completion rates of bulk-payment based studies. Comparing studies with different population, burden, total payment, questionnaires, devices, etc. is difficult, if not impossible to make in a rigorous manner. Therefore our comparison is merely a high-level comparison of these studies.

Table 8 compares our study against two studies that seem to involve a similar amount of burden [15, 21]. Participants in [21] were required to take measurements, using an on-body sensor, and complete a 58-item questionnaire every 45minutes, during waking hours. Participants got prompted roughly 18-20 questionnaires per day, for which they were compensated about \$42 per day. The study was split into three days for training and feedback, and three for actual data collection. Of the 157 participants in the study, the authors

report that 81% of participants completed at least 75% of all prompts. [15] was a similar study, with the main differences being a slightly shorter questionnaire (45 items), and lower compensation (\$33 per day). This study comprised of 2 periods of 3 day monitoring. The authors report that 88% completed at least 57% of all prompts. Our study offered \$20 per day, which is less than what the others offered. However, our study lasted only 3 days rather than six days, and our questionnaires were shorter and involved only 25 items.

Comparing completion rates across these studies, we see that our compliance is similar to that reported in [21] (81% of participants completed at least 75% of all prompts), and higher than numbers from [15] (they reported 88% of participants completing at least 57% of prompts, whereas ours was 100% for the same metric). This is despite the fact that our study did not involve any pre-screening of participants to enroll only compliant participants unlike [15, 21].

These observations indicate that the completion rates that we obtain are at least comparable, if not better, than those from other bulk payment studies. However, we note that making an authoritative comparison between micro-payment and bulk-payment incentive schemes will require a careful study designed to experiment with both types of incentive schemes under identical conditions.

IMPLICATIONS AND LIMITATIONS

Uniform vs Variable incentives: In terms of overall performance, we see that both incentive schemes perform very well across the board. While one could conclude that this implies the *Uniform* incentive scheme is good enough, the *Variable* incentive scheme can be useful to measure the reservation wage of the population, which can be used to enable better parameter settings for future studies for similar users. For example, we found that the overall cost of the *Uniform* strategy could have been reduced by 50% by using information gleaned from the *Variable* study that similar compliance could be achieved by lower incentives. Thus, we feel that *Variable* is a more powerful incentive strategy than *Uniform*: it provides insight into the relationship between wages and compliance for the population, is robust to poor choices of remuneration, and can be used strategically to improve compliance.

Hidden incentives: The *Hidden* incentive scheme was designed to create a sense of anticipation for larger rewards in participants. However, our results clearly showed that it did not perform well. Participants in this group were less attentive to the questions resulting in lower consistency, and perhaps the frustration over finishing a questionnaire only to find out that the remuneration was less than they had hoped carried over to future questionnaires. One explanation for this is that unlike a lottery where the payoffs are large, the potential payoffs with micro-incentive strategies are relatively modest. Overall, the lack of good performance from the *Hidden* incentive scheme is consistently seen across our results, and suggests that this was, in retrospect, not such a good idea.

Personalized micro-payments: Our analyses show that variable micro-payment incentive schemes provide powerful knobs that study designers can tune in order to affect different performance metrics. One can even imagine taking this one step further and using dynamic pricing models to personalize micro-payment incentives. The incentives can adapt to individuals' reservation wages in order to obtain high levels of compliance while minimizing cost to the study designer. Similarly, micro-payments may be adapted over time to improve compliance and retention for longer studies. We believe that our work lays the foundation for future studies that explore personalization at greater depth, and we plan to pursue these directions in future work.

Optional questions: Our analysis shows several ways in which micro-payments and optional questions interact. Response rate for optional questions can be increased by careful selection of micro-incentives, but variable micro-payments may also lead to adverse behavior. Optional questions can be used, but we recommend some fraction of them be validated through context inferencing on the phone. Examples include: "Are you talking now?", and "Are you walking?", where accurate context inferencing can be performed using accelerometer or other sensor data. Such questions can help determine if an individual is more likely to lie or tell the truth, and can help provide confidence in the data. Our future studies will explore the effectiveness of such techniques.

Limitations and Implications for Future Work: While our study reveals several unanticipated considerations with micro-incentives, it also has several limitations. Perhaps the key limitation of our study is the size of the participant pool. Since we wished to compare several micro-incentives, we had to divide the participant pool into smaller groups, which gave us less data for some comparisons. In addition, this limited us to an exploration of three incentive models; clearly, there are several variants of these models as well as other incentive models that could have been attempted. In particular, a comparison of micro-payment schemes to commonly-used bulk payments would help identify what advantages each approach has over the other. In addition, we need to understand the reactivity or impact of micro-payments on the affective state of the participant. Do high rewards increase participant happiness and vice-versa? In studies that involve affect, this could be undesirable, since affect triggered by micro-incentives could confound the data and reduce its ecological validity. Ultimately, our work should be viewed as a pilot that articulated and explored several questions that need to be addressed for designing more robust micro-payment based incentive mechanisms. We hope that future work drills into these considerations through larger studies.

CONCLUSIONS

In conclusion, this paper presents an in-depth look at several micro-incentive based strategies for high-burden user studies. Overall, we find that micro-incentive strategies play a central role in keeping users engaged and attentive in studies. We saw very high levels of compliance, and little if no retention issues, despite the high burden on the individual to participate in our study. Our study also revealed sev-

eral interesting observations about how incentives impact various aspects of user studies, which we explained using concepts from labor economics. Some of the most interesting observations include: a) our variable incentive strategy allow us to empirically measure if compliance would drop if we had reduced wages, and can be an effective strategy to reduce the cost of user studies, b) we demonstrate that the micro-payment amount and its relation to the reservation wage influences the propensity for users to answer optional questions, and the truthfulness of these answers, and c) we show that the data quality is influenced by the incentive strategy, with consistency of answers likely to suffer if not carefully designed. Our results lay the framework for larger and longer-term incentive studies that we wish to pursue, and more in-depth understanding of how to design micro-incentive strategies that are low-cost and yet provide high compliance, and high data quality.

