The Effects of Pay-to-Quit Incentives on Crowdworker Task Quality

Christopher G. Harris
Department of Computer Science
SUNY Oswego
Oswego, NY 13126
christopher.harris@oswego.edu

ABSTRACT
Companies such as Zappos.com and Amazon.com provide financial incentives for newer employees to quit. The premise is that workers who will accept this offer are misaligned with their company culture, which will therefore negatively affect quality over time. Could this pay-to-quit incentive scheme align workers in online labor markets? We conduct five empirical experiments evaluating different pay-to-quit incentives with crowdworkers and evaluate their effects on mean task accuracy, retention rate, and improvement in mean task accuracy. We find that the number of times a user is prompted for the inducement, the type and frequency of performance feedback given to participants, the type of incentive, as well as the amount offered can help retain high-performing workers but encourage poor-performing workers to quit early. When we combine the best features from our experiments and examine their aggregate effectiveness, mean task accuracy is improved by 28.3%. Last, we also find that certain demographics contribute to the effectiveness of pay-to-quit incentives.

Author Keywords
Risk Aversion; Human Computation; Crowdsourcing, Relevance Assessment; Incentive Schemes; Pay-to-quit Incentives

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Interaction Styles; J.4 Social and Behavioral Sciences: Economics

General Terms
Human Factors; Economics; Experimentation

CSCW 2015, March 14–18, 2015, Vancouver, BC, Canada.

http://dx.doi.org/10.1145/2675133.2675185

INTRODUCTION
Online labor markets have facilitated the performance of *microtasks* – concise, simple, and easily-defined tasks that humans can perform well but computers cannot, such as annotating images, performing relevance assessments, or validating common sense facts. These labor markets have demonstrated value, dramatically lowering costs, reducing task completion times and, provided the task is well explained and quality checks are in place, the results can be of high quality. For these reasons, online labor markets, such as Amazon Mechanical Turk (MTurk), have proliferated, serving as matchmaker between task requester and worker.

For centuries, economic theory has held that rational workers will choose to improve their outcome in response to incentives [22, 31]. Over the same period of time, techniques to enhance worker motivation have been discussed, such as in the writings of Industrial Age philosophers Robert Owen and Jeremy Brentham, as well as through experiments by behavioral psychologists such as Frederick Herzberg and B.F. Skinner. More recently, companies known for their service have using incentives early on to separate employees motivated by personal financial gain from those who are not, under the premise that the former will not provide the same commitment to corporate culture and loyalty than the latter. Zappos.com, a company regarded for its top-notch customer service, has provided a pay-to-quit incentive of $4,000 to newly-hired employees, called "The Offer" [21, 28]. This offer provides an opportunity for newbies to quit with no questions asked. However, this opportunity is rarely acted upon, with fewer than three percent of new employees accepting the offer. Other companies have offered similar incentives with varying results. In May, 2012, start-up wine seller L18.com offer employees pay-to-quit incentives of up to a month's salary with six of the company's 85 employees taking the offer [18]. Amazon also offers its employees $1,000 per year of employment, up to $5,000, as a pay-to-quit incentive [26]. In June 2014, Riot Games, maker of the popular "League of Legends" game, announced it would offer employees up to $25,000 to quit [3].

Many researchers that have examined workplace employment incentives have concluded they must work, since they are still being offered and these offers have expanded to new firms. This gives rise to some important questions: Can these same pay-to-quit incentives be structured for online labor markets to obtain similar increases in productivity? Also, which motivational techniques, such as financial incentives, can be used in online labor markets to improve work quality?

The study of incentive schemes, particularly those that attempt to discriminate between high-performing workers and poorly-performing workers, has received scant scholarly attention. Although there is a fair amount of theoretical guidance in the literature, remarkably few empirical studies have been conducted on incentive schemes. We offer the following contributions. First, using a game format, we conduct a series of experiments to see what factors may induce poorly-performing crowdworkers to quit before the end of the task, improving the mean crowdworker accuracy. Second, borrowing from game show design and established theories of economic behavior, we introduce pay-to-quit inducements and observe if they can encourage poorly-performing workers to withdraw from the tasks early while retaining the high-performing workers. This is the first empirical evaluation of the effects of pay-to-quit incentive schemes in crowdsourcing tasks. Third, as collaborative work continues its global expansion, an understanding of the role of these incentive schemes plays across demographics may influence how task designers develop incentive schemes that increase task quality. To this end, we examine worker demographics and examine their role in the performance of incentive schemes.

This paper is organized as follows. In the next section, we discuss the related work, background and motivation for our study. We then develop our research questions and explain our experimental design. Last, we discuss our findings and conclude and indicate planned future work.

BACKGROUND AND MOTIVATION

In this section, we discuss some of the elements involved in our work, and why they are important to our study.

Online Labor Markets

Online labor markets, such as MTurk, are crowdsourcing platforms in which requesters list tasks called HITs or "human intelligence tasks" along with the compensation to be provided. The compensation is usually in the range of a few cents per task. Individual workers (called crowdworkers or simply workers) can then elect to perform a task; upon completion workers are compensated by the task requester. Requesters also may provide additional incentives to workers in the form of a bonus payment.

Incentives

Under certain circumstances the provision of financial incentives can undermine "intrinsic motivation" (e.g. enjoyment, desire to help out), possibly leading to poor outcomes [11, 16]. Locke [24], presented evidence that it is not monetary incentives, time limits, or knowledge, but a person's intrinsic motivation (such as their goals and intentions) that have the largest effect on task performance. To this end, we examine the role of feedback in pay-to-quit inducements.

In crowdsourcing, Mason and Watts examined financial incentives on two MTurk tasks in [27]. They found that greater financial incentives increase the quantity, but not the quality, of work performed. The authors indicate this could be due to the result of the "anchoring" effect, where workers who were paid a small amount believed the value of their work to be greater, and were less motivated to produce quality than those doing it without compensation. Shaw, Horton and Chen examined the role of incentives on non-expert raters using controlled experiments on MTurk in [33]. They evaluated 14 incentive schemes designed to increase mean task accuracy. Two of these incentive structures significantly improved results. We use information from their study design as a foundation for our own experiments, expanding the number of treatment options on their "punish-agreement" incentive scheme. Harris examined various types of financial incentives (only positive, only negative, or both positive and negative incentives with MTurk in a resume evaluation task finding the hybrid scheme outperformed the other incentive schemes [14].

