Employer Reputations and Effort Reciprocity

Lee Huang, Jonathan Jin, Anthony Pan, Geoffrey Wang, Jaewon Yoon

06/01/2015

Abstract

Constructing effective incentive schemes to generate worker effort involves many confounding factors. Incentive schemes involving gift-exchanges should be effective in theory and has been proven in several field and lab experiments. However, the specific motivations behind such gift-reciprocity remains ambiguous. Several theories predict different motivating factors in worker effort, including fair wage-effort, employer regard, and reputation. We outline relevant existing literature and potential empirical evidence and design a field experiment around Amazon's Mechanical Turk to investigate motivating factors behind gift-reciprocity in labor markets–specifically, the extent to which worker reciprocity comes from concerns for worker reputation as opposed to regard for the employer.

Contents

List of Figures

List of Tables

1 Motivation

1.1 Literature

List (2005) claimed that gift-reciprocity in the field is caused by reputational concern rather than reciprocal regard for the buyer. Specifically, in his field experiment examining whether sports card/ticket-dealer matched quality of the traded good accordingly to the price offered by the customers, List found that sellers reciprocated to the value of payment given by buyer only when the quality of the traded good impacted the seller's reputation [22]. That is, the quality was verifiable by the customers (third-party quality check was available) and the seller's reputation would be affected accordingly as the seller expects repeated interaction with the customer (local dealer).

List's data "[does] not necessarily preclude social preferences from having import within other economic domains." In particular, it would be interesting to examine whether it is social preferences or self-interest that guides gift-exchange in the labor market. Employee reciprocity in the labor market has been a longtime conundrum to economists. In the labor market, employers pay higher wages than the market-clearing wage in the hopes of motivating workers to exert more effort. From papers by Akerlof and Fehr, we see that contrary to neoclassical theory – which assumes rationality of economic agents – would predict, workers do exert more effort in response to higher wage [1, 11].

Previous literature addressed the issue by examining whether it is our inherent preference for fairness or our impulse to reciprocate to another agent's kindness that drives the gift-effect [10]. However, based on List (2005) we propose that it is a self-oriented reputation concern that guide's employee's gift reciprocation in the labor market. In other words, we believe that the labor market is also a domain where agents are driven primarily by self-interest rather than social preferences. If this assumption holds to be true, this would imply that not all workers would reciprocate to higher wages with increased effort. For instance, it would not be a good attempt to motivate one-time contract workers with an unexpected increase in wage, if these workers do not expect their interaction with the specific employer to affect his/her future reputation.

1.2 Theoretical Models

In a meta-analysis of gift-exchange experiments, Charness and Haruvy documented various models explaining the gift-exchange based on altruism, inequity aversion, or reciprocity [8]. The following models adapted from their paper, which incorporate reciprocity and reputation concern in the agents' utility function respectively, are particularly relevant for our study.

The utility functions in the table include $R(g)$ and $R(g, r)$. Both are coefficients that indicates the regard the worker has for the employer. In the model of social-preference theory, the regard is only affected by the size of the gift(g). However, in the model of reputation-concern theory the regard is determined by both the size of the gift(g) and the degree to which the worker's reciprocation impacts the worker's reputation(r).

Table 1: The study's theoretical models

		Social Preference / Reciprocity	Self-interest / Reputation
Employee's concern		- Buyer/firm's welfare	- Own reputation
Determining factors		- How nicely sellers/workers have been treated	- How nicely sellers/workers have been treated - Possibility of quality verification by buyer/employer
Model	Utility	U(e) = w + g - C(e) + R(g)(ve - g - w)	U(e) = w + g - C(e) + R(g, r)(ve - g - w)
	FOCs	C'(e) = R(g)v	C'(e) = R(g, r)v

Table 2: Explanation of model variables

Parameter	Explanation
w	Lump-sum wage
g	Gift
e	Employee effort quotient
C(e)	Effort-cost function
R(g)	Employee's regard to firm
v	Employer's valuation of employee effort
r	Reputation concern

We are interested in determining whether gift-exchange is driven by the worker's social preference to return the employer's favor or the worker's concern about how his/her respond to gift would impact his/her reputation. In other words, our research would address whether reputation concern, r, is indeed included in the regard coefficient, $R(.)$, by examining whether reputation concern (r) impacts the worker's degree of reciprocation (excess effort or e). In particular, we predict that if $r = 0$, that is reputation is not at stake, $R(.) = 0$.

