Pay It Backward: Per-Task Payments on Crowdsourcing Platforms Reduce Productivity

Kazushi Ikeda
KDDI R&D Laboratories, Inc.
Fujimino, Japan
kz-ikeda@kddilabs.jp

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

ABSTRACT
Paid crowdsourcing marketplaces have gained popularity by using piecework, or payment for each microtask, to incentivize workers. This norm has remained relatively unchallenged. In this paper, we ask: is the pay-per-task method the right one? We draw on behavioral economic research to examine whether payment in bulk after every ten tasks, saving money via coupons instead of earning money, or material goods rather than money will increase the number of completed tasks. We perform a twenty-day, between-subjects field experiment (N=300) on a mobile crowdsourcing application and measure how often workers responded to a task notification to fill out a short survey under each incentive condition. Task completion rates increased when paying in bulk after ten tasks: doing so increased the odds of a response by 1.4x, translating into 8% more tasks through that single intervention. Payment with coupons instead of money produced a small negative effect on task completion rates. Material goods were the most robust to decreasing participation over time.

Author Keywords
Crowdsourcing; incentives; motivation; crowd work.

ACM Classification Keywords
H.5.3 Information interfaces and presentation (e.g., HCI): Group and Organization Interfaces.

INTRODUCTION
The rise of paid microtask crowdsourcing platforms such as Amazon Mechanical Turk [1] has spread a norm of piecework (per-task) payment. Each microtask on the platform is priced individually, and workers are paid a base rate multiplied by the number of correctly completed tasks. This norm of payment-per-task has remained relatively unchallenged, with platforms such as Crowdflower, mClerk [3] and Clickworker adopting the same model. Researchers have largely focused their investigations within this model, for example finding that higher payment leads workers to complete more tasks [26] and that incentivizing agreement with other workers has no impact on accuracy [34].

In this paper, we ask: is piecework payment a well-grounded approach? We draw on the behavioral economics literature to suggest several alternatives. First, *goal-setting* experiments have established that specific, ambitious goals lead to higher task performance than general, easy goals [24], in part because people wish to avoid losing funds or effort that they have already invested [6]. Second, not all payments must come as earnings: coupons can also attract participation [18]. So, discounting the cost of a necessity such as a monthly phone bill may be a viable alternative to traditional payment schemes. Third, material goods (e.g., catalog gifts) can also function as incentives [31]. Might crowdsourcing markets offer material goods as an alternative to cash? These investigations can challenge the default design of crowdsourcing marketplaces.

The present study aims to examine whether these alternate incentive approaches implied by economic and psychological theory result in higher performance. It measures the likelihood of responding to time-limited mobile crowdsourcing tasks under incentive conditions that (1) reward per task vs. reward in bulk only after ten tasks have been completed, (2) reward with money vs. reward with a coupon for a cellular phone bill, and (3) reward with material goods vs. money. Behavioral economics would predict that payment in bulk increases participation [24], and both coupons [18] and material goods [31] are less effective than money.

Compared to being paid per task, participants increased the odds of completing a task by 1.4x when paid in bulk, translating to 8% more tasks completed. This result means a 16% relative increase in task completion relative to how crowdsourcing markets operate today through this change to the payment approach. Counter to prediction, completion rates stay higher over time with material goods than with cash earnings. Finally, as predicted, payment with a coupon resulted in a small (non-significant) negative effect on task completion rates.

The rest of the paper is organized as follows. We first review studies on motivation in crowdsourcing and economic theories to develop our research question and

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858327

hypotheses. We then introduce the experimental design of our field experiment and present quantitative and qualitative results. Finally, we describe limitations and future work.

CROWDSOURCING PLATFORMS AND MOTIVATION

Whatever else their motivations, microtask crowd workers seek to earn money [4]. They seek out tasks that maximize their expected earnings [16]. Online marketplaces offer them quick and efficient payment [42] in exchange for their effort. Requesters (clients) on the platform likewise aim for rapid and correct results. However, both this rapidity and accuracy can break down on paid crowdsourcing markets.

These breakdowns can be mediated through monetary or non-monetary means. Monetary incentives tend to motivate workers to perform more tasks. The simplest approach to increasing productivity is simply to pay more: this was one of the first widely-cited results in crowdsourcing [26]. However, what is the most effective way to deploy that extra money to maximize participation? More complex schemes now include banking bonuses and paying them out periodically instead of immediately [11]. Beyond this, game theory and auction theory offer shared incentives and conditions under which workers are properly incentivized to participate [36, 41]. Payment need not even be certain: lotteries can attract many participants for the task [30]. However, participants contribute more hours with piecework payments than with lotteries. Or, sometimes the way to achieve more work is to temporarily require less work: taking time to relax during long sequences of tasks alleviated worker fatigue and significantly improved worker retention rate [32].

Surprisingly, monetary incentives can even increase short-term contributions to *intrinsically* motivated projects such as citizen science [27]. However, this mixing of incentives still lessens intrinsic motivation, especially for newcomers [27, 14]. Competition-based payment works well for skilled workers, whereas norm-based payment motivates novice workers who dislike competition [28]. This combination of approaches can work in reverse, too: designing for social engagement amongst workers (thus producing an intrinsic motivation to participate) improves retention rates [40]. Paid crowd workers, properly incentivized, can match or exceed the work of intrinsically motivated unpaid volunteers [25].

Monetary incentives are also a common route to ensure accuracy. If workers perform poorly on Amazon Mechanical Turk (AMT), requesters can reject their work and refuse to pay. However, simply offering higher payment does not increase quality [26]. Neither does paying only when workers agree with each other [34]. Offering to train workers and give feedback can improve quality [40, 12], as can including gold standard ("attention check") tasks with known answers [22]. These strategies gave rise to even more unorthodox approaches that increase accuracy, for example paying accuracy-based bonuses only when the task changes form [39] or even offering financial incentives for new workers to stop and quit [15].

