Into the Peanuts Zone: Leveraging Hyperbolic Time Discounting to Motivate Transient Crowdworkers to Return

Peter Kinnaird

ABSTRACT

The boom of studies on crowdsourcing platforms like MTurk is in full swing, yet some types of studies are hampered by the challenge of motivating individual workers to return for follow-up work later. Purely increasing incentives has been shown to provide mixed results. Results and techniques from behavioral economics suggest a way forward. I describe the results of an experiment aimed at testing two prominent results from behavioral economics, hyperbolic time discounting and framing effects. Further, I tested incentive compatible price elicitation methods against fixed incentives in hopes of increasing cost efficiency. We contribute an early look at 3 methods for establishing return rates along two dimensions: cost and rate of return.

Author Keywords

Crowdsourcing, Mechanical Turk, Incentives, Behavioral Economics, Cognitive Science, Framing effects, Time Discounting

ACM Classification Keywords

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

General Terms

Human Factors; Design; Measurement.

INTRODUCTION

The boom of psych studies on crowdsourcing platforms like MTurk is in full swing [5,6,7], yet some types of studies are hampered by the challenge of motivating individual workers to return for follow-up work later. Behavioral economists have extensively demonstrated that people typically undervalue future earnings according to a hyperbolic discount function for ordinary or large sums of money[3,4,10,11]. Behavior and utility estimates remain undetermined for very small amounts of money since those amounts are often deemed trivial by participants in experimental environments. I designed and conducted an experiment to estimate the parameters of the time-discount function that Turkers apply to revisiting a task at a later time which can be used to adequately incentivize workers to return.

Microtask marketplaces like MTurk can provide workers with a marketplace filled with uncertainty and transient jobs for pay. An abundance of comparable tasks reduces the likelihood that a worker will seek out subsequent jobs from a specific employer since there is little incentive to do so. Technical capabilities exist within MTurk to email prior workers to ask them to return for a new job. Unfortunately, even if they do return for a new job, the manager has little control over when they will return.

Behavioral economics experiments typically involve thought experiments by participants or small sums of money/goods. They are often asked questions like, “Would you prefer $50 now, or $100 in 4 weeks?” Alternatively, participants may be offered a coffee mug, a nice pen, or a free lunch. Through numerous studies [3,13], behavioral economists have consistently demonstrated that people typically discount psychologically distant rewards according to a hyperbolic function. That is, the further in time that a reward will be realized, the less value people assign that reward. The value decreases exponentially (hyperbolically) according to its distance. One weakness in many of these experiments, however, is that they lose consistency when the reward amounts become too large or too small. Asking someone whether they prefer $15 million now or $30 million dollars at the end of the month can provide unexpected results. Likewise, when people are asked to work with amounts of money they deem extremely small, they often respond inconsistently. The small end of the scale is sometimes called the “peanuts zone” because people say, “I don’t know, that’s just peanuts.”

Given that nearly all transactions in MTurk and similar environments are within this area, it’s unclear whether Turkers will behave according to ordinary models. The explanation for the peanuts effect is that the monetary amounts in question are trivial – yet Turkers consistently perform work (albeit of varying quality) for these trivial amounts.

I conducted an experiment that employed techniques and models from behavioral economics in an attempt to motivate workers to return for a follow up task later. The proposed experiment contributes an estimation of parameters for Mechanical Turk workers in theoretical hyperbolic time discounting functions and implications for the design of crowdsourced tasks.

BACKGROUND

Bernstein et al. successfully employed a retainer model that enabled them to keep workers on standby until a specific moment when they would be asked to work. The longest the experimenters kept a worker on retainer was 30 minutes [2]. This solution could be framed in terms of getting workers who complete a task at time t0 to return at t1 for a new task. The design implemented for the study (using a javascript alert to draw attention to the task window), however, is less realistic for scenarios in which the t1 is much further in the future than t0 like, for example, 2 days since workers might not even be at the computer, or the computer might have been restarted. Other alternatives have been tried, but these are all aimed at solving the problem of high latency in work response time rather than improving retention rates for returning workers.

Researchers have shown that paying more for a task elicits greater volunteerism but no increase in quality [9]. Similarly, Shaw, Horten, and Chen tested a variety of incentive structures designed to increase quality with few useful results [12]. Neither of these studies examined the question of motivating workers to return to a task later.

Behavioral economists have long known that people respond to time variant decisions according a hyperbolic function [3]. While the general curvature is hyperbolic, the precise parameters are unknown and likely subject to a number of contextual factors such as framing [8]. In general, this means that people prefer a smaller amount of money now to a larger amount of money later. The further out this goes, the larger the amount has to be to dissuade people from taking the smaller amount now. These results suggest that motivating people to perform a task in the future will require larger amounts of money for more distant events and smaller amounts of money for tasks couched in terms of losses rather than gains (since people are risk averse) [13].