2 Question

Our research question can be phrased as follows: *is effort reciprocity in labor markets mediated by employees' concern for their own reputation (self-interest) rather than their reciprocal regard for employer (social-preference)?*

3 Experimental Design

We recruit participants through Amazon Mechanical Turk to complete a simple but non-trivial task with easily trackable progress, whose economic value to the firm is nevertheless obvious and non-trivial.

3.1 Amazon Mechanical Turk

Amazon's Mechanical Turk allows us to crowd-source our study participants and achieve a larger subject pool. Additionally, it allows for greater participant transparency since there is a built in reputation mechanic to mitigate non-ideal or abusive behavior in participants. As a result, due to Mechanical Turk's built in properties, there is no need need to physically moderate our study, personally source participants, or physically relocate — all things that increase the overall cost of the experiment. Additionally, using Amazon Mechanical Turk will solve the problem of a "muted" reputation effect from a one-shot experiment since the built-in reputation mechanic will cause participants to be more concerned about their reputation.

The biggest trade-off in our decision to use Mechanical Turk to source study subjects is that the decision renders us unable to evaluate "softer" metrics such as body language, facial expression, and the like. However, given the empirical basis of our experimental design, this trade-off are not significant and pose no threat to the validity of our experiment.

3.2 Task

We ask participants to extract key information from unstructured personal records. Such key information would include, potentially among other things, name, email, phone number, etc.

We frame the task in the context of a higher education institution consolidating its alumni database for the purpose of later requesting alumni gifts and donations. To this effect, our experiment is best performed in partnership/collaboration with a reputable institute of higher education, such as the University of Chicago.

This framing gives our task plausible credibility in Turkers' eyes–they see the task as legitimate and not a "toy" task contrived solely for study purposes. It additionally grants us as experimenters a precedent by which we might valuate each task instance to the Turker. We inform Turkers that, on average, each gift requested alumnus are expected to respond to the request with $\$x$ of donation, where x is a value to be obtained from the partner institution. We inform Turkers thus in order to make the economic value of each task constant across individual task instances.

A sample task instruction is included in Appendix 7.1.

3.3 Treatments

The study will be a 2x2 design utilizing the following two treatment dimensions:

- Whether or not the employee is given a gift; and
- Whether or not the employee's work quality impacts his/her reputation

Therefore, the participants will be separated into one of the four treatment groups as shown in Figure 1.

Figure 1: Treatment Group

	No Gift	Gift
Verification	Effort exerted with no gift with verification	Effort exerted with a gift with verification
Non-Verification	Effort exerted with no gift without verification	Effort exerted with a gift without verification

3.3.1 Gift vs. No-Gift

The Gift group will be manipulated by whether or not the Turkers receive a higher wage than the one announced in the job advertisement. As shown in Appendix 7.2, both Gift and No-Gift groups will be induced to expect to receive a total of $2 for working 10 minutes. However, their actual wage rates revealed just before they begin their task will differ as shown in Appendix 7.1. As researchers have consistently replicated that employees show enhanced effort in response to gift in lab and field [11, 15, 19, 20], we expect a higher degree of employee effort in the Gift condition compared to the No-Gift condition.

3.3.2 Verification vs. Non-verification

Reputational concerns will be manipulated by assuring the Turkers that the quality of their work either will or will not be verified and reflected in their evaluation. As employers on MTurk, we will be able to leave either positive or negative feedback for Turkers, which contributes to their overall reputation as an employee on MTurk. While employee reputation is not public to other users, employees can view their own reputation and employers can put restrictions on their tasks such that only employees with a certain reputation can complete those tasks.

3.4 Parameter of Interest

We are interested in measuring the amount of effort that participants in each treatment group display. In particular, we will be interested in observing the difference in effort that the participants in the Gift treatment group shows compared to the No-Gift treatment group depending on whether or not their reputation is on stake. To assess effort, we assume that the quality and quantity of work done in a given time is a function of the effort put in. The quantity of work will be measured by the number of customer info-sheets processed and quality of work will be measured by the number of fully correct info-sheet. From an experiment run by Kube 2012, quantity and quality increase at the same time as higher reciprocity gifts are introduced [19]. However, in order to make our results more calculable and to reduce confounds involving measuring quantity and quality, we normalized our measurement to a quality ratio, represented by the ratio: $\frac{\text{correct entries}}{\text{total entries}}$. The use of such normalized measurement is adapted from Kube (2012), which also looked at worker reciprocity to gifts. We will examine the effects of our Gift treatment and Verification treatment on workers' quality ratios for our task in order to distinguish the causes of worker effort reciprocity. One potential concern is that workers may exhibit increased effort primarily through output quantity at the expense of output quality. By emphasizing the importance of the quality of work (the expected value of incorrectly entered information is 0), we ensure that increased effort would indeed reveal itself through increased perunit quality.