Outside of marketplaces, crowd participation rates are likewise an issue. Participants will increase their effort when the system reminds them of how unique their contribution is and when given challenging goals [7]. For prosocial platforms such as elder volunteers, social contributions and identities on the community are key factors for continuous participation [18]. Sudden social needs such as the disappearance of an academic colleague can mobilize thousands of volunteers [2]. Lacking this exogenous motivation, many have instead turned to games to motivate contribution. The ESP Game [36] pioneered this design space with an image-labeling game. Foldit [9] transformed arcane protein structuring tasks into games, drawing thousands of people to help fold proteins in weeks that took scientists years. Lacking a social need or an available game design, coercion (for security purposes) works as well: ReCAPTCHAs [37] are tasks that determine whether a user is a human or a bot, used to digitize books on the side. Blocking users from accessing an application until they contribute does boost participation rates, but it also causes many users to leave [35].

Some of these non-monetary methods have already made their way into commercial services. However, since their mechanisms are often task dependent, their effects are limited to specific tasks and are difficult to apply to new goals. For this reason, most existing general crowdsourcing platforms use paid incentives. In this paper, we examine alternative incentives within a paid crowdsourcing framework.

Taken together, this literature paints a picture where increased salaries or game-theoretic techniques are necessary to increase worker productivity in paid crowdsourcing markets. In volunteer crowdsourcing communities, however, behavioral manipulations from social psychology have been successful. There is a rich literature of behavioral economics that suggests similar effects may be possible here. This observation motivates our research question:

RQ. Can behavioral economic techniques increase worker productivity in microtask crowdsourcing markets?

In this paper, we explore behavioral economic theories and study their impact on workers.

THE BEHAVIORAL ECONOMICS OF PRODUCTIVITY

The preceding results were mostly generated *tabula rasa* within the crowdsourcing literature. However, there is a rich history of research in behavioral economics and psychology that can orient us toward effective approaches to increase effort. We focus on interventions that test either *when* to pay (bulk payment) and *what* to pay (coupons and gifts). This section presents a review of studies on these behavioral economic results in order to establish hypotheses.

Payment in bulk

Goal setting theory proposes that specific and ambitious goals motivate us [23]. When completing tasks under a time limit, harder tasks and specific goals — instead of telling participants to "do your best" — led to higher work efficiency. Goal setting theory has been applied to organizations in several economic studies. For example, when truck drivers were asked to carry logs with their best effort, the average weight was 60% of the legal maximum truckloads. When they set a specific goal of 94% of the legal maximum weight, the average weight was increased to 90% [21].

The sunk cost effect may be one of the mechanisms behind the effectiveness of goal-setting when money is at stake. Once people have invested funds or effort, it becomes difficult for them to leave without achieving their final goal [6]. In a canonical sunk cost study, participants were more likely to recommend spending one million dollars to complete a project that had already cost ten million, and was likely to fail, than to spend the same one million dollars on the same project when the ten million dollars were not already invested.

Badges can also operate as challenging yet specific goals within social computing sites. Badges increase participation in sites such as Stack Overflow, and steer users' behavior toward acquiring the badge [5].

This literature prompts the hypothesis:

H1. Bulk payment after several tasks rather than per-task in paid crowdsourcing platforms will lead to increased effort in terms of the number of tasks completed.

Payment with coupons

We may consider not only what increments to pay in, but also what form that payment should take. Payment on Amazon Mechanical Turk is escrowed in an Amazon Payments account for use on Amazon.com, but can in many cases be transferred to a bank. In that sense, Mechanical Turk is a blend between a coupon (to Amazon.com) and cash. We seek to identify which method will be most effective. Comparing cash to coupons has precedence in health adherence studies. For example, 83% of participants offered $15 in cash and 66% of participants offered a $15 food coupon attended an AIDS prevention session [10]. This result prompts H2:

H2. Paying with coupon instead of earning new cash in paid crowdsourcing platforms leads to decreased effort in terms of the number of tasks completed.

Payment with material goods

Some efforts, for example summer reading or fundraising drives for elementary school children, succeed by using material goods such as catalog gifts rather than cash. While most workers on crowdsourcing platforms seek money [4], concrete material goods may in some cases be substitutable. However, in other domains, cash has more impact on participation rates than goods [31, 25, 13]. For example, cash incentives resulted in higher response rates for a face-to-face and mail-based questionnaire than material goods [31]. A survey of medical studies indicates that material incentives do increase participation, but cash may increase it more [25].

From these previous results, non-monetary incentives might motivate crowd workers, but cash may do so more. In particular, this research suggests:

H3. Material goods of equivalent value instead of cash payments in paid crowdsourcing platforms will lead to decreased effort in terms of the number of tasks completed.

EXPERIMENTAL DESIGN

To investigate whether the behavioral economics theories of bulk payment, coupons and material goods can be applied to increase crowdsourcing productivity, we ran a between-subjects mobile crowdsourcing experiment in April 2015. In the experiment, we recruited subscribers of a mobile telecommunications company to install a paid crowdsourcing application that sent them tasks intermittently throughout the day, and randomized the incentives for responding across participants. We measured whether users responded to each task within a prespecified time limit. We summarize our experimental conditions in Table 1 and describe the details as follows.

Participants

We recruited participants (N=300, 41.2% female, aged 19 – 61, mean age=36.3, std. dev=8.0) who were cellular phone subscribers with a major telecommunications company in Japan. Japanese residents engage heavily with their cellular phones while commuting, waiting for others, or standing in line. Collaboration with this telecom company was beneficial because it meant all participants had a necessity payment that we could discount in the experiment: their monthly cellular phone bill. Japanese citizens are also familiar with tiered gift catalogs (with, $10 gift options, $20 options, and so on), and purchasing gift credit for someone else is a common gift strategy. We recruited participants through email advertisements sent to approximately 5,000 staff working for a recruiting company. We required that participants have an Android or iPhone to participate. Participants' demographics (including age and income) are similar to those of average workers in Japan.