Risk Attitude

According to economic theory, risk attitude plays an essential role in evaluating incentives under uncertainty. Several studies provide insight on people's view of risk. Kahneman and Tversky investigated how risk is evaluated by people in decision-making events [20]. They found that people prefer avoiding losses to acquiring gains by a factor of nearly two-to-one. Recent studies by others have shown loss aversion does not occur as frequently as previously thought. Gal illustrated that loss aversion phenomena are more likely to result from inertia than from loss/gain asymmetry, but found that loss aversion may be more salient when people are in a competition [9]. Gil and Prowse [10] demonstrated that people are loss averse in a competitive environment that involves real effort. Holt and Laury examined risk aversion and the effects of incentives and found that when the payout was tangible (i.e., paid in cash), people became more risk averse [17]. Others have investigated risk and loss aversion in game show environments. Hartley *et. al.* and [15] Lam *et. al.* [23] examined risk attitude on the popular game show "Who Wants to Be a Millionaire." Each discovered that risk aversion is affected by the scale of the financial incentive at risk as well as the probability of success. We note that game show design typically involves a rapid increase in task difficulty as the game proceeds, making the amplified risk apparent to the player through various clues. Traditional online labor market task design, on the other hand, rarely provides an increasing level of difficulty or reward to the

worker; however, as a worker proceeds through a set of tasks, their effects on task quality become magnified for the requester.

Pay-to-Quit Inducement Amounts

In the literature, most discussion regarding pay-to-quit inducements (also called *reservation* or *walk-away* amounts) comes from negotiation theory and game show risk attitude studies. Van Puocke and Beulens explain in [34] that a reserve amount is an indifference point where a negotiator should be indifferent between accepting the offer or ending the negotiation (i.e., accepting the walk away amount). It therefore represents the lowest outcome a negotiator is willing to accept. In this paper, we regard the pay-to-quit incentive as the compensation offered to a worker if they choose to quit before the completion of the task. This amount is typically a percentage of the cumulative pay + bonus earned by that worker. Since the inducement represents "certain compensation," the amount should be set so that highly-performing workers would be better off refusing the inducement and continuing with the task, while poorly-performing workers would be better off accepting it and quitting the task early.

RESEARCH QUESTIONS

Our objective in this study is to evaluate pay-to-quit incentives in online labor markets and evaluate how they affect non-expert worker quality. Specifically, we examine the following five research questions.

1. Does performance improve if participants are given explicit windows in which they can accept the pay-to-quit incentive, as Amazon.com does, or is permitting the right to leave with an incentive at any time the better approach? If explicit windows are the better choice, how many should be offered?
2. Understanding the participant's choice to keep playing or to take a pay-to-quit incentive involves rational decision-making, which works best with performance feedback. How frequent should this feedback be provided in order to discriminate the two groups? Does the information provided need to be truthful?
3. When requesters list a task on an online labor market, they often seek to maximize worker quality (with time and cost also being considerations). How should incentive bonus schemes be structured so that we only retain the highest-performing workers? Should it be additive (start with zero bonus and add to it for each successive correct answer) or subtractive (start with a large bonus and deduct for each incorrect answer)?
4. What amount should the bonus be to induce the poorly performing participants to quit but not the highly-performing participants?
5. Are there certain demographics that are more responsive to pay-to-quit incentives than others?

EXPERIMENTAL DESIGN

Beat the Clock or Walk

To evaluate the role of incentive structure in mean task accuracy, we used a game, "Beat the Clock, or Walk" (abbreviated BTCOW). In relevance assessment tasks, games have shown the ability to provide quality inputs with less spam and at lower cost than crowdsourcing [8]. Games with a purpose (GWAP) are often designed for repetitive tasks such as labeling images [35] or evaluating common sense facts [36]. Perhaps most importantly, the game format allows us to dynamically evaluate different bonus incentives with the crowd that would be difficult to accomplish through a standard crowdsourcing platform.

In each round of BTCOW, a worker is presented with an image or short movie description, along with five potential movie titles, with only one being a correct choice. The player (worker) is instructed to select the best choice. Figure 1 displays a screenshot from the BTCOW game.

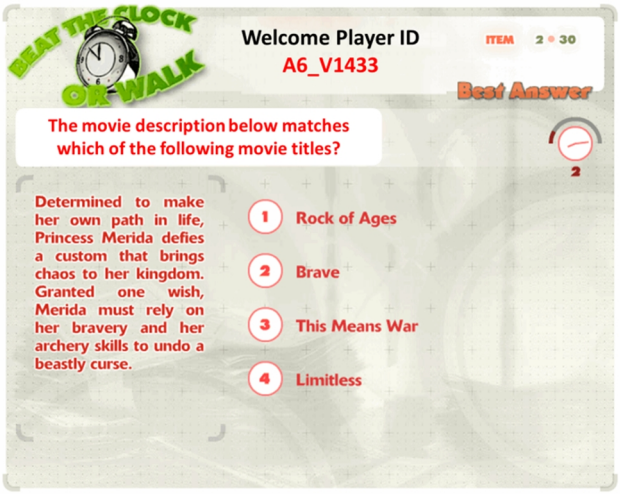

Figure 1. A screenshot from the BTCOW game

Participant Assignments

We recruited 1200 participants from MTurk. Experiments 1 through 4 used 60 participants for the control group and 60 for each treatment (some experiments have multiple treatments). Experiment 5, which applied the best features from each experiment against a baseline, used 120 participants for the baseline and 120 for the treatment group.

Participants were randomly assigned to a control condition or one of several treatment conditions. The same thirty questions and the answer choices for each question were randomly assigned to each user to avoid question order and position bias (each question is considered a *round;* all thirty questions comprise the *task*). Participants could only play the BTCOW game once; our game logged the MTurk User ID and the IP address of each participant. Only the participant's first attempt was evaluated; any suspected duplicate efforts were removed from our study and the task was relisted for other participants.

For each of the 30 rounds, workers have 15 seconds to make a selection. Three consecutive non-selections indicate abandonment of the game, this occurred in 116 (9.6%) of games played; when this occurred, the participant is considered to have withdrawn from the task, the data was not used, and the task was relisted. This time limit is provided to enhance *flow*, a concept introduced by Csikszentmihalyi [5] to describe the delicate balance between anxiety (due to a task that is too difficult) and boredom (due to a task that is too easy). In an earlier pilot task, we found that 80 percent of participants could make a comparable selection within 15 seconds. We believe that this, coupled with the financial incentive offered in the game, provide a sufficient ongoing balance of flow to a majority of participants. There is also an additional benefit to imposing a time limit: In the game, descriptions of movies are provided and participants must select the movie title to match the description; providing a time limit restricts the ability to do external searches and thus limits the ability for participants to cheat.

Participants were compensated $0.30 for completing the task and answering the pre- and post-task surveys. This represents $0.01 per question – a typical compensation offered on MTurk for this level of effort. The amounts that are under risk ($0.30) are minimal, but as Gal and Prowe pointed out in [10], people in a competitive situation where payments are tangible (i.e., in cash) often make decisions without carefully considering the magnitude of their losses. We found our participants made decisions within the game that showed effort beyond the amount of money under risk.