We would also need to look at a number of other information to ensure that our randomization across treatment groups were valid. In particular, we would want to consider if the baseline reputation of the employees and previous experience with data entry work between groups are roughly similar. These information will collected in a pre-screening survey prior to task distribution.

3.5 Recruiting

One potential shortcoming of using the Mechanical Turk for a between-subject design study is that the system itself does not have means to either eliminate participants who had previously taken the study or randomize participants' task. However, this issue can be resolved by including a screening process in our task web page, which would screen out past participants based on the Turker's worker ID.

Once a participant decides to work for our HIT (Human Intelligence Task), he/she will be forwarded to a pre-screening web page, where we will ask him/her a few background questionnaires and confirm that his/her MTurk worker ID is not in our records. Once it has been confirmed that the Turker did not participate in our experiment before, he/she would be directed to one of the four treatment pages. The specific procedure can be seen in Appendix 7.2.

4 Sample Size, Power, and Cost Calculations

We want to find a sufficiently large and reasonably sized sample that will allow us to test the effects of the different treatments on the quality ratio of work completed by Turkers. For our study, we will aim for a power of .80 and a confidence level of .95, following the standards in the literature [23]. We will investigate the minimum sample sizes needed in order to detect reasonable differences between treatment means, and then scale those sample sizes to make use of our experimental budget of $5000, taking into account the fact that Amazon charges 10% of what the employer pays the workers as an additional MTurk fee.

4.1 MTurk Pilot Study

To calculate optimal sample sizes for each of our treatment cells, we will run a pilot study to obtain preliminary numbers for the mean and standard deviations of the quality ratios pertaining to the each of the treatment groups. The pilot study will collect thirty responses from each of the four tasks above, and will provide us with the expected means and standard deviations of quality ratios for the treatment groups.

With estimated means and standard deviations of the treatment groups, we will be able to calculate the sample size needed in order to discern a minimum detectable difference between treatment groups with a power of .80 and a confidence of .95.

4.1.1 MTurk Pilot Study Cost

For our pilot study, our target sample size is 120 participants (30 participants for two tasks). The compensation per Turker will be a lump sum of $2 for No-Gift treatments and $3 for Gift treatments. Coupled with the 10% MTurk fee, we calculate the total pilot study cost as $330, as shown below:

$$(\$2 \times 60 + \$3 \times 60) \times 1.1 = \$330 \tag{1}$$

4.2 MTurk Field Experiment

Since we have not yet run the pilot study, we will look back to previous literature in order to estimate the optimal sample sizes. Because of our particular study design, we have four minimum detectable differences to consider when calculating our overall required sample-size.

Table 3: Parameter of Interest by Treatment Groups[3]

Treatment	*No Gift*	*Gift*
Non-Verification	P_A	P_B
Verification	P_C	P_D

We want to detect three main results in our study. Namely:

1. Gift effect exists in both Verification and Non-Verification treatments
2. Reputation effect exists in both Gift and No-Gift treatments
3. Gift effect in Verification treatment is larger than gift effect in Non-Verification effect

So we would want to detect a meaningful difference between

1. Gift effect
   (a) P_A and P_B
   (b) P_C and P_D
2. Reputation effect
   (a) P_A and P_C

[3]Here, P is the quality ratio and the subscript refers to the different treatment groups. For example, A is No-Gift, Non-Verification, and B is Gift, Non-Verification.

(b) P_B and P_D

3. $P_B - P_A$ and $P_D - P_C$

We can find optimal sample size of treatment cells for detecting each minimum differences with a given significance level and power. For example, we can get an optimal sample size of the No Gift/Non-Verification group for both gift effect detection (1-(a)) and reputation effect detection (2-(a)). Then we can find our desired sample size by choosing the maximum required sample size in each treatment group. The procedure is represented in the following table:

Table 4: Optimal Sample Sizes, by treatment group [4]

Treatment	*No Gift*	*Gift*
Non-Verification	$max(n_A^{1-(a)}, n_A^{2-(a)})$	$max(n_B^{1-(a)}, n_B^{2-(b)})$
Verification	$max(n_C^{1-(b)}, n_C^{2-(a)})$	$max(n_D^{1-(b)}, n_D^{2-(b)})$

We look at previous literature to set a minimum detectable difference for sample size calculation purposes. For gift effect conditional on No Verification(1-(a)) and reputation effect conditional on No Gift(2-(b)), we will take into account observations made by Kube, et al. (2012) and Al-Ubaydli, et al. (2008), respectively. Unfortunately, there were no previous literature on gift effect conditional on Verification (1-(b)) or reputation effect conditional on Gift(2-(b)). However, the sample size calculations for these cases can be done once we have results from our pilot study.