Method

Since we cannot easily examine alternative incentives on existing crowdsourcing platforms, we ran the experiment on our own crowdsourcing platform. The platform operates similarly to others such as Gigwalk in offering payment for location-sensitive tasks. We believe that the findings from this platform generalize, since the incentives are not based on the nature of the task. However, applicability to each crowdsourcing platform should be examined.

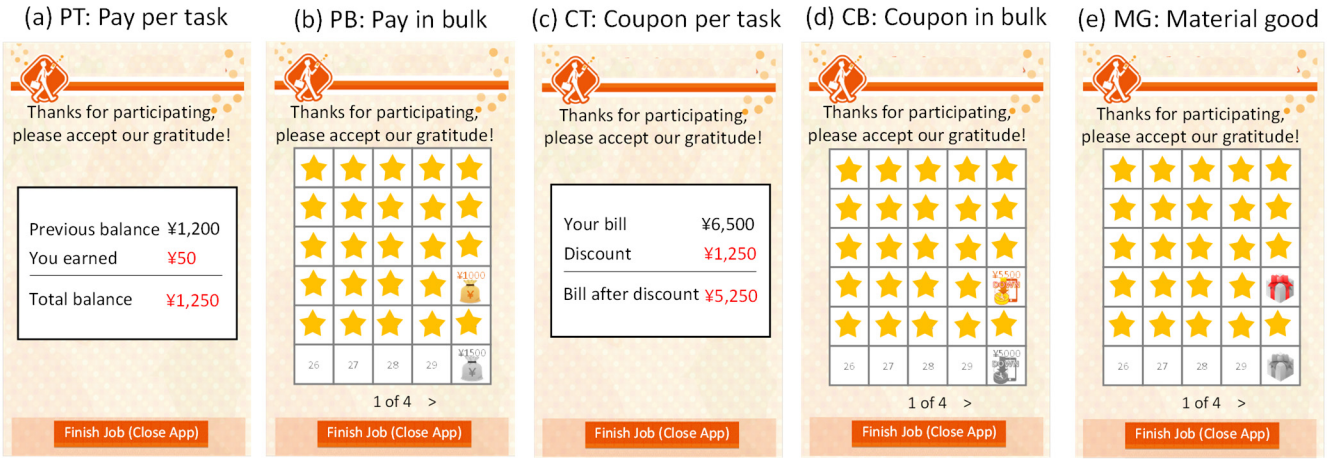

Figure 1. Screenshots of the proposed incentive methods. (a) PT: Pay per task (existing method), (b) PB: Pay in bulk, (c) CP: Coupon per task, (d) CB: Coupon in bulk, (e) MG: Material good. (Messages on screens are translated from Japanese.)

Task	Quality of service survey
Required time	2~3 minutes / task
Notification	Average 5 times / day
Incentive price	50 yen / task (or 1 stamp / task)
Measurement	Task completion and task quality
Participants	300
Duration	20 days in Mar 30 ~ May 11, 2015

Table 1. Summary of experimental settings. We evaluated task completion rate and task quality as indicators of worker motivation.

All participants installed an application on their phones that was branded as a quality-of-service survey application from the telecommunications company. They continued to use it for twenty days. During these three weeks, the application passively tracked workers' location using the phone's GPS. When participants were in an area with low quality of service (known to have few phone reception "bars" or many dropped calls), it would trigger a phone notification with a survey. The survey asked workers about their current cell phone quality of service (Table 2). When workers received the notification, they could either accept it or ignore it. Since tasks are location dependent, the invitation expired five minutes after notification.

The application would trigger whenever the user was in a low quality-of-service area. However, it was limited to a minimum (3) and maximum (5) of notified tasks per day so that all participants were notified roughly the same number of times. The task took two to three minutes to complete, and was worth ¥50 (roughly forty cents, or $10/hr of work). This payment rate aligned with current payment goals on Amazon Mechanical Turk [33], and minimum wage in Tokyo (¥900/hour). Since the average response rate was about 50% in pilot experiments, the average participant was expected to earn ¥3,000 through the study. Upon installing the application, all participants answered a demographic survey and received ¥500 (or 10 stamps) in response.

1. What are you doing now? (Single selection) A. ○ Moving ○ Staying in one place B. ○ Train ○ Outside ○ Inside
2. How is your quality of service? (Single selection) ○ Excellent ○ Good ○ Fair ○ Poor ○ Very poor
3. Please describe the details of your situation (Free text) e.g. "I'm at station [X] on the [Y] line"

Table 2. Quality of service survey task (translated from Japanese).

Conditions

We anchored our study in a 2 (pay or discount) x 2 (per task or bulk) design, adding a fifth material good condition. Participants were randomly assigned to a condition at the start of the experiment. All participants were fully informed of the mechanism and the form of payment about the assigned method before starting the experiment. After participants submitted each task, the application showed a summary page detailing the participant's earnings (Figure 1).

Each condition had a different payment screen. Our control condition was *pay per task* (PT), representing current practice on microtask labor platforms. In this condition, participants saw the incremental payment for the task they completed, as well as their total earnings so far. At the end of the experiment, workers received their accumulated payment to the participants' Amazon account. Given the topic of this paper, we acknowledge that there are differences between payment through cash transfers and a certificate. In this case, however, the payment path mirrors how Mechanical Turk operates. *Pay in bulk* (PB) showed a stamp sheet interface and only paid out in ¥500 increments, after workers completed every ten tasks. Completing fewer than ten tasks in a set at the end of the study resulted in no payment for that set. For most participants, ten tasks would mean responding to every request for two days.