ACKNOWLEDGEMENTS

We acknowledge the contributions of the following individuals. Anind Dey and Mustafa alAbsi contributed to the study design, while Mary Read, Kelly Peck, Andrew Hoff, Ken Ward, and Satish Kedia helped conduct the study. Emre Ertin, Nathan Stohs, and Siddharth Shah developed the AutoSense sensor suite. Dan Siewiorek, Asim Smailagic, Patrick Blitz, Brian French, and Scott Fisk contributed to the study smartphone software. This work was supported in part by NSF grants CNS-0910878, CNS-0910900, CNS-0855128 funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5), CNS-0939652 and NIH Grant U01DA023812 from NIDA.

REFERENCES

1. Amazon mechanical turk. <http://www.mturk.com>.
2. Gamification.org. <http://gamification.org/>.
3. P. Cahuc and A. Zylberberg. *Labor Economics*. MIT Press, 2004.
4. T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, et al. The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing*, pages 32–41, 2008.
5. T. Christensen, L. Barrett, E. Bliss-Moreau, K. Lebo, and C. Kaschub. A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, 4(1):53–78, 2003.
6. A. Church. Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57(1):62, 1993.
7. R. Cialdini. *Influence: Science and practice*. Allyn and Bacon, 2001.
8. S. Consolvo and M. Walker. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing*, pages 24–31, 2003.
9. C. Ferster and B. Skinner. *Schedules of reinforcement*. Appleton-Century-Crofts, 1957.
10. C. Grady, N. Dickert, T. Jawetz, G. Gensler, and E. Emanuel. An analysis of US practices of paying research participants. *Contemporary clinical trials*, 26(3):365–375, 2005.
11. J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 203–212. ACM, 2010.
12. J. Hopson. Behavioral Game Design, April 2011. http://www.gamasutra.com/view/feature/3085/behavioral_game_design.php.
13. G. Hsieh, I. Li, A. Dey, J. Forlizzi, and S. Hudson. Using visualizations to increase compliance in experience sampling. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 164–167. ACM, 2008.
14. S. S. Intille, E. M. Tapia, J. Rondoni, J. Beaudin, C. Kukla, S. Agarwal, L. Bao, and K. Larson. Tools for studying behavior and technology in natural settings. In *In Proceedings of UBIComp 2003*, pages 157–174. Springer, 2003.
15. T. Kamarck, M. Muldoon, S. Shiffman, K. Sutton-Tyrrell, C. Gwaltney, and D. Janicki. Experiences of Demand and Control in Daily Life as Correlates of Subclinical Carotid Atherosclerosis in a Healthy Older Sample. *Health Psychology*, 23(1):24, 2004.
16. S. Magari, J. Schwartz, P. Williams, R. Hauser, T. Smith, and D. Christiani. The association between personal measurements of environmental exposure to particulates and heart rate variability. *Epidemiology*, 13(3):305–310, 2002.
17. W. Mason and D. Watts. Financial incentives and the performance of crowds. *ACM SIGKDD Explorations Newsletter*, 11(2):100–108, 2010.
18. J. McGonigal. *Reality is Broken: Why games make us better and how they can change the world*. Penguin Pr, 2011.
19. R. Nelson and J. Merz. Voluntariness of consent for research: an empirical and conceptual review. *Medical care*, 40(9):V, 2002.
20. K. Plarre, A. Raji, S. Hossain, A. Ali, M. Nakajima, M. Al'absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, D. Siewiorek, A. Smailagic, and L. Wittmers. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, pages 97–108, april 2011.
21. D. Polk, T. Kamarck, and S. Shiffman. Hostility explains some of the discrepancy between daytime ambulatory and clinic blood pressure. *Health Psychology*, 21(2):202, 2002.
22. R. Pritchard, D. Leonard, C. Von Bergen, et al. The effects of varying schedules of reinforcement on human task performance* 1. *Organizational Behavior and Human Performance*, 16(2):205–230, 1976.
23. K. Pryor. *Don't shoot the dog!: the new art of teaching and training*. Interpet Publishing, 2002.
24. M. Rahman, A. Ali, K. Plarre, M. al'absi, E. Ertin, and S. Kumar. mconverse: Inferring conversation episodes from respiratory measurements collected in the field. In *In Proceedings of the 2nd ACM Wireless Health Conference, San Diego, CA*, 2011.
25. S. Reddy, D. Estrin, M. Hansen, and M. Srivastava. Examining micro-payments for participatory sensing data collections. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 33–36. ACM, 2010.
26. R. Reid. The psychology of the near miss. *Journal of Gambling Studies*, 2(1):32–39, 1986.
27. R. Rivest and A. Shamir. PayWord and MicroMint: Two simple micropayment schemes. In *Security Protocols*, pages 69–87. Springer, 1997.
28. E. Simmons and A. Wilmot. Incentive payments on social surveys: A literature review. *Social Survey Methodology Bulletin*, pages 1–11, 2004.
29. E. Singer. The use of incentives to reduce nonresponse in household surveys. *Survey nonresponse*, pages 163–177, 2002.
30. T. Yamabe, V. Lehdonvirta, H. Ito, H. Soma, H. Kimura, and T. Nakajima. Applying pervasive technologies to create economic incentives that alter consumer behavior. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 175–184. ACM, 2009.
31. B. Yang and H. Garcia-Molina. PPay: micropayments for peer-to-peer systems. In *Proceedings of the 10th ACM conference on Computer and communications security*, pages 300–310. ACM, 2003.