4.2.1 Testing Gift Effect

In order to test our gift effect, we can use past literature that has tested the effects of Gift vs. No-Gift, given Non-Verification (P_A and P_B). We are unable to test the effects of Gift vs. No-Gift, given Verification because such literature has not been published.

Kube (2012), just as our proposed experiment, examined worker reciprocity but in the context of transcribing books; they measured the effect of No-Gift and lump-sum Gift treatments on the quality ratio of work completed. As mentioned in ?refsec:parameter-of-int we borrowed our definition of quality ratio from Kube (2012), which defined the term as the ratio of correctly entered entries to the total number of entries. We will use the means and standard deviations of the quality ratio observed by Kube (2012) to find the sample size necessary to test the No-Gift/Non-Verification treatment against the Gift/Non-Verification treatment.

They found that the difference in mean quality ratios between the two treatment groups was about 0.05 with standard deviations of 0.1168 for the No-Gift case and 0.0890 for the Gift case. We will use 0.05 as our minimum discernible difference, 0.1168 as the standard deviation for the No-Gift/Non-Verification group, and 0.0890 as the standard deviation for the other three treatment groups. Using the Sadoff, Wagner, and List sample size formula to calculate the optimal sample size for these numbers, we obtain the following sample size:

$$N = \left(\frac{t_{\alpha/2} + t_\beta}{\delta}\right)^2 \left(\frac{\sigma_1^2}{\frac{\sigma_1}{(\sigma_1+\sigma_2)}} + \frac{\sigma_2^2}{\frac{\sigma_2}{\sigma_1+\sigma_2}}\right) = \left(\frac{1.96+0.84}{.05}\right)^2 \left(\frac{.1168^2}{\frac{.1168}{.1168+.0890}} + \frac{.0890^2}{\frac{.0890}{.1168+.0890}}\right) \approx 133 \quad (2)$$

This is the estimated total number of participants we need in order to observe a minimum detectable gift-effect as demonstrated by Kube (2012) with the significance level of 0.05 and power of 0.80. We can further specify this number by treatment cell based on the predicted standard deviation by cell. In particular, since the cost per participant across treatment differ ($2.2 vs. $3.3), the ratio of the number of participants in Gift and No-Gift group can be calculated as the following:

[4] The subscript here refers to the treatment group (A,B,C,D) as defined in Table 3, while the superscript refers to the difference detected as defined above. For example, a superscript of 1-(b) is referring to the gift effect as measured by P_C and P_D.

$$\frac{n_1}{n_0} = \sqrt{\frac{c_0}{c_1}}\frac{\sigma_1}{\sigma_0} \tag{3}$$

$$\frac{n_{\text{Gift}}}{n_{\text{No Gift}}} = \sqrt{\frac{2.2}{3.3}}\frac{0.0890}{0.1168} \approx 0.6222 \tag{4}$$

Therefore, rounding up the calculation, the Gift-group should have 52 participants whereas No-Gift group should have 82.

4.2.2 Testing Reputation Effect

In order to test our reputation effect, we can use past literature that has tested the effects of Verification vs. Non-Verification, given No-Gift (P_A and P_C). We are unable to test the effects of Verification vs. Non-Verification, given Gift because such literature has not been published.

Al-Ubaydli, et al (2008) also examined worker reciprocity in a gift-exchange environment with workers packing envelopes. Further, Al-Ubaydli takes into account possible worker reputation concerns by having some treatments possibly lasting two days instead of one, contingent on the fact that the worker produces sufficient-quality work on the first day. Thus, the worker may have an incentive on the first day to put in more effort to preserve their reputation in order to be hired for the second day.

To examine worker effort, Al-Ubaydli measured the quantity of envelops processed as well as the error ratio, which is analogous to our quality ratio. However, this study has a key difference from ours in that the complexity of their task allows different types of errors at multiple levels. On the other hand, our task is simple enough that we only allow one type of error. In our study an entry with an error is counted as an incorrect entry, while an entry with no errors is counted as a correct entry. Nevertheless, we will assume that the difference in experimental design is insignificant for the results and focus on Al-Ubaydli's "recording error ratios" for our pre-pilot-study calculations to find a sufficient sample size for the Verification treatment. Recording errors are the errors committed in the final stage of Al-Ubaydli's experimental task, which involved a simple administrative task of verifying each letter that was sent. The recording errors are most relevant to our parameter of interest, because it too is based on a documenting task that is simple to complete accurately.