Coupon per task (CT) looked like pay per task, but instead of increasing their total earnings, participants worked to discount their monthly cellular phone bill. After each task, the application would show how much of their phone bill remained. The cost of the bill was initialized to their previous month's charge. Since it was difficult to programmatically interface with the company's payment infrastructure, at the end of the experiment these workers received a gift certificate that they could apply to their cellular bill. *Coupon in bulk* (DB) looked like pay in bulk, but likewise metered down their bill every ten tasks rather than counting up the earnings after each task.

The *material good* (MG) condition looked like the bulk interface, but each set of ten tasks upgraded the gift that was available in a popular gift catalog. Catalog gifts are an appropriate choice here because they are popular on celebratory occasions in Japan: senders pay the fee of a catalog gift and receivers can freely select any one gift from the hundreds of gifts on the catalog. Catalog gifts have different price tiers. More expensive tiers offer higher quality gifts. We allowed workers to work toward gift tiers in the price range of ¥500 to ¥5,500, at the same ¥500-increments as the bulk condition. However, since knowledge of gift prices may reduce their effect on worker motivation, we hid the actual price of each tier and instead allowed participants to view the content of the catalog gifts on their mobile phones. Available gifts included clothes, accessories, furniture, electronics, kitchenware, gourmet and vacation experience. Material goods could not be crossed with the other dimensions because goods cannot be sent out in partial quantities.

Measures

We recorded an observation each time the application notified the participant that a new task was available. Our primary dependent variable is *task completion,* which we operationalize as a binary variable: whether each task notification is accepted and completed. Task completion rate is good measurement to compare the incentive methods, and similar methods have been used in prior work (e.g., Rogstadius et al. [29]). Since all experimental conditions are completely the same except for the method of payment, conditions that result in more completed tasks are indicative of stronger motivation.

We also measured task correctness and response time as *task quality*. For task correctness, we manually verified free-text answers reporting the detail locations of the workers (e.g. "I'm at station [X] on the [Y] line") because it was the most time consuming question. We sampled 2000 tasks (5 methods x 20 users x 20 tasks), and manually classified the answers into three classes; (i) high quality, (ii) low quality, and (iii) unknown or unclassifiable comparing the GPS location collected with the answers. We dropped the unknown labels (iii) and defined a task correctness rate as (i) / ((i) + (ii)). Two people labeled these responses blind to condition; any disagreements were resolved by a third rater.

Finally, we recorded participants' gender, income, phone use frequency, phone use hours, commuting method, and commute length.

Following the twenty days of the experiment, we debriefed participants via a survey. In the survey, we described all five conditions and asked participants to rank their preference if they were to continue using the system. We then asked participants with an open-ended survey what they liked most about their top-rated method, and what they liked least about their bottom-rated method.

Method of Analysis

To determine how incentive scheme relates to task completion rate, we performed a logistic regression with task completion as the dependent variable and conditions encoded as three binary variables (pay or discount, per task or bulk, and money or material goods). The variables were coded such that the intercept (control) was pay per task, corresponding to the typical approach on microtask crowdsourcing marketplaces. We added controls for demographic variables: age, gender, income, phone usage frequency, phone usage time, and commute time. We also added a control for day of the study, which would help model any novelty effects. Before running the model, we removed data from participants who were not in enough low quality-of-service locations to trigger the application, corresponding to fewer than forty recorded notifications throughout the twenty-day experiment.

We built a separate linear regression model with the same independent variables and controls, but open-ended response length as the dependent variable. In this model, we have removed participants from analysis who completed tasks fewer than 10 times throughout the experiment period.

We also calculated basic summary statistics such as the mean task completion rate across users, in aggregate and per day, in each of the five conditions.

Finally, we compared the subjective survey rankings to the empirical effectiveness of the five conditions. We correlated these two rankings against each other via a rank correlation. We also grouped and themed the free text responses per strategy.

RESULTS

A total of N=275 participants completed the study and had more than forty notifications across the twenty days. Among this group, the median participant received 87 notifications, or 4–5 per day. Kruskal-Wallis tests confirmed that randomization was effective and no demographic categories were significantly different between conditions (all $p>0.05$).

	Average task completion rate		
Condition	**Mean**	**Std. Dev**	**Median**
Pay per task	0.49	0.228	0.515
Pay in bulk	0.57	0.238	0.595
Coupon per task	0.45	0.227	0.417
Coupon in bulk	0.53	0.241	0.548
Material good	0.54	0.263	0.588

Table 3. Mean, standard deviation, median of participants' average task completion rate across conditions. Paying in bulk increased participants' mean task completion rate.

	Task completion		
Variable	**Odds ratio**	**SE**	**p-value**
(Intercept)	0.52	0.37	0.076 .
Main effects			
Bulk	1.36	0.13	0.024 *
Coupon	0.82	0.14	0.137
Material Good	1.18	0.19	0.376
Controls			
Day of study	0.98	0.00	0.000 ***
Age	1.01	0.01	0.114
Gender (1=Male)	1.17	0.16	0.326
Income (in ¥1,000,000)	1.01	0.05	0.853
Phone use frequency	1.01	0.00	0.096 .
Phone use time (hour)	1.11	0.06	0.090 .
Commute (min)	1.00	0.00	0.193

Table 4. The logistic regression (N= 19,520) predicting task completion. Paying in bulk per ten tasks increased the odds of responding to a task by 1.36 times.

Task Completion

First, we examine the relationship between incentive condition and task completion. Aggregating by user, Figure 2 and Table 3 report the average task completion rate in the five incentive conditions across the twenty days of the study.

Our main analysis examines how three variables, 1) per task or bulk, 2) pay or discount, and 3) money or material goods, impacted the probability of responding to the task notification within five minutes and completing the task. Table 4 reports the logistic regression coefficients for the fitted model. Odds ratios describe a multiplicative factor on the baseline intercept odds by which a notification is more or less likely to be responded to. In our model, pay per task is the intercept, or baseline. For example, suppose the baseline odds were 2:1, meaning two nonresponses for every response. If a factor has an odds ratio of 1.1, it means that the odds when that factor is true would shift to (2*1.1) : 1, or 2.2 : 1.