Performance Measurements

Mean accuracy is the most appropriate quality indicator for this task; we therefore measure *mean accuracy* across the k participants for each treatment as follows.

$$Accuracy_i = \sum_{i=1}^{k} \frac{Correct_i}{Answered_i} \quad (1)$$

In order to evaluate the effects of removing those participants who accept our pay-to-quit incentive offer, we provide two different calculations for mean accuracy; one in which answers selected by participants accepting the incentive are removed, another where they are not removed. Answers selected by participants who withdraw from the game, but do not accept the offered incentive, are not considered.

The task was designed to be challenging to differentiate performance levels between participants. In order to evaluate the number of participants that complete the entire task, we also evaluate the *retention rate*, which is the percentage of participants who answer each question. Last, we look at the *gain* in mean accuracy between those that accept the pay-to-quit incentive and those that complete the task. A larger difference indicates the feature being tested is more effective to separate the two groups.

Demographic Survey of Participants

A brief survey of all participants was conducted, asking questions about gender, country of residence, income, experience with games, and if crowdsourcing was their primary source of income. Of the 1200 workers participating in our main study, all provided this information. Once the game was completed participants were given a three-question survey on the player's game experience. A total of 1153 workers provided this information. We asked if the time given to answer questions was sufficient, if the task was challenging enough, and for those players that chose to leave before the game's completion, we asked their reasons for doing so.

EXPERIMENTS

We conduct a series of four sequential experiments to evaluate our research questions, using the outputs of the earlier experiments as inputs to the later experiments.

Experiment 1 –The Role of Explicit Pay-to-Quit Offers

All participants are able to quit the task at any time; however, they are only able to receive $0.01 per question if they leave when prompted by the game. How many pay-to-quit windows are appropriate? The number of windows is likely to affect flow; too many would be a distraction to the task's objective, too few would have participants finish the task that otherwise may be in their best interest to quit. Therefore, we want to determine how many were appropriate in a thirty-question task.

Participants (N=240, Female = 48%) were randomly assigned to either a control group (N = 60) or to one of three treatment groups (N = 60 each). The control group was provided a message before the task began indicating they could quit at any time and receive the $0.01 per question, but did not provide any prompts during the task.

Figure 2. A screenshot of the pay-to-quit pop-up window, displaying the information participants see.

The first treatment group was provided a pop-up window after the 15th question (i.e., half-way through the task), requiring the participant to either accept or refuse the offer

in order to proceed (see Figure 2 for a screenshot of the pay-to-quit incentive pop-up). The second treatment group was provided the same pop-up at one-third and two-thirds through the task (i.e., after the 10th and 20th questions). The third treatment group was offered an identical pop-up after the 7th, 14th and 21st questions. There was no time limit to accept/decline the pay-to-quit incentive. This addresses our first research question.

Experiment 2 – The Role of Performance Feedback

Participants (N =300, Female = 45%) are randomly assigned to a control group or one of four treatment groups. The control group is only provided feedback at the end of the task. Feedback consisted of the number of questions the participant got correct and the average number correctly chosen for all participants. Treatment group 1 is provided feedback during two pop-ups (after the 10th and 20th questions) and asked if they wanted to quit and receive an incentive. Treatment group 2 is accurately provided feedback after responding to each of the 30 questions. Treatment group 3 is provided correct feedback on the number correctly answered, but told they outperformed the average, regardless of their true performance, during two pay-to-quit pop-ups. Treatment group 4 is provided the opposite of Treatment group 3; they are told their performance was worse than the average, regardless of their true performance. This addresses our second research question.

Experiment 3 – The Role of Incentive Type

Participants (N = 180, Female = 47%) were randomly assigned into a control group or one of two treatment groups. The control group was not provided any additional pay-to-quit incentive, but at two pay-to-quit windows during the game, they are offered $0.01 for every question answered. Treatment group 1, modeled after Amazon.com's pay-to-quit incentive, is offered an *increasing* incentive during the two pay-to-quit windows; they are told that they would receive $0.01 per question answered plus $0.01 per question they got correct. Treatment group 2 is offered a *decreasing* incentive during the two pay-to-quit windows; they are offered *n* cents, with *n* defined as:

$$n = c * (30 + \#incorrect\ answers - \#answered) \quad (2)$$

where the magnitude constant, c, is 0.5 for this experiment. We believe this decreasing incentive will induce participants to quit early if they initially perform poorly. This addresses our third research question.

Experiment 4 – The Role of Incentive Size

Participants (N = 240, Female = 44%) are randomly assigned to a control group or one of three treatment groups. The control group is provided with $0.01 per question answered and no additional incentive. Treatment group 1 is offered a smaller *decreasing* incentive (i.e., c in Equation 2 is set to 0.5) in the two pay-to-quit pop-up windows that appear after the 10th and the 20th questions). Treatment group 2 is offered a moderate *decreasing* incentive (i.e., c = 0.75). Treatment group 3 is offered a larger *decreasing* incentive (i.e., c = 1). This addresses our fourth research question.

Experiment 5 – Putting it All Together

In this last experiment, we take the best strategies from each experiment and combine them into a single "best" strategy, and run them against a control group. The control group allows participants can quit at any time and receive payment of $0.01for each question attempted, but no other incentives are offered.

RESULTS AND DISCUSSION

Experiment 1 – The Role of Explicit Pay-to-Quit Offers

Mean accuracy for the control group and each of our 3 treatment groups is reported in Table 1. The mean and standard deviation for mean accuracy for *all participants* (the answers for quitters are not removed), for *finishers* (only those not taking the pay-to-quit incentives are included; the answers for quitters are removed), and for *quitters* is provided. A bolded mean value indicates the difference from the control group is statistically significant at $p<0.05$.

A two-tailed independent samples t-test was used to evaluate the effects of the number of prompts on mean accuracy. There was a significant difference between our control group at the $p<0.05$ level for the two-prompt condition, $t(118) = 2.138$, $p = 0.0346$, for the finishers, $t(91) = 3.100$, $p = 0.0024$, for the quitters, $t(25) = 2.426$, $p = 0.0168$. Cohen's *d* was 0.41, 0.63, and 0.51 for these three values, respectively, which indicates a moderate effect size. None of the other treatment groups provided a significant difference from the control group for mean accuracy.

Table 1. Mean accuracy and standard deviations for the control group (N=60) and three treatment groups (each N=60), each offering a different number of inducement prompts.