In order to get an idea of how reputation affects worker effort, we will be looking at the treatment groups 1-day-piece (workers hired for one day with a lump-sum wage plus a small piece-rate wage per envelope completed) and 2-day-piece (workers hired for two days with a lump-sum wage plus a small piece-rate wage, provided that the worker produces sufficient quality on the first day). Unfortunately, the differences of error ratio between these conditions are insignificant. The standard deviations are larger than the means: Al-Ubaydli found that the difference in mean errors between these two treatment groups was about .084, and the standard deviations were .403 for the 1-day-piece group (no reputation concerns) and .278 for the 2-day-piece group (adding reputational concerns). We believe this can be attributed to a number of factors. As the authors have noted, workers may have exhibited effort primarily in the form of increased number of output at the expense of quality of work (less error rate). The concern for reputation may not have been as impactful in the given study design, as the worker's relationship with the employer is temporary (one-time), and the workers may have more profitable job opportunities than coming back to the 2nd day of the experimental task.

Nevertheless, because the study has the most analogous measures of effort and using these numbers would give us a generous upperbound, we conduct our pre-pilot-study calculations with Al-Ubaydli's data. Using the Sadoff, Wagner, and List sample size formula to calculate the optimal sample size for these numbers, we obtain the following sample size:

$$N = \left(\frac{1.96 + 0.84}{.075}\right)^2 \left(\frac{.403^2}{\frac{.403}{.403+.278}} + \frac{.278^2}{\frac{.278}{.403+.278}}\right) = 515.29 \approx 516 \tag{5}$$

Since our verification of work will not incur additional costs, the sample size ratio between the

Verification Group and Non-verification group can be calculated as the following:

$$\frac{n_1}{n_0} = \frac{\sigma_1}{\sigma_0} \tag{6}$$

$$\frac{n_V}{n_{NV}} = \frac{.403}{.278} \approx 0.6898 \tag{7}$$

Rounding up the numbers, we conclude that the estimated required sample size for the Verification treatment is 211 and Non-Verification treatment is 306. These numbers are significantly larger than our numbers for the gift effect, which is most likely a limitation of the Al-Ubaydli paper that we examined. The paper differs in experimental design from our experiment because their design allows for multiple errors per task, while ours doesn't, which may inflate the standard deviations of the quality ratios they observed and inflate our calculated sample size. We expect these problems to be solved after running the pilot study, but for the purposes of this paper, we will proceed with these preliminary calculations.

4.2.3 Overall Sample Size and Cost

From the above two equations, we calculate that the following sample sizes per cell are required to detect the desired gift and reputation effects, respectively:

Treatment type	**Control group**	**Treatment group**
Gift, assuming No Verification	52	82
Verification, assuming No Gift	306	211

There is a lack of previous literature that deals with the interaction of the gift and reputational effects on worker reciprocity, so we do not have estimates for sample size needed to distinguish the gift effect across Verification vs. Non-Verification groups. Although Al-Ubaydli does have some treatment groups that could be interpreted as a Gift/Verification group, we will not use their numbers for standard deviation because we believe the differences in our experimental designs will lead to misleading estimates. When we obtain results from the pilot study, optimal sample sizes can be calculated similarly to the two calculations above.

Because we do not have the pilot study results to account for the interaction between the gift and reputation effects, for now we will make a preliminary sample size calculation assuming that the minimum detectable difference in gift-effect across Verification groups is larger than the minimum detectable difference in reputation effect given No Gift. Since detecting reputation effect requires the largest sample size under this assumption, we determine the sample size of each cell based on our calculations in 4.2.2. Thus, we will need the following optimal sample sizes for our treatment cells:

Treatment	*No Gift*	*Gift*
Non-Verification	306	306
Verification	211	211

This leads to a total sample size of 1034 employees, which will cost us the following:

$$\$2 \times 1.1 \times (306 + 211) + \$3 \times 1.1 \times (306 + 211) = \$2843.5 \tag{8}$$

However, this leaves $\$5000 - \$300 - \$2843.5 = \$1826.5$ of our experimental budget untouched. We will use our budget fully by proportionally scaling the sample size of each treatment group until we utilize our entire budget. Our treatment sample sizes thus become the following:

Treatment	*No Gift*	*Gift*
Non-Verification	502	502
Verification	346	346

We now have a total sample size of 1,696 workers. We believe that this is a reasonable sample size because a past study obtained 270 participants in 48 hours by offering 80 cents for completing a 30

minute survey [4]–we are offering a much more competitive wage and we will be able to post the job for a longer period of time.