Paying in bulk has a statistically significant increase in the odds of a response, increasing the odds by a factor of 1.36 ($p<0.05$). In absolute terms, participants paid in bulk had an 8% higher task completion rate than those paid by task. This increase corresponds to a 16% relative increase in response rate. Thus, H1 was supported: bulk payment increases task completion rates.

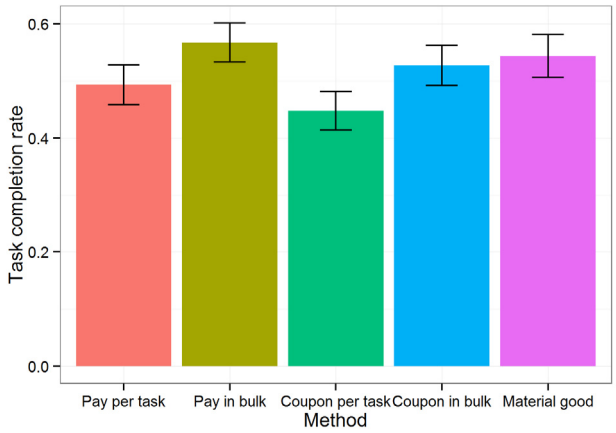

Figure 2. Task completion rates. Payment in bulk had a higher completion rate than the baseline payment per task.

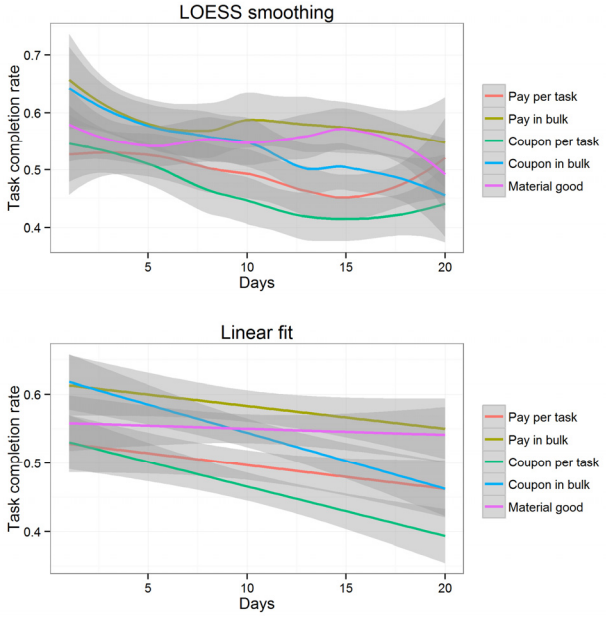

Figure 3. Time series variation of task completion rate (Upper: smoothing with LOESS curve. Lower: fitting with linear method). Material good maintains task completion rate throughout the experiment period whereas task completion rates are declining in other methods as time goes.

Providing the payment with a coupon for necessity payment rather than with cash had a slight trend toward a negative impact (4% absolute), but not statistically significant ($p>0.05$). The material good condition had a moderate absolute effect (5%), but due to the large variance across participants, the effect was likewise not significant. H2 and H3 were not supported: framing payment as discount rather than earnings, or as material goods rather than money, does not significantly affect task completion rates in our data.

	Average task correctness		
Condition	**Mean**	**Std. Dev**	**Median**
Pay per task	0.892	0.120	0.900
Pay in bulk	0.886	0.122	0.900
Coupon per task	0.863	0.177	0.922
Coupon in bulk	0.866	0.163	0.900
Material good	0.898	0.100	0.912

Table 5. Mean, standard deviation and median of task correctness across conditions. Incentive methods did not affect task quality significantly.

	Average response time		
Condition	**Mean**	**Std. Dev**	**Median**
Pay per task	49.7	17.6	45.5
Pay in bulk	47.4	15.3	44.3
Coupon per task	53.8	23.2	49.6
Coupon in bulk	54.4	22.7	50.7
Material good	42.9	10.2	42.1

Table 6. Mean, standard deviation and median of response time (in seconds) across conditions. Incentive methods did not affect task quality significantly.

Figure 3 shows the task completion rates per day across the conditions. As the novelty of the study wore off, task completion rates declined across all conditions. However, the material good condition (best fit slope = -0.0009% per day, or nearly 2% over the twenty days) declines at a slower rate than the others, for example than pay per task (best fit slope = -.00348% per day, or 7% over twenty days) or pay in bulk (best fit slope = -.00330% per day; 7% over twenty days).

Task Quality

We measured task correctness and response time as indicators of task quality. Table 5 shows mean, standard deviation and median task correctness rates for the five incentive methods throughout the experimental period. In this experiment, participants are required to answer their location and situation using free text. We manually verified the task correctness by comparing the free-text answers with the GPS location collected with the answers. An ANOVA revealed no significant difference between the methods ($p>0.05$, $F(4, 94)=0.40$, $p=0.80$).

Table 6 shows response time, where we analyzed mean time to complete tasks for each worker. Mean time and standard deviation of response time are in seconds. Likewise task correctness, no significant difference was observed ($p>0.05$, $F(4, 233)=1.53$, $p=0.19$).

These models found no significant main effect for the bulk, coupon or material good variables. Thus, there was no observable effect of incentive mechanism on task correctness and the amount of extra time that workers invested in the task. This result echoes prior work in that payment can impact volume of work but not accuracy [26].

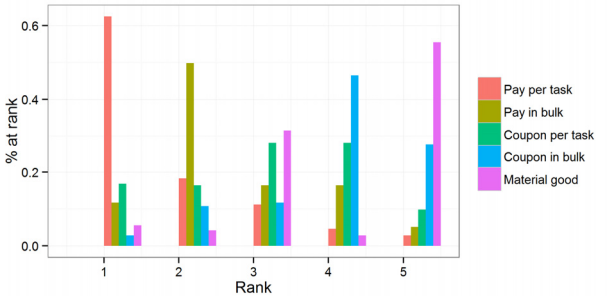

Figure 4. Payment per task ranked the highest in a post-study survey, though in practice it was not the strongest condition.