	Control		1 Prompt		2 Prompts		3 Prompts	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
All	.647	.175	.631	.187	**.708**	.135	.612	.165
Finishers	.653	.171	.639	.193	**.738**	.126	.610	.178
Quitters	.645	.194	.597	.164	**.575**	.111	.619	.095

In the control group, as well as in the three-prompt treatment, there was little difference between the mean accuracy for the participants that chose to accept the pay-to-quit incentive and those that refused it. This is surprising, as the rational choice for most high-performing participants would be to continue. There is a slight difference for the single-prompt treatment and a significant difference in the two-prompt treatment. In the single-prompt case, even the high-performing participants could obtain payment for half their work (since they were halfway through the task when the prompt appears) if they are risk-averse. In the three-prompt treatment, our data shows many of the poorly-performing participants did not take any of the three offers, even though they were consistently underperforming from the beginning. Having multiple opportunities to quit may

downplay their significance more than if only one or two opportunities are offered. This indicates that of the four models explored, the two-prompt model appears to be the best at discriminating the highly-performing and the poor-performing participant groups.

Although Table 1 indicates how using two-prompt treatment is significantly better for both finishers and quitters than the control group, we are more interested to see if we can *discriminate* between the two cases; that is, see a large non-overlapping difference between the finishers and quitters in our treatment group. If the task finishers and task quitter groups are very distinct, it indicates that particular treatment is helpful for raising mean accuracy, since we can eliminate the results from the group with lower mean accuracy. Figure 3 illustrates the differences between the finishers (F) and quitters (Q) for the control and the two-prompt treatment groups. In both cases, as expected, the finishers outperform the quitters. Additionally, the amount of overlap between the two unshaded boxplots on the right (representing the two-prompt model) is smaller than the overlap between the two shaded ones on the left (representing the control group). This visually indicates our two-prompt treatment is effective to retain the better performers while eliminating the poorer performers.

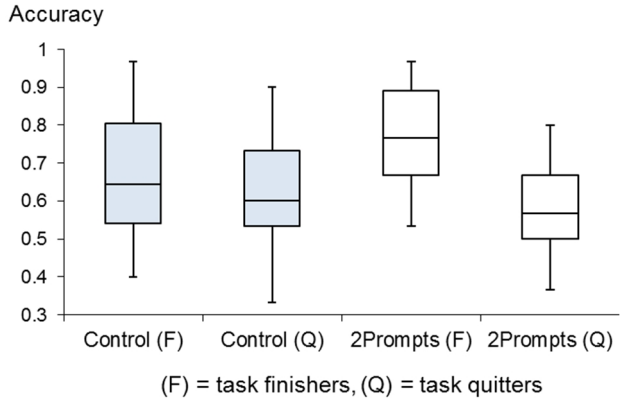

Figure 3. Box-plot comparison of mean accuracy between the control (light blue) and two-prompt group (unshaded)

We also examined the retention rate. Overall, our four groups for this experiment averaged a 79% retention rate, with the control group obtaining a 77% retention rate, and treatment groups 1 2, and 3 obtaining retention rates of 82%, 78% and 80% respectively. This consistency across groups indicates that as we offer a greater frequency of pay-to-quit incentive prompts, it does not encourage more participants to quit.

The gain in mean accuracy for the finishers as compared with all participants was -1.9%, 2.3%, 2.4% and 0.7% for our control group and each of our three treatment groups, respectively. Thus, the second treatment group, offering two prompts to accept the pay-to-quit incentive, provides the best discrimination power, with the one-prompt treatment group closely following. The control group actually eliminated its better-performing participants before the end of the game, lowering the overall mean accuracy score. We note going forward that the two-prompt treatment group with its higher mean accuracy scores, and better discriminatory power, is the best of these choices.

Experiment 2 – The Role of Performance Feedback

We report the mean accuracy for the control group and each of our 5 treatment groups in Tables 2 and 3. In each table, we report the mean and standard deviation for mean accuracy for all participants, for finishers and for quitters.

In Table 2, we compare the mean accuracy for those in the control group (where feedback is provided only once at the end of the task), Treatment group 1, where it is reported at the end of each question, and those in Treatment group 2, where feedback is reported thrice (after the 10th question, the 20th question, and at the end of the game). A bolded mean value indicates the difference from the control group is statistically significant at $p<0.05$. Boxplots to illustrate the differences between groups shown in Figure 4

Table 2. Mean accuracy and standard deviations for the control group (N=60) and two treatment groups (N=60), each offering a different frequency of feedback to the participant.

	Control		Each Round		Each Prompt	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**
All	.629	.159	.649	.176	**.686**	.153
Finishers	.636	.157	.668	.177	**.712**	.151
Quitters	.603	.167	.583	.171	.559	.159

A two-tailed independent samples t-test was used to evaluate the effects of the feedback frequency on mean accuracy. There was a significant difference between our control group at the $p<0.05$ level for the providing feedback at each pay-to-quit prompt condition for all participants, $t(118) = 2.001$, $p = 0.0477$ and for the finishers, , $t(96) = 2.631$, $p = 0.0096$. Cohen's *d* for these values were 0.37 and 0.49, respectively, indicating a moderate effect size. None of the other values provided a significant difference from the control group for mean accuracy.

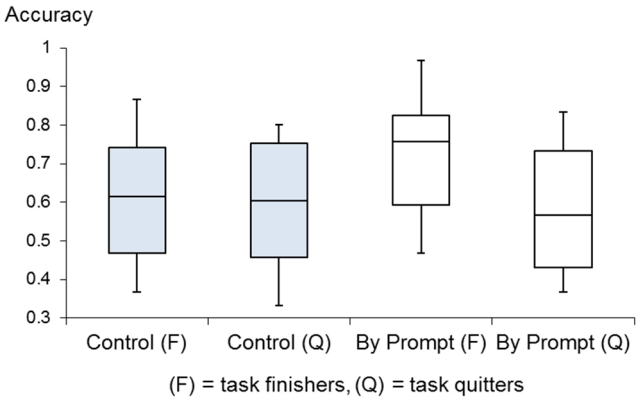

Figure 4. Box-plot comparison of mean accuracy between the control (light blue) and feedback after each prompt (unshaded).

The retention rate varied only slightly between the three treatment groups, with the accurate, feedback each round, and feedback each prompt treatment groups retaining 77%, 77%, and 73% respectively. The gain in mean accuracy for the finishers as compared with all participants was 0.9%, 2.3%, and 3.1% for each of our three treatment groups, respectively. The treatment the largest discrimination power is the feedback at each prompt treatment. We carry this forward to the next part of Experiment 2.

In Table 3, we evaluate the effects of feedback on *relative* performance on a participant's choice to accept a pay-to-quit incentive during two pay-to-quit prompts. We compare the group in which they are given an accurate indication of performance (Treatment group 2), those told that they are outperforming the group average (Treatment group 3) and those that are told they are underperforming the group average (Treatment group 4). A bolded mean value indicates the difference from Treatment group 2 is statistically significant at $p<0.05$.

Table 3. Mean accuracy and standard deviations for three treatment groups (each N=60), each providing a different feedback message during the pay-to-quit incentive prompt.