The total experiment cost, utilizing the entire budget, is shown below:

$$\$2.5 \times 1696 \times 1.1 + \$330 = \$4664 + \$330 = \$4994 \tag{9}$$

4.2.4 Minimum Detectable Effects

With our new sample sizes, we are able to have a smaller minimum detectable difference between treatment group means than the .05 and .084 found in Kube, et al (2012) and Al-Ubaydli, et al (2008).

The minimum detectable differences between for the gift effects and the Verification (reputation) effects are calculated by solving two the equations below:

$$\left(\frac{1.96+0.84}{x}\right)^2 \left(\frac{.1168^2}{\frac{.1168}{.1168+.0890}} + \frac{.0890^2}{\frac{.0890}{.1168+.0890}}\right) = 502 \tag{10}$$

$$\left(\frac{1.96+0.84}{y}\right)^2 \left(\frac{.403^2}{\frac{.403}{.403+.278}} + \frac{.278^2}{\frac{.278}{.278+.403}}\right) = 346 \tag{11}$$

We find that x, the minimum detectable difference for the gift effect given our new sample size, is .0154, and y, the minimum detectable difference for the reputation effect given our new sample size, is .0613.

5 Predictions

Our experiment aims at identifying the determining factors of effort reciprocity when workers are inserted into a gift exchange setting.

5.1 Manipulation Check

In order to check that our gift manipulations and reputation manipulations are effective such that each of the factors impact worker effort, we will first run a pilot study with a total of 120 samples.

If our gift of $1 additional lump-sum wage is effective, workers in the Gift treatment would exhibit better per unit quality of work compared to the workers in the No-Gift treatment.

Although we are pretty confident that the gift would be effective, as lump-sum gifts have led to effort reciprocation in numerous previous literature [8, 11, 15, 20], there is a possibility that our gifts turn out to be ineffective. Indeed, some previous researchers have found that while workers to reciprocate to negative gifts, that is exert less effort in response to wage decrease, they do not respond with more effort to positive gifts [20]. If that is the case in our study, one potential cause may be that our gift was monetary. A number of previous research has demonstrated that monetary gifts are less effective compared to non-monetary gifts such as water bottles or origami [17, 19].

Also, if our verification manipulation is effective, we expect that workers in the Verification treatment to exhibit better per unit quality of work compared to the workers in the Non-Verification treatment, as they would consider the impact that their quality of work has on their reputation. Based on List, 2005, in which card dealers' concern for reputation was determined by whether or not the consumers had access to experts who could examine the quality of the card, we predict that the Turkers in our experiment would also show increased effort in the Verification treatment compared to the Non-Verification treatment.

5.2 Hypotheses and Expected Results

Assuming that our manipulations are effective, in our actual experiment we would focus on identifying how the reputation effect interacts with the gift-effect. To do so, we will examine whether the overall gift effect is driven by the Verification group or the Non-Verification group. We have three hypotheses. Namely, when there is verification, the per unit quality difference between Gift treatment and No-Gift treatment would either increase, be unaffected, or decrease compared to when there is no verification.

Our main prediction is that the per unit quality across Gift and No-Gift treatments are driven by the Verification group. This would imply that gift-effect in the labor market is primarily driven by the worker's reputation concern. The prediction is supported by the List, 2005 study that showed that reciprocity in commodity market is absent when the seller's reputation is not at stake. A series of dictator games in labs have also demonstrated that people do not show altruism when their action is not subject to evaluation, implying that our regard for others may inherently be driven by the possibility of evaluation.

Alternatively, the overall gift effect may be similarly driven by both Verification and Non-Verification groups. This result would imply that there is an intrinsically motivated regard for others behind our reciprocation to gifts, but that reputation also matters. In lab economic games, participants have shown altruism in void of other preconditions, but have shown even more generosity toward their partners when they are subject to evaluation [16, 25, 28]. In wage-effort matching games, employees have a general tendency to match their effort to the wage they are given, but does so to a greater degree when they are playing sequential games with the same partner rather than repeated one-shot games [7, 13]. Therefore, we may predict that both intrinsic regard for the employer and concern for own reputation drives the workers to reciprocate to gifts, in a way that reputation concern does not crowd out altruistic concerns.

Finally, gift-effect may even be driven by the Non-Verification group. This would imply that reciprocity is primarily driven by concern for others, and that the concern for reputation actually crowds out the gift-effect. If the willingness to reciprocate to another person's kindness is an intrinsic motivation that can be crowded out by the extrinsic concern for reputation, the Verification condition may even cause an over-justification effect, or the de-motivation of an intrinsically motivated activity by making the extrinsic motivation salient [12, 21]. If this is the case, we would find no gift effect in the Verification group. Indeed, it has been demonstrated in previous literature that extrinsic incentives such as reputation can crowd out altruistic behavior [2, 13]. For instance, offering monetary rewards for blood donation decreases the number of donors to one-half [24].