Condition	**Mean Rank**	**Median Rank**	**Actual**
Pay per task	1.67	1	4
Pay in bulk	2.53	2	1
Coupon per task	2.98	3	5
Coupon in bulk	3.86	4	3
Material good	3.99	5	2

Table 7. Participants ranked payment per task as the option they would most likely take, though it earned less money for them in our study.

Subjective Rankings

Figure 4 shows the distribution of participants' subjective ranks when they were shown all five incentive conditions and asked to rank their preferences after the study. Of the 275 participants who finished the study, 213 participants (77.5%) answered the questionnaire. Table 7 shows the mean, standard deviation and median ranks, split by the participants' actual condition.

Averaging rankings across participants from most desired to least, participants preferred pay per task most, then pay in bulk, coupon per task, coupon in bulk, and finally material good. More than half of participants (62%) reported that they would most prefer the pay per task model in general. Half (50%) listed pay in bulk as the second most preferred method, and more than half (55%) listed material good as their least preferred method. However, looking at average task completion rates from the experiment, the empirical ranking that produced the most completed tasks is: pay in bulk, material good, coupon in bulk, pay per task, and coupon per task. We do not yet have sufficient data to argue statistically that this ranking is robust, but the differences are striking. For example, the worst two conditions in our study were ranked #1 and #3 in the survey. The Spearman's rank-order coefficient between the two lists is -0.2.

Notably, the ranking correlation is negative, indicating that empirical earnings and subjective preferences are trending in reverse order from each other. In many studies, such a result would indicate that participants were more performant in one condition but likely would enjoy another condition more. In this study, the line is blurrier because if money is their main goal, workers maximize their own utility by choosing the option that has the highest earning

Pay per task: positive	**Pay per task: negative**
Easy to understand the price of payment	Don't feel much sense of accomplishment with small price
Can buy whatever I like at Amazon	I don't use Amazon frequently
Pay in bulk: positive	**Pay in bulk: negative**
Feel motivated to achieve milestones	Difficult to understand the price of payment
Fun, like a game	Any unused fraction is ignored
Can buy whatever I like at Amazon	I don't use Amazon frequently
Coupon per task: positive	**Coupon per task: negative**
Easy to understand the payment system	Difficult to understand the price of payment
Never forget to use the payment	I don't feel like I'm getting money
I want to save on my mobile phone charge	Don't want to save on my mobile phone charge
Coupon in bulk: positive	**Coupon in bulk: negative**
Easy to understand milestones	Difficult to understand the price of payment
Feel motivated to achieve milestones	I don't feel like I'm getting money
Easy for automatic discount	Any unused fraction is ignored
Material good: positive	**Material good: negative**
Easy to understand what I can get	The number of items in the catalog is limited
Fun to get a tangible goods	I can't find attractive item on catalog
Easy to understand visually	Any unused fraction is ignored

Table 8. Representative opinions from the follow-up survey, translated from Japanese.

potential. Future work might show workers the quantitative results from this study before asking them to choose.

Analysis of the free text explanations produced some insight into these rankings. Table 8 reports common opinions from the survey, translated from Japanese. Paying in bulk produced a sense of achievement: *"Since there were clear milestones, I thought I wanted to do more tasks to reach the next goal."* Discounting the cellular phone bill produced split opinions, with some workers claiming it as a major positive, (*"I must pay a mobile phone charge every month no matter what I do"*) and others wanted payment in return for their contribution (*"A discount on my cellular phone bill does not motivate because I don't feel like I'm earning additional income"*). This result suggests that preference for coupon may have significant individual differences in its motivational effect. There was a similar trend in the material good condition: some participants felt motivated by tangible goods (*"I could enjoy my participation with a goal of my preferred gift. Tangible goods made me feel rewarded for my work."*) whereas other participants did not want to use it because they could not find many items that they liked in the catalog.

DISCUSSION

We began with three hypotheses: H1, bulk payment leads to increased effort in terms of the number of tasks completed; H2: coupons lead to decreased effort; and H3: material goods of equivalent value lead to decreased effort.

While we observed a significant increase in task completion with concrete, challenging goals (H1), more discussion is required to better understand the risk it trades off for its stronger incentive. The bulk method can be unfair because a worker cannot receive a reward if they quit before reaching 10 tasks, and a worker may not complete the batch. We report worker sentiments in Table 8. There may be ways to split the difference — e.g., obtain the psychological motivation but minimize potential harm — by offering some percentage of the full payment if the worker chooses not to complete the full set of ten tasks.

Our hypothesis that coupons reduces the productivity on crowdsourcing (H2) was not statistically supported in the experiment. One possible reason for this result is that coupons for necessities such as cellular phone bills can feel similar to cash. Future work can establish whether other coupons might be suitable replacements to cash in some paid crowdsourcing platforms.

We hypothesized that material goods decrease the productivity of crowd workers compared to cash (H3). This hypothesis was not supported. Some economic studies insist that the motivation caused by material goods is minor compared to monetary payment [31, 7]. When asked to answer a face-to-face questionnaire and offered $5 cash or a park ticket worth $12 as gratuity, the response rate was higher with cash than a park ticket [31]. The suggestion is that motivations caused by material goods are risky because different people have uneven values they would attach to material goods. Likewise, response rates for mailed questionnaires are higher when the incentive is sent with the questionnaire than when it is sent afterwards, and it is higher with monetary payment than with material goods [8]. Given that our results suggested material goods might be

more robust to novelty effects, future work can engage in a longitudinal evaluation longer than three weeks.

Based on the large individual differences in preference ranking, future systems might consider allowing workers to choose their preferred awards. For example, in our survey responses, some participants commented that mobile phone charge reductions and catalog gifts are not major motivators for them. It is entirely possible that the effect of coupons and material goods would be higher with different bills or goods. For example, the monthly cellular phone bill was larger than participants could make through participation in the study, and this likely threatened motivation.