	Accurate		Outperform		Underperform	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**
All	.658	.153	**.717**	.157	.627	.151
Finishers	.712	.151	.727	.155	.695	.155
Quitters	.559	.159	**.698**	.167	.502	.143

A two-tailed independent samples t-test was used to evaluate the effects of the feedback type given to participants on mean accuracy. There was a significant difference between Treatment group 2, which provided accurate feedback, at the $p<0.05$ level for the outperform condition for all participants, $t(118) = 2.08$, $p = 0.039$, and for those that chose to accept our pay-to-quit incentive, $t(23) = 2.37$, $p = 0.025$. Cohen's d was 0.37 (a moderate effect) and 0.85 (a strong effect), for all participants and the quitters, respectively. None of the other treatment groups provided a significant difference in mean accuracy from Treatment group 2.

In Figure 5, we compare the finishers and quitters for those who receive accurate reporting of their performance (Treatment group 2) with those who are told they are outperforming the group average (Treatment group 3). The distinction between finishers and quitters is clear in both models, but more pronounced for Treatment group 3.

The retention rate varied between more these treatment groups than any other, with the accurate, outperform and underperform groups retaining 73%, 85%, and 65% respectively. This is a bit surprising, since all participants are paid based on their own performance, regardless of how they perform relative to other participants.

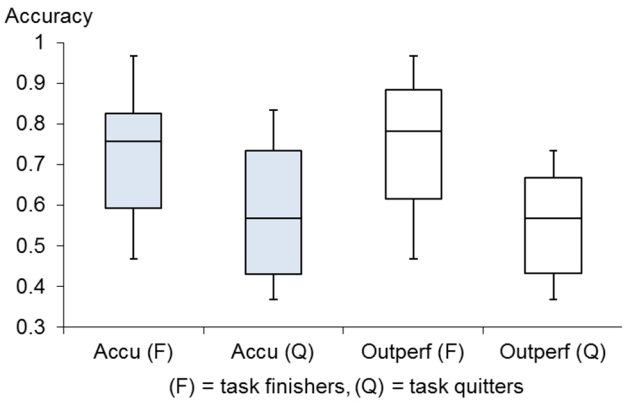

Figure 5. Box-plot comparison of mean accuracy between those accurately reporting task performance (light blue) and those told they are outperforming the group average (unshaded).

The difference between the mean accuracy for finishers and all participants was 5%, 1%, and 7% for each of our three treatment groups, respectively. Thus, the treatment the largest discrimination power is the under-encouraging treatment, with the policy of accurately providing feedback close behind. The 1% gain by providing over-encouraging feedback does not persuade the best performers to improve their performance any further, but the 7% gain by the underperforming feedback group weeds out many of the poorly-performing participants, leaving many of the stronger participants to complete the task.

Experiment 3 – The Role of Incentive Type

Mean accuracy for the control group and our two treatment groups is reported in Table 4. A bolded mean value indicates the difference from the control group is statistically significant at $p<0.05$.

Table 4. Mean accuracy and standard deviations for a control group (N=60) and two treatment groups (each N=60), each offering a different incentive type.

	Control		Increasing		Decreasing	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**
All	.648	.156	.682	.171	.653	.151
Finishers	.656	.152	.707	.172	.699	.149
Quitters	.618	.170	.573	.164	.525	.158

A two-tailed independent samples t-test was used to evaluate the effects of the frequency of feedback on mean accuracy. There were no a significant differences between our control group and the treatment groups at $p<0.05$. Neither the increasing nor the decreasing incentive types were significant, therefore we do not provide a box-plot comparison.

The retention rates for this experiment were 78%, 82%, and 72% for the control, increasing and decreasing incentives, respectively. This indicates that a greater number of participants (28%) chose to take the pay-to-quit incentive for the decreasing incentive, which we believe demonstrates its effectiveness.

The gain in mean accuracy for the finishers as compared with all participants was 0.8%, 2.5%, and 4.9% for the control group and two treatment groups, respectively. Thus, the decreasing incentive is the treatment with the largest discrimination power. Overall, the decreasing incentive is the most valuable to us, and therefore we use it in Experiments 4 and 5.

Experiment 4 – The Role of Incentive Size

Mean accuracy for the control group and our three treatment groups, which use a different value of our magnitude constant, *c,* to set our decreasing incentives, is reported in Table 5. Bolded mean values indicate the difference from the control group is significant at $p<0.05$.

A two-tailed independent samples t-test was used to evaluate the effects of the frequency of feedback on mean accuracy. There was a significant difference between our control group and finishers given the medium incentive, $t(91) = 2.181$, $p = 0.032$ and between our control group and finishers given the larger incentive, , $t(94) = 2.366$, $p = 0.020$, at the $p<0.05$ confidence level. Cohen's *d* was 0.57 and 0.63, respectively, indicating a moderate effect size. None of the other values provided a significant difference from the control group for mean accuracy.

Table 5. Mean accuracy and standard deviations for the control group (N=60) and two treatment groups, (each N=60), each offering a different incentive size.

	Control		Small (0.5)		Med (0.75)		Large (1.0)	
Group	M	SD	M	SD	M	SD	M	SD
All	.649	.154	.673	.159	.679	.141	.681	.139
Finishers	.651	.156	.691	.161	**.719**	.144	**.722**	.138
Quitters	.621	.146	.599	.149	.568	.131	.569	.142

In Figure 6, we look at the quitters and finishers for the control group, medium-sized incentive, and large incentive. We see that the medium-sized and large incentives have the best discriminatory power between finishers and quitters, with the large incentive treatment being slightly better.

The retention rate varied between more these treatment groups than any other, with the control, small incentive, medium incentive, and large incentive groups retaining 78%, 80%, 73% and 68% respectively. The gain in mean accuracy for the finishers as compared with all participants was 0.8%, 1.8%, 4.0% and 4.1% for these groups, respectively. Thus, the treatment with the largest discrimination power is the large incentive treatment, with the medium incentive treatment very close behind. The medium incentive, with nearly the same metrics as the large incentive, but it provides a higher retention rate and a lower cost; therefore it slightly edges out the larger incentive as the best treatment.

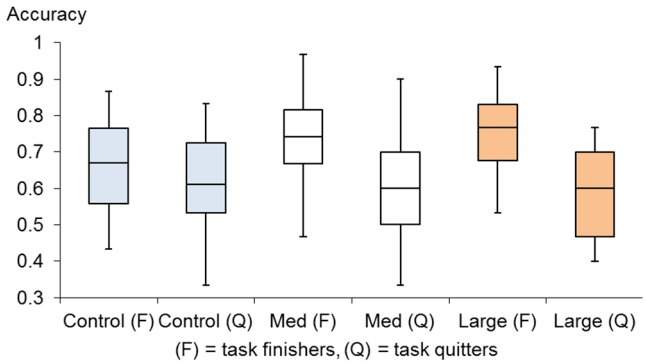

Figure 6. Box-plot comparison of mean accuracy between our control group (shaded light blue, on left), those offered a medium-sized incentive (unshaded, center), and those offered a large incentive (shaded orange, right).