5.3 Linear Regression Model

In order to estimate the average quality ratio in each of our four treatment groups, we will be using an OLS linear equation. Our model is specified in Equation 12:

$$\mathbf{P}_i = \beta_0 + \beta_1 \mathbf{G}_i + \beta_2 \mathbf{V}_i + \beta_3 (\mathbf{G}_i \times \mathbf{V}_i) + \epsilon_i \tag{12}$$

Where $\mathbf{P}_i$ is the quality ratio, $\mathbf{G}_i$ is a dummy variable for the two Gift treatments, and $\mathbf{V}_i$ is a dummy for the two Verification types. Using this model, we see that different sums of the coefficients will be measures for the average quality ratio in each of the treatment groups, as listed in Table 5.

Table 5: Regression Coefficient Interpretation

	$\mathbf{G}_i = 0$	$\mathbf{G}_i = 1$
$\mathbf{V}_i = 0$	β_0	$\beta_0 + \beta_1$
$\mathbf{V}_i = 1$	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

The sums of certain regression coefficients above relate to the different treatment groups that we are analyzing. For the No-Gift, Non-Verification condition, β_0 is the average predicted quality ratio. For No-Gift, Verification condition, $\beta_0 + \beta_2$ is the average predicted quality ratio, and so on. Putting these coefficients in terms of our main prediction, $\beta_0 = 0$, meaning that no one would be willing to work, or there would be no correct entries, with No-Gift and Non-Verification. $\beta_1 = 0$ in our main prediction as well because we predict that the addition of a gift will have no effect if work is not being verified. However, we expect $\beta_2 > 0$ since we expect the verification effect to improve quality ratio. Lastly, we expect $\beta_3 > 0$ since we expect there to be a slight increase in quality ratio with both a gift and verification. For our 1st alternative, $\beta_1, \beta_2, \beta_3, \beta_4 > 0$. This means that we expect workers to exert some positive effort even if not given a gift and in the Non-Verification condition. We also expect increases across the boar for each additional treatment condition (Gift, Verification, and both). For alternative 2, $\beta_0 = 0$, $\beta_1 > 0$, $\beta_2 > 0$, $\beta_3 = 0$. This means that workers will not exert any kind of effort without verification and without a gift. However, workers will increase their quality ratio if given a gift ($\beta_1 > 0$) and if given a Verification

condition ($\beta_2 > 0$). However, $beta_3 = 0$ means that having both verification and a gift will not increase quality ratio, meaning that the Gift and Verification variables are independent.

We chose not to include any control variables (such as Age, Sex, and other demographics) because we are assuming that our sample is sufficiently randomized. As such, including additional covariates would not change the estimates on the treatment indicators. After gathering the data from our experiment and running this regression, however, we will investigate the typical OLS assumptions such as normality of residuals and heteroskedasticity in order to determine if any further transformations or controls need to be included in the model.

6 Concerns and Further Investigation

Through our literature review on the benefits and shortcomings of using Mechanical Turk to conduct experiments, we have identified and addressed some concerns that may come about from using Mechanical Turk. First, a survey conducted has shown that only 70.79% of workers on Mechanical Turk list their reason for working as "Compensation", while other reasons for working include boredom, curiosity, fun, and education.[4] Since approximately 30% of workers are not using Mechanical Turk for primarily financial reasons, there exists the possibility that the gift effect will be "muted" across a large percentage of our sample. Second, we believe there may be some concerns regarding the external validity of our experiment giving the demographics of MTurk workers. Studies have shown that MTurk workers are generally from the US and India, and that the US workers over-represent Asians and under-represent Blacks and Hispanics [27]. Third, the issue of non-random attrition resulting from the ability of MTurk workers to quit working in the middle of the task is a point of concern. If a worker believe that a certain treatment task is too complicated or boring, the participant may be likely to drop out, and a confound is introduced into the experiment. Different treatment groups will have a type of self-selection problem since the pools of subjects who participated will be different, so experimenters should be careful to make sure that attrition rates are similar across treatments. [29].