Other crowdsourcing platforms and firms may modulate the design of these proposed incentives. For example, our study was unusual in that we could pair with a company that charges users a monthly bill. Others who charge monthly payments for infrastructure or subscriptions could do this. Alternatively, workers could perform tasks to pay off train or bus charges, or mobile payment platform (e.g., Venmo) debts, while waiting in line. These scenarios may work well but require proper evaluation.

Material goods succeed when they enables workers to view their contributions in a non-monetary frame, or to imagine a concrete return for their work. In this paper, we used graded catalog gifts worth 500 yen to 5,500 yen so that workers could find something they liked. We expect that this method could apply to digital stores and in-app purchases as well.

Our worker pool and the tasks used in the study may not be representative of regular paid crowd workers and tasks on crowdsourcing platforms. Strictly speaking, all crowdsourcing platforms have different worker demographics according to the amount and types of tasks provided on the platforms. Even Mechanical Turk and Crowdflower have vastly different demographics currently. However, we suspect that the differences we saw in our study would hold externally. We recruited our participants from a manpower supply company similar to those that power crowdsourcing platforms such as Clickworker, and their demographics (including age and income) were similar to those of average workers in Japan, just as Mechanical Turk's American worker demographics bear some similarity to those in the United States. As shown in Table 4, in addition, we haven't observed any significant difference in the preference of incentive methods among demographics. Since our participants did not contain students, retired or unemployed workers, the effect of the proposed incentives for these people may differ because their reasons for working on crowdsourcing platforms may differ.

Another limitation is that our task may not be representative of all tasks on platforms such as Mechanical Turk. However, there is some evidence for generalizability: "survey" is one of the top twenty keywords on Mechanical Turk [17]. While there are of course many different kinds of surveys on AMT, there are tasks which are similar to ours, and the results extend to traffic reporting, store stock reporting, and many others. In addition, much like on other crowdsourcing marketplaces, the task was repeated many times, just the input (context) changed. Unlike regular paid crowdsourcing tasks where workers can work whenever they want, our task was a mobile sensing task where workers could work only when they are notified of new tasks. This difference may reduce the effect of our interventions because workers could not "streak" and complete many tasks at once. Complementing this study on more traditional crowdsourcing marketplaces may lend additional insight.

CONCLUSION

In this paper, we have investigated whether incentive methods of bulk, coupons and material goods that are based on behavioral economic theories could be applied to crowdsourcing systems. These approaches, respectively, suggest (1) setting a multi-task goal and paying all at once, (2) working toward discounts rather than earning cash, and (3) offering material goods rather than money. We evaluated these approaches through an experiment deployed using mobile crowdsourcing quality-of-service survey tasks to 300 participants. The conventional pay per task method was not the most performant. Instead, payment in bulk after sets of ten tasks had an 8% higher task completion rate than (16% increase relative to) payment per task. The material good method showed a high retention, where participation decreased very little throughout the experimental period. The coupon conditions appeared to have a negative effect on task completion rates, but this effect was not significant due to large variation between participants. Survey results after the study suggested that participants' opinions were not always in line with the empirical participation rates.

Today's crowdsourcing markets overwhelmingly orient themselves around piecework payment. Our research suggests that the literature on behavioral economics would be a powerful complement to today's focus on computational workflows and mechanism design in the creation of crowdsourcing markets.

We summarize the contributions of our study as follows:

- We have proposed alternative incentive methods that could improve workers' motivation beyond the dominant approach in crowdsourcing markets today.
- We have carried out a mobile crowdsourcing field experiment with 300 participants and tested the advantages of the proposed methods.
- We have proposed methods that can be applied to existing crowdsourcing platforms by altering the payment mechanism, and can lead to more work throughput and (as a result) higher earnings by workers.

So far, we have carried out our mobile crowdsourcing experiment in Japan, targeting subscribers of a Japanese telecommunication company. We believe this choice was instructive, as crowdsourcing is an increasingly global

phenomenon. However, we note that the effects of our proposed methods may differ by country and culture [36]. We seek to draw on cross-cultural studies of goal setting theory, coupons, and material goods to test these incentives across cultures a global crowdsourcing market. In addition, we aim to identify factors other than the theories in this paper in order to improve satisfaction and effort in crowd work. For example, some crowdsourcing studies have reported that social relations motivate workers [19]. Other economic studies have focused on awards as non-monetary incentives [20]. Combining the proposed methods with these findings and providing a pro-social, long-term sustainable career option for crowd workers is core to the future of crowd work.

ACKNOWLEDGEMENTS

We would like to thank Yasuyuki Nakajima, Masatoshi Suzuki, Yasuhiro Takishima, Kazunori Matsumoto and Hiromi Ishizaki of KDDI R&D Laboratories, Inc. for supporting the study. This work was supported by a National Science Foundation award IIS-1351131. We would like to thank Stanford HCI group members for their valuable comments. We also thank all of the participants in our experiment.