Experiment 5 – Putting it All Together

Next, we compare the performance of our best model, which uses two pay-to-quit incentive prompts after the 10th and 20th questions, provides accurate feedback on performance during the prompt, and provides a decreasing incentive with the magnitude constant, c, at the moderate value of 0.75. For this experiment, we doubled the number of participants to 240. Participants were randomly assigned to the control group (N=120, Female = 46%) or the best strategy group (N=120, Female = 45%).

Mean accuracy for the control group and our treatment group is reported in Table 6. A bolded mean value indicates the difference from the control group is statistically significant at $p<0.05$.

A two-tailed independent samples t-test was used to evaluate the effects of the number of prompts on mean accuracy. There was a significant difference between our control group at the $p<0.05$ level for the two-prompt condition, $t(238) = 2.773$, $p = 0.006$, for finishers, $t(196) = 3.791$, $p = 0.002$. Cohen's *d* was 0.41 and 0.51 for these two values, respectively, which indicates a moderate effect size. None of the other treatment groups provided a significant difference from the control group for mean accuracy.

Table 6. Mean accuracy and standard deviations comparing a control group (N=120) and a treatment group (N=120) comprised of the best strategies from Experiments 1 through 4

	Control		Single Best Strategy	
Group	M	SD	M	SD
All	.637	.146	**.687**	.133
Finishers	.642	.145	**.717**	.133
Quitters	.618	.151	.578	.131

The retention rates for our control group and our best model were 78% and 73%, respectively. This indicates that more than a quarter of our participants (27%) chose to take the pay-to-quit incentive for the decreasing incentive, which we believe demonstrates its effectiveness.

The mean accuracy gain for the finishers as compared with all participants was 0.5% and 3.0% for the control group and the treatment group, respectively. Comparing the finishers of our *best* model compared with all users in the control group, a baseline representing a typical crowdsourcing study design is 8.0%, or a 28.3% gain in mean accuracy. Since it is highly unusual for crowdsourcing tasks to offer any incentives to quit early, we feel that comparing the finishers with our best treatments applied with all participants in the baseline represents an appropriate comparison.

Figure 7 shows box-plots for All (A), task finishers (F) and task quitters (Q) for the "Control" baseline group (shown in light blue shading) and the combination of treatments shown to provide the "Best" performance in Experiments 1-4 (shown in white). The distinction between the finishers and quitters is greatest in the "Best" model – this discrimination allows us to improve our score by comparing the "Best (F)" model with the "Control (A)" model resulting in the 28.3% reported earlier.

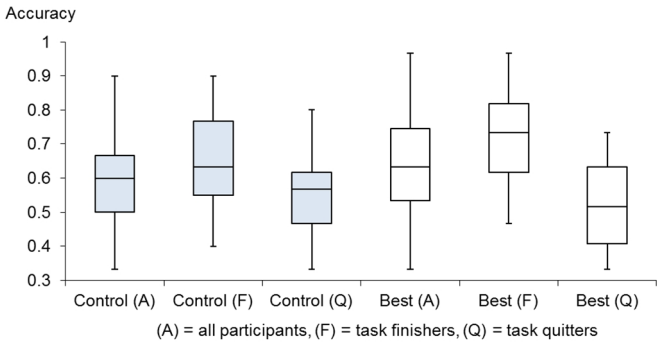

Figure 7. Box-plot comparison of mean accuracy between our baseline (Control) and the combination of treatments from Experiments 1 to 4 that provided the best performance (Best).

We also captured the time taken to answer each question. Based on the *sunk cost fallacy*, more time taken per question may suggest a greater investment in the process and therefore a greater commitment to the task. Conversely, just as with those employed by companies such as Amazon.com or Zappos.com, participants who rushed through the game may be enticed by some easy money to quit early. Thus, we wanted to see if participants who took more time per question were more likely to stay in the game. There was a significant positive correlation between average time taken per question and retention, $r(1200) = +0.489$, $p < .001$. This correlation indicates that participants who took more time to answer each question, on average, were more likely to refuse the offer to quit the game when provided with the pay-to-quit offer.

The Role of Demographics

We examined the demographic information reported in the 1200 pre-task and 1153 post-task surveys to see if any incentive-related patterns existed. We found that females, who comprised 46% of our study participants overall, were significantly more risk averse, accepting the pay-to-quit incentives in far greater frequency than men, despite providing nearly the same overall mean accuracy (two-tailed t-test, $p<0.001$). This may indicate that females are more sensitive to opportunity costs, or it may also indicate greater risk-aversion, as a number of other studies indicate (e.g., [6, 19, 32]). Likewise, Indian and "other South Asian" residents (together comprising 36% of our participants) were more likely to accept the pay-to-quit incentive than residents of other countries or regions (two-tailed t-test, $p = 0.012$), which implies the same aversion to risk [2] or awareness of greater opportunity cost of completing the entire task.

Workers who played games more than 5 hours per week (19% of participants), were significantly less likely to take the reserve bonus than those reporting fewer hours of gaming, even when they were performing poorly (two-tailed t-test, $p<0.001$). This may indicate a greater focus on the game's entertainment value than on the bonus, or it may be due to issues such as cognitive bias – experienced by people who begin to lose at a slot machine gambling and make an irrational attempt to win back their initial investment [12, 37]. Participants who relied on crowdsourcing as their primary source of income were less likely to take the reserve bonus than those who did not (two-tailed t-test, $p = 0.024$). Interestingly, workers in the highest reported income band (exceeding $40K per year) were more likely to take the reserve bonus than those whose reported income was in the lowest band (less than $10K year) (two-tailed t-test, $p = 0.017$). While we expect people to be risk-neutral when relatively small amounts of money are involved, this difference may be due to factors such as greater opportunity cost, competitiveness, or the entertainment value of the game itself.

ANALYSIS

Many of the experiments conducted here examine a few features of a single game for a small amount of compensation. Therefore, it would seem that the implications for pay-to-quit incentives are limited. However, as Experiment 5 illustrates, the value of proper design of incentives can encourage the high-performers to complete the task while encouraging the poor-performing participants to accept an inducement to quit early and is backed up by other research, e.g. [13]. We believe that applying these techniques on a larger scale, with more at risk for participants, is likely to magnify the effects of these treatments.