METHODS

I conducted a between subjects experiment on Mechanical Turk limiting workers to the United States. I recruited workers for a brief task in exchange for $0.15 and randomly assigned them to condition. I wanted the task itself to be a believable one for which the requester might actually want the Turker to return at a later time. As such, I designed a brief questionnaire about grocery shopping habits which quickly hones in on questions about potato chips and Doritos. Turkers then advance to a silly Doritos ad from the Superbowl 2010 and answer some questions about it. Only after they advance to the third page are they informed that we will ask them to join a followup HIT. The idea I'm trying to implicitly convey is that we will want to see if the ad convinced them to buy Doritos. If they don't understand that, however, no confounds are introduced. At this point in the task Turker's experiences will differ based on which condition they were randomly assigned to.

The experimental design was a 4 x 3 x 2. I tested four bonus conditions: no bonus offered; a fixed bonus offered ($0.20); and two incentive compatible bonus elicitation methods: a common bid, and a BDM[1] bid. Incentive compatible methods, in general, are methods for eliciting fair prices from participants. If I were to ask a participant to simply name a price for which they would return, they might ask for an unrealistic amount ($1 million). Conversely, if I informed participants that too high of a bid would result in a severe punishment, I would depress prices to an unrealistic amount ($0). The text of the BDM method that I used is as follows:

You might be wondering how much of a bonus you've won. We have a bonus amount in mind that we believe is fair compensation for the followup HIT. Enter a bid that will motivate you to return for the followup HIT. If the amount you suggest is higher than the amount we have in mind, you will no longer be eligible for the followup HIT and won't win the bonus. If the amount you suggest is less than or equal to the amount we have in mind, you'll be eligible. We'll tell you if you're eligible right away. It's in your best interest to tell us the smallest amount that will motivate you to return.

The common bid text is as follows:

You might be wondering how much of a bonus you've won. We're not sure how much of a bonus to offer for the followup HIT. Enter a bid that will motivate you to return for the followup HIT. We'll invite Turkers back for the followup HIT starting with the lowest bids. We'll tell you if you're eligible right away. It's in your best interest to tell us the smallest amount that will motivate you to return.

Although these are both complex constructions, I would suggest that they remain simpler than Shaw, Horten, and Chen's Bayesian Truth Serum formulation [12]. These bid amounts are a primary dependent variable for analysis.

Workers then have the opportunity to request up to four different kinds of assistance which I use as a surrogate for likelihood to return. They can ask for help making a bookmark to the followup HIT or adding a reminder to their electronic calendar. We can also send them an email immediately with all of the information about the followup HIT or we can send them an email some amount of time (user specified) prior to their window of eligibility.

Workers were asked about one of three time frames: 15 minutes later with a 2 minute eligibility window, 1 day later with a 30 minute eligibility window, or 3 days later with a 1 hour 30 minute eligibility window. The windows grow with the time because a substantial subset of return tasks might be more concerned with getting a returning worker at about the right time than at an exact time. This fuzziness is likely to increase with the delay since the cost of losing a returning worker (to task completion) increases with the delay.

Finally, workers were shown one of two framings: gains or losses. The text I used for gains was:

Congratulations! You have the opportunity to win a bonus!

The text I used for losses was:

Congratulations! You've won a bonus, but you might lose it!

Table 1. Participants (N) per condition (framing italicized).

	No Bonus	Fixed Bonus	Common Bid	BDM Bid
15 min.	N = 24	*Gains*: N = 23 *Losses*: N = 23	*Gains*: N = 24 *Losses*: N = 23	*Gains*: N = 24 *Losses*: N = 23
1 day	N = 24	*Gains*: N = 23 *Losses*: N = 23	*Gains*: N = 23 *Losses*: N = 23	*Gains*: N = 23 *Losses*: N = 23
3 days	N = 24	*Gains*: N = 23 *Losses*: N = 23	*Gains*: N = 24 *Losses*: N = 23	*Gains*: N = 23 *Losses*: N = 23

No follow up task was actually provided to any workers. When they reached the point in the experiment at which they would be provided the link, workers are informed that our budget has (sadly) run out and we are unable to offer the follow up HIT and bonus.

Primary dependent variables are the bid requests and the number of requested reminders (as a surrogate for motivation to return). Hypotheses are as follows.

H1: Participants will require larger rewards for more distant follow up hits and request more reminders.

H2: Participants who are oriented in a loss oriented frame will request more reminders than those in a gains oriented frame.

H3: Workers in incentive-compatible bid elicitation conditions will request more reminders than those in fixed and no-bonus conditions.

RESULTS

418 Turkers completed the HIT and are included in the analysis. These were divided among conditions as shown in table 1.

The distribution of bonus bid amounts requested conformed approximately to a power law so I transformed this data by taking the log. The logged bid amounts are very close to a normal distribution.

I evaluated H1 with an overall ANOVA and three planned contrasts. The overall ANOVA results along bid amount are insignificant with F(2,2) = 2.37; $p < .09$. The planned contrasts show that participants request a significantly higher bid in the 3 days condition in comparison to 15 minutes, but the 1 day condition is not significantly different from 15 minutes or from 3 days. The overall ANOVA for total reminders requested is significant with F(2,2) = 4.31; $p < .01$. The planned contrasts reveal that significantly fewer reminders are requested for 15 minutes than 1 or 3 days. There is no significant difference between the number of reminders requested for 1 or 3 days. Therefore, H1 is partially supported.