REFERENCES

1. Amazon Mechanical Turk, https://www.mturk.com/
2. Help Find Jim. http://www.helpfindjim.com
3. Aakar, G., Thies, W., Cutrell, E. and Balakrishnan, R. mClerk: enabling mobile crowdsourcing in developing regions. In Proc. CHI 2012, 1843-1852.
4. Antin, J. and Shaw, A. Social desirability bias in reports of motivation for US and India workers on Mechanical Turk. In Proc. CSCW, 2011.
5. Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec, J. Steering user behavior with badges. In Proc. of WWW, 2013, 95-106.
6. Arkes, H. R. and Blumer, C. The psychology of sunk cost. Organizational behavior and human decision processes, 35, 1, 1985, 124-140.
7. Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P. and Kraut, R. Using social psychology to motivate contributions to online communities. In Proc. CSCW, 2004, 212-221.
8. Church, A. H. Estimating the effect of incentives on mail survey response rates: A meta-analysis. Public Opinion Quarterly, 57, 1, 1993, 62-79.
9. Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Fay, A. L., Baker, D. and Popovic, Z. Predicting protein structures with a multiplayer online game. Nature, 466, 7307, 2010, 756-760.
10. Deren, S., Stephens, R., Davis, W. R., Feucht, T. E. and Tortu, S. The impact of providing incentives for attendance at AIDS prevention sessions. Public health reports, 109, 4, 1994, 548.
11. Difallah, D. E., Catasta, M., Demartini, G. and Cudré-Mauroux, P. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In Proc. HCOMP, 2014.
12. Dow, S., Kulkarni, A., Klemmer, S. and Hartmann, B. Shepherding the crowd yields better work. In Proc. CSCW, 2012, 1013-1022.
13. Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Pratap, S., Wentz, R. and Kwan, I. Increasing response rates to postal questionnaires: systematic review. 324, 7347, 2002, 1183.
14. Frey, B. S. and Jegen, R. Motivation crowding theory. Economic Surveys, 15, 5, 2001, 589-611.
15. Harris, C. G. The effects of pay-to-quit incentives on crowdworker task quality. In Proc. CSCW 2015, 1801-1812.
16. Horton, J. J., and Chilton, L. B. The labor economics of paid crowdsourcing. In Proc. of the 11th ACM conference on Electronic commerce, 2010, pp. 209-218.
17. Ipeirotis, P. G. Analyzing the amazon mechanical turk marketplace. XRDS: Crossroads, The ACM Magazine for Students, 17, 2, 2010, 16-21.
18. Kane, R. L., Johnson, P. E., Town, R. J. and Butler, M. A structured review of the effect of economic incentives on consumers' preventive behavior. American journal of preventive medicine, 27, 4, 2004, 327-352.
19. Kobayashi, M., Arita, S., Itoko, T., Saito, S. and Takagi, H. Motivating multi-generational crowd workers in social-purpose work. In Proc. CSCW, 2015, 1813-1824.
20. Kosfeld, M. and Neckermann, S. Getting more work for nothing? Symbolic awards and worker performance. American Economic Journal, 3, 3, 2011, 86-99.
21. Latham, G. P. and Baldes, J. J. The practical significance of Locke's theory of goal setting. Applied Psychology, 60, 1, 1975, 122-124.
22. Le, J., Edmonds, A., Hester, V. and Biewald, L. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In Proc. CSE, 2010, 21-26.
23. Locke, E. A. Toward a theory of task motivation and incentives. Organizational Behavior and Human Performance, 3, 2, 1968, 157-189.
24. Locke, E. A. and Latham, G. P. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. American psychologist, 57, 9, 2002, 705-714.
25. Lutge, E. E., Wiysonge, C. S., Knight, S. E. and Volmink, J. Material incentives and enablers in the

management of tuberculosis. Cochrane Database System Review, 1, 2012.

26. Mason, W. and Watts, D. J. Financial incentives and the "performance of crowds". In Proc. HCOMP, 2009, 77-85.
27. Massung, E., Coyle, D., Cater, K. F., Jay, M. and Preist, C. Using crowdsourcing to support pro-environmental community activism. In Proc. CHI, 2013, 371-380.
28. Preist, C., Massung, E. and Coyle, D. Competing or aiming to be average?: Normification as a means of engaging digital volunteers. In Proc. CSCW, 2013, 1222-1233.
29. Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. and Vukovic, M. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In Proc. ICWSM, 2011.
30. Rula, J. P., Navda, V., Bustamante, F. E., Bhagwan, R. and Guha, S. No one-size fits all: Towards a principled approach for incentives in mobile crowdsourcing. In Proc. HotMobile, 2014.
31. Ryu, E., Couper, M. P. and Marans, R. W. Survey incentives: Cash vs. in-kind; face-to-face vs. mail; response rate vs. nonresponse error. Public Opinion Research, 18, 1, 2006, 89-106.
32. Rzeszotarski, J. M., Chi, E., Paritosh, P. and Dai, P. Inserting micro-breaks into crowdsourcing workflows. In Proc. HCOMP 2013.
33. Salehi, N., Irani, L. C. and Bernstein, M. S. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In Proc. CHI 2015, 1621-1630.
34. Shaw, A. D., Horton, J. J. and Chen, D. L. Designing incentives for inexpert human raters. In Proc. CSCW 2011, 275-284.
35. Tomasic, A., Zimmerman, J., Steinfeld, A. and Huang, Y. Motivating contribution in a participatory sensing system via quid-pro-quo. Proc. CSCW, 2014, 979-988.
36. von Ahn L. and Dabbish L. Labeling images with a computer game. In Proc CHI, 2004, 319-326.
37. von Ahn, L., Maurer, B., Mcmillen, C., Abraham D. and Blum M. reCAPTCHA: Human-based character recognition via Web security measures. Science, 321, 5895, 2008, 1465-1468.
38. Weber, E. U. and Hsee, C. Cross-cultural differences in risk perception, but cross-cultural similarities in attitudes towards perceived risk. Management Science, 44, 9, 1998, 1205-1217.
39. Yin, M., Chen, Y. and Sun, Y. A. Monetary Interventions in Crowdsourcing Task Switching. In Proc. HCOMP 2014.
40. Yu, L., André, P., Kittur, A. and Kraut, R. A comparison of social, learning, and financial strategies on crowd engagement and output quality. In Proc. CSCW 2014, 967-978.
41. Zhang, Y. and Schaar, M. Reputation-based incentive protocols in crowdsourcing applications. In Proc. INFOCOM 2012, 2140-2148.
42. Zyskowski, K., Morris, M. R., Bigham, J. P., Gray, M. L. and Kane, S. K. Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In Proc. CSCW 2015, 1682-1693.