In addition to confounds of the current experiment, there are a couple of areas for future investigations that we would like to consider. First, the use of MTurk, as stated above in methods, prevents us from observing many "soft" factors of the participants such as their mood and the speed at which they are completing a task. However, experiments in the past, such as Gneezy and List, have observed that the gift effect fades over time [15]. In order to determine if the effort reciprocity effect will carry over long-term, we propose a move away from MTurk in order to calculate the extraction rate of the participants over a certain period of time. Another alternative to moving away from MTurk would be to develop a custom-built experiment application in order to monitor the rate at which our participants are working. Second, and similarly to measuring extraction rate, we would like to observe the long-term gift effect of repeated interaction between employees and employers. In Gachter and Falk's investigation on reputation and reciprocity, they observed that reciprocity was strengthened in a repeated game compared to a one-shot game, and we would like to determine if this repeated effect holds for our treatments as well by using a repeated games. [11]. Additionally, Kube has found that an in-kind gift, such as gifting a thermos, results in a larger, more significant gift effect compared to a monetary gift. [19] If we want a larger gift effect, we could also use a physical gift although this may require a move away from MTurk as well or a alternative delivery system.

7 Appendix

7.1 Task Instruction Samples

In this section, we list out sample of the instructions we will supply to our participants. The first sample (General) is supplied to all participants and gives participants information about the task that they will be conducting, what their work will b.e used for, and the monetary value of the task to the employer. The next four instruction samples are the instructions that will be added on to the general instructions depending on the treatment group that our participant is in. These add-ons give information about whether the participant's work will by verified or not, and whether or not there will be a gift, phrased as a "bonus", for a participant completing the work.

Figure 2: Task Instructions, General

General

Alumni Database Update

The University of Chicago is updating its alumni records to initiate its new alumni gift requesting campaign based on personalization. Researches have shown that people respond more readily to tailored information/advertisement. Therefore, the university hopes to increase the effectiveness of their alumni gift request by sending personally tailored gift letters. These letters would refer to each recipient alumni's college experience, such as the programs they enjoyed or student organizations they were a part of.
Your task is to update the alumni database of the University of Chicago by reading hand-written alumni surveys and entering data in specific data fields. The information updated will be used to request alumni gifts. The university has run similar campaigns every year, and the university expects this new program to increase the amount of donation by 30%. Based on this information, we expect to collect roughly $0.30 per alumni with this new campaign.
Therefore, every page of information you encode accurately will worth roughly $0.30 to the university.

Figure 3: Task Instructions, by treatment

Non-Verification, No-Gift

We will not be reviewing your work.
However, please try to be as accurate as possible. Because the new gift-request scheme relies heavily on tailoring to the individual, any incorrect information would hurt the University's gift collection.
Regardless of the amount of data entry you get done, you will be given $2 after the 10 minute session.

Non-Verification, Gift

We will not be reviewing your work.
However, please try to be as accurate as possible. Because the new gift-request scheme relies heavily on tailoring to the individual, any incorrect information would hurt the University's gift collection.
Also, we are offering a $1 bonus to you for helping us out.
Regardless of the amount of data entry you get done, you will be given $2 plus a $1 bonus after the 10 minute session.

Verification, No-Gift

We will be reviewing your work, and will reject your submission if the completed entries are not sufficiently accurate.
Please try to be as accurate as possible. Because the new gift-request scheme relies heavily on tailoring to the individual, any incorrect information would hurt the University's gift collection.
Regardless of the amount of data entry you get done, you will be given $2 after the 10 minute session.

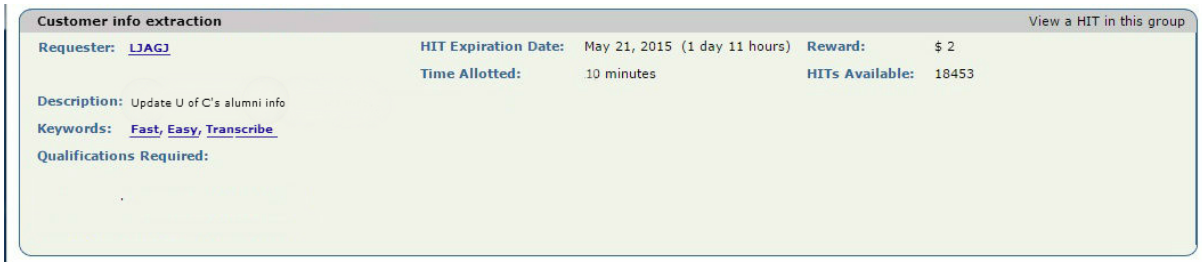
Verification, Gift

We will be reviewing your work, and will reject your submission if the completed entries are not sufficiently accurate.
Please try to be as accurate as possible. Because the new gift-request scheme relies heavily on tailoring to the individual, any incorrect information would hurt the University's gift collection.
Also, we are offering a $1 bonus to you for helping us out.
Regardless of the amount of data entry you get done, you will be given $2 plus a $1 bonus after the 10 minute session.