I evaluated H2 with an overall ANOVA and found no significant difference in the number of reminders requested based on framing.

I also evaluated H3 with an overall ANOVA and found no significant differences in the number of reminders requested based on bid elicitation method.

Post-hoc analysis showed no significant differences in elicited bids between common bid and BDM bid.

Two questions inserted into the grocery shopping survey explored participants' risk aversion. These items were evaluated on a 5 point labeled Likert scale {Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, Strongly Disagree}. The questions were:

1. In general, I'll go out on a limb and take risks.
2. I usually like to play it safe.

Inverting either question results in a correlation of .60 suggesting the creation of a scale. No interaction was observed between this risk aversion scale and either the total number of reminders requested ($p < .28$) or the bid elicited ($p < .18$).

DISCUSSION

H1 received partial support and was observed to trend towards full support. The results demonstrate that distant returns do require a larger bonus than immediate returns. I would speculate that Turkers may perceive return events in two bins rather than linearly. That is, there is some relatively fixed cost required for an immediate (or near immediate) return event. The proximate return cost is likely estimable based on Turkers' perceptions of the worthwhileness of the first HIT in comparison with other HITs available at that time. The more distant cost may also be relatively fixed and increase because of the additional effort required to return at a specific time. These results suggest that testing a greater number of periods may enable modeling in this "peanuts zone." Though the absence of an interaction between risk aversion and framing may be due to the peanuts effect.

Framing effects may have had a smaller effect in this context since I am technically incapable of actually giving the workers a bonus and then revoking it. Instead, I relied on the substantially weaker effect of simply telling the worker that they had already won the bonus, or they are eligible to win the bonus.

The effects of the manipulations on the total reminders requested is somewhat disappointing but ultimately may not be an indicator of actual likelihood to return. Future work will explore this possibility.

FUTURE WORK

Given that there were no significant differences for framing effects or bid elicitation method, I will drop those from future work. I plan to rerun the experiment with alternative conditions. I will test a greater number of time periods in 3 bins: less than a day, 1 day to 1 week, and longer than 1 week. I will also incorporate a control in which I simply surprise the Turker with an email asking them to do a follow up HIT. I will also increase the sample size so I can increase the power of the findings. Finally, I will construct a follow up HIT so I can measure real return rate as a primary dependent variable.

ACKNOWLEDGMENTS

Thanks to Paul Andre, Niki Kittur, George Loewenstein, and Sara Kiesler for feedback on the ideas and helpful discussions and support.

REFERENCES

1. Becker, G.M., DeGroot, M.H., and Marschak, J. Measuring Utility by a Single-Response Sequential Method. *Behavioral Science 9*, 3 (1964), 226-232.

2. Bernstein, M.S., Brandt, J., Miller, R.C., and Karger, D.R. Crowds in two seconds. *Proc UIST 11*, ACM Press (2011).

3. Frederick, S., Loewenstein, G., and O'donoghue, T. Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature 40*, 2 (2002), 351-401.

4. Frederick, S. and Loewenstein, G.F. Time Discounting : A Critical Review. *Ariel 40*, 2 (2001), 1-83.

5. Heer, J. and Bostock, M. Crowdsourcing Graphical Perception : Using Mechanical Turk to Assess Visualization Design. *Area 152*, 4 (2010), 203-212.

6. Ipeirotis, P. Crowdsourcing using Mechanical Turk: Quality Management and Scalability. *Crowdsourcing for Search and Data Mining CSDM 2011*, (2011), 5.

7. Jakobsson, M. Experimenting on Mechanical Turk: 5 How Tos. *ITWorld September 3*, 2009, 2009. http://www.itworld.com/internet/76659/experimenting-mechanical-turk-5-how-tos.

8. Kahneman, D., Knetsch, J.L., and Thaler, R.H. Anomalies The Endowment Effect , Loss Aversion , and Status Quo Bias. *Journal of Economic Perspectives 5*, 1 (1991), 193-206.

9. Mason, W. and Watts, D.J. Financial incentives and the "performance of crowds."*ACM SIGKDD Explorations Newsletter 11*, 2 (2010), 100.

10. McClure, S.M., Ericson, K.M., Laibson, D.I., Loewenstein, G., and Cohen, J.D. Time discounting for primary rewards. *Journal of Neuroscience 27*, 21 (2007), 5796-5804.

11. Reuben, E., Sapienza, P., and Zingales, L. Time discounting for primary and monetary rewards. *Economics Letters 106*, 2 (2010), 125-127.

12. Shaw, A.D., Horton, J.J., and Chen, D.L. Designing Incentives for Inexpert Human Raters. *CSCW*, (2011), 1-8.

13. Tversky, A. and Kahneman, D. Advances in prospect theory. *Journal of Risk and Uncertainty*, (1992).