In Experiment 1, we note that two prompts appear to be the optimal number to offer for our thirty-question task. In Experiment 2, we identify that feedback at the time of a pay-to-quit incentive is important to the decision maker – and the relative performance matters as much as the absolute performance in the decision to accept the pay-to-quit incentive. In the short-term, we note that telling participants they are underperforming (e.g, performing

worse than their peers) will help to discriminate the better performing workers from the poorly-performing ones. When a worker is aware they are underperforming, it normally can be rationally explained; when others mention it, it impacts the decision to continue with work. In the long-term, however, people are likely to be aware of the performance of others and the underperformance message is unlikely to carry as much weight. In an anonymous work environment, such as an online labor market, this façade is more difficult for workers to verify.

Experiment 3 illustrates the importance of designing the incentives so poorly-performing workers have a better incentive to leave. We note, however, that the decreasing incentive model that performed well in this study could be reverse engineered by workers over time, and a new model would need to be introduced. However, the underlying idea of providing enough compensation for the poorer performers to leave remains the same.

Many of the incentives we offered, including the large incentive in Experiment 4, produced a range of possible pay-to-quit incentives between $0.20 and $0.30 at the first prompt, and between $0.10 and $0.30 at the second prompt. The poorest-performing participants were offered larger incentives. Since all participants are aware the entire task is worth at most $0.30, and a poorly-performing participant who has played two-thirds of the game has a good idea of the maximum compensation they could receive at the game's conclusion, why wouldn't they take an incentive that provides a greater income than continuing to play?

Based on the feedback provided in the 1,153 completed post-hoc surveys, a few explanations come to mind. First, cognitive overload, often a problem in game design, could obfuscate the decision to be made at the prompt during a game designed to induce a stressful environment. Second, the participant may be less interested in the compensation than participating in the task. Third, because of the perception of a limited downside to the task, they could continue to play without giving the pay-to-quit incentive serious consideration. Fourth, it could be due to self-selection; the people who would benefit from the incentive are risk-seekers and want to continue to take a risk instead of take the "safe" incentive.

Each of these has important considerations for pay-to-quit inducements in online labor markets. First, because tasks in online labor markets are often simply designed and are known to serve as experimental test-beds for deception studies and other information gathering studies, participants may not consider an incentive to quit to be valid. They may rush through the prompt without giving it careful thought. This is backed up by studies on crowdsourcing that indicates even the most basic quality checks are frequently overlooked by workers [7]. Second, when tasks are designed as a game and advertised in online labor markets, the task requesters are looking for people to participate more than once, since more work is accomplished at no additional cost. The value of actual compensation offered then becomes secondary to task enjoyment [8]. This draws a certain type of worker, which is not representative of crowdworkers overall. It also has parallels to job-related employment in the real world.

Third, microtasks, through their anonymity, are often thought of as "disposable work", with little expectation for worker and requester to continue to work together in the future. Last, the type of people who are looking for work on MTurk may be more risk-seeking than the population as a whole [4, 30], affecting their decision to take an incentive to quit a task early. This type of risk-seeking behavior as a response to a run of bad luck has been studied in stock markets [25, 29] and in game shows [1], where frequently people make decisions that may seem irrational to others.

One may wonder if the better-performers are discouraged by being excluded from receiving pay-to-quit rewards. We believe this is not the case; the incentives offered by these firms as attempts to signal to customers, employees, and investors that top-notch customer service and dedicated employees are an essential part of their enterprise that is worth a sum of money to protect. This increases the likelihood that the better performers will see themselves as part of an exclusive group, which in turn will likely encourage them to maintain high performance. Similarly, the few hiring mistakes these companies make are corrected as early as possible, limiting any potential long-term negative effects. On a smaller scale, these same qualities are valuable to task requesters in online labor markets as well for many of the same reasons.

Implications for Collaborative Work

Incentive schemes, when properly designed and applied, can provide an important role in aligning the goals of workers and requesters alike. However, in our study, we demonstrate that not all incentive schemes can improve performance as expected. Most notably, a pay-to-quit incentive does not always discriminate between those that rationally should quit early and those that rationally should continue with the task. However, by taking steps to identify the features most relevant to the online labor market pool, we can discriminate between the high-performing and poor performing workers as companies like Zappos.com, Amazon.com and L18 have attempted to do.

Some workers continued with the task even though the most rational option would have been to quit and accept the incentive. Their reluctance to quit may be due to the "pseudo-certainty" effect, where people overestimate the likelihood of an uncertain event (in our case, obtaining more compensation at the game's conclusion).

In more diverse collaborative groups, an understanding of the group's collective view on risk may provide information on the most effective incentive to offer workers. Even in simple tasks with little at stake, such as with multiple-choice questions with the possible reward of $0.30, workers

tend to demonstrate either risk-seeking or risk-averse behavior. We believe that future designers of crowdsourcing tasks will need to incorporate protocols that will be more complex, involve a more diverse set of workers with different attitudes and motivations towards work, and encompass incentives that are far more complex than are currently offered.

CONCLUSION

Companies known for hiring the best people, such as Zappos.com, have implemented pay-to-quit incentives to best align work attitudes towards corporate objectives. These companies have found that these incentives help enhance the corporate culture and increase productivity. In this study, we have observed the effects of different types of incentives on worker mean task accuracy through a game.

Through a series of five experiments, we examined different aspects of pay-to-quit incentive schemes in online labor markets. The first evaluated the frequency workers should be prompted with the pay-to-quit incentive and found that making offers every 10 questions (approximately every 5 minutes) appears to be correct in a 30-question task. The second experiment examined the amount of feedback to provide, and found that providing feedback at the time of the incentive offer worked best. This experiment also demonstrated that feedback on relative performance did matter, with poorly-performing workers who were told that they were underperforming accepted the incentive more often than groups that were provided accurate information or told that they were outperforming other workers.

The third experiment looked at incentive types based on performance, one providing incentives that increased based on performance and one that offered incentives that decreased based on performance. The latter provided the best incentives for the poorly performing workers to leave. Our fourth experiment looked at the size of the incentive, finding that the size of the incentive matters, but there is little difference between a moderately-sized incentive and a larger one. In our fifth experiment, we sought to validate our findings. We took the best strategies uncovered from each of the four earlier experiments and ran this experiment with a control group. The workers we retained provided an increase in mean task accuracy of 28.3% over the mean accuracy of all workers in our control group. We believe this represents a true measure of performance improvement over a task without the pay-to-quit incentives.

Given the increasingly multi-national scope of collaborative work, an understanding the worker pool demographics is essential. Examining our worker demographics, we observe that females, South Asian residents and people earning higher wages are more likely than other groups to accept the pay-to-quit incentive when offered, possibly indicating greater risk aversion or greater opportunity cost for their time. People who play games regularly, who depend on income from crowdsourcing, and workers in the lowest income band are less likely to accept the reserve bonus. Our results indicate that certain demographics may be more risk averse than other groups, and this can be explored through additional analysis.

One limitation to this study is we have only examined financial incentives. A number of studies, including the aforementioned one by Locke [21] have shown that financial incentives are less important to workers than other types of incentives, such as recognition of effort or achievement. This is particularly true in collaborative work environments. Financial incentives, coupled with other incentive types, are likely to further improve performance in collaborative environments.

To this end, we believe that this pay-to-quit incentive model can also be applied to non-financially rewarding crowdsourcing works, such as games. A participant in a war game, for example, may choose to quit at certain milestones in the game and keep their earned rewards if they get the feeling that a particular mission will not go well (or is not in their best interest). We also believe this model can apply even better to participant models that have different roles or levels, rather than a single level as we have explored in this study. Not only could we offer financial compensation as an incentive, but also the promotion/demotion between levels. For many, earning titles in Wikipedia, for example, such as steward, sysadmin or ombudsman, are coveted by those that devote considerable time in making Wikipedia edits; keeping (or promoting to) those titles incentivizes them to be aligned with Wikipedia's objectives, which parallels the incentives faced by employees at Zappos.com, Amazon.com and elsewhere.

In future work, we plan to continue our examination of incentives including an evaluation of non-financial incentives. Games provide numerous advantages over static tasks for applying treatments. We plan to examine this area in greater detail and provide participants with more complex tradeoffs in order to examine the decisions they make.

REFERENCES

1. Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008). *Risk aversion in game shows* (Vol. 12, pp. 359-404). Emerald Group Publishing Limited.
2. Binswanger, H. Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, 62:3 1980, 395-407
3. Cavuto, N. (2014, June 24) *Pay to Go Away: Bribing Some Workers to Just…Disappear*. Retrieved from http://www.foxbusiness.com/economy-policy/ 2014/06/ 24/pay-to-go-away-bribing-some-workers-to-just-disappear/
4. Charness, G., Gneezy, U., & Kuhn, M. A. (2013). Experimental methods: Extra-laboratory experiments-

extending the reach of experimental economics. *Journal of Economic Behavior & Organization*, *91*, 93-100.

5. Csikszentmihalyi, M. Flow: The psychology of optimal experience. Harper Perennial, 1991.
6. Eckel, C. C. and Grossman, J. Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, 1, 2008, 1061-1073.
7. Eickhoff, C., & de Vries, A. (2011). How crowdsourcable is your task. *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM),* pp. 11-14.
8. Eickhoff, C., Harris, C. G., de Vries, A.P. and Srinivasan, P. Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments. *Proc. of SIGIR'12*, ACM, New York
9. Gal, D. A psychological law of inertia and the illusion of loss aversion. *Judgment and Decision Making*, 1, 1 2006. 23-32.
10. Gill, D. and Prowse, V. A structural analysis of disappointment aversion in a real effort competition. *The American economic review*, 102:1 2012, 469-503.
11. Gneezy, U. and Rustichini, A. Pay enough or don't pay at all. *Quarterly Jour. of Econ.,* 115:3 2000, 791-810.
12. Griffiths, M. D. (1994). The role of cognitive bias and skill in fruit machine gambling. *British Journal of Psychology*, *85*(3), 351-369.
13. Harinck, F., Van Dijk, E., Van Beest, I. and Mersmann, P. When gains loom larger than losses: reversed loss aversion for small amounts of money. *Psychological science,* 18:12, 2007, 1099-1113.
14. Harris, C. G. You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. *Proc. WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (Hong Kong, China)* 2011.
15. Hartley, R., Lanot, G. and Walker, I. Who really wants to be a millionaire? Estimates of risk aversion from game show data. 2006. Working paper.
16. Heyman, J. and Ariely, D. Effort for payment a tale of two markets. *Psych Science*, 15, 11. 2004, 787-793.
17. Holt, C. A. and Laury, S. K. Risk aversion and incentive effects. *Amer. economic review*, 92:5. 2002, 1644-1655.
18. Jeffries, A. *Lot18 Is Straight-Up Paying Unhappy Employees to Quit*. https://betabeat.com/2012/05/lot18-founder-offered-to-pay-unhappy-employees-to-quit-today-and-six-of-them-did/
19. Jianakoplos, N. A. and Bernasek, A. Are Women More Risk Averse? *Economic Inquiry*, 36:4 1998, 620-630.
20. Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 1979, 263-291.
21. Kopelman, R. E., Chiou, A. Y., Lipani, L. J., & Zhu, Z. (2012). Interpreting the success of Zappos. com, Four Seasons, and Nordstrom: Customer centricity is but one-third of the job. *Global Business and Organizational Excellence*, *31*(5), 20-35.
22. Laffont, J. J. The theory of incentives: the principal-agent model. *Princeton Univ Pr,* 2002.
23. Lam, S. K., Pennock, D. M., Cosley, D. and Lawrence, S. 1 Billion Pages= 1 Million Dollars? mining the web to play" who wants to be a millionaire?". *Morgan Kaufmann Publishers Inc*., New York, 2002.
24. Locke, E. A. Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance*, 3, 2 1968, 157-189.
25. March, J. and Shapira, Z. Managerial perspectives on risk and risk taking. *Mgmt science.* 1987, 1404-1418.
26. Martin, D. *Amazon Pays Employees $5,000 to Quit Their Jobs.* http://www.cbsnews.com/news/amazon-pays-employees-5000-to-quit/, 2014.
27. Mason, W. and Watts, D. J. Financial incentives and the performance of crowds. ACM. New York, USA, 2010.
28. McFarland, K. (2013). Why Zappos Offers New Hires $2,000 to Quit. *http://www.businessweek.com/stories/2008-09-16/why-zapposoffers-new-hires-2-000-to-quit*
29. Miller, E. M. Risk, uncertainty, and divergence of opinion. *Journal of finance*. 1977, 1151-1168.
30. Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411-419.
31. Prendergast, C. The provision of incentives in firms. *Journal of economic literature*, 37, 1 1999, 7-63.
32. Schubert, R., Brown, M., Gysler, M. and Brachinger, H. W. Financial decision-making: are women really more risk-averse? *The American economic review*, 89:2 1999), 381-385
33. Shaw, A. D., Horton, J. J. and Chen, D. L. Designing incentives for inexpert human raters. *Proc. of the ACM 2011 conference on Computer supported cooperative work (Hangzhou, China*). ACM, New York, NY. 2011.
34. van Poucke, D. and Buelens, M. Predicting the outcome of a two-party price negotiation: Contribution of reservation price, aspiration price and opening offer. *Journal of Economic Psychology*, 23, 1 2002, 67-76.
35. von Ahn, L. and Dabbish, L. Labeling images with a computer game. ACM, New York, 2004.
36. von Ahn, L., Kedia, M. and Blum, M.Verbosity: a game for collecting common-sense facts. ACM New York 2006.
37. Walker, M. B. (1992). Irrational thinking among slot machine players.*Journal of Gambling studies*, *8*(3), 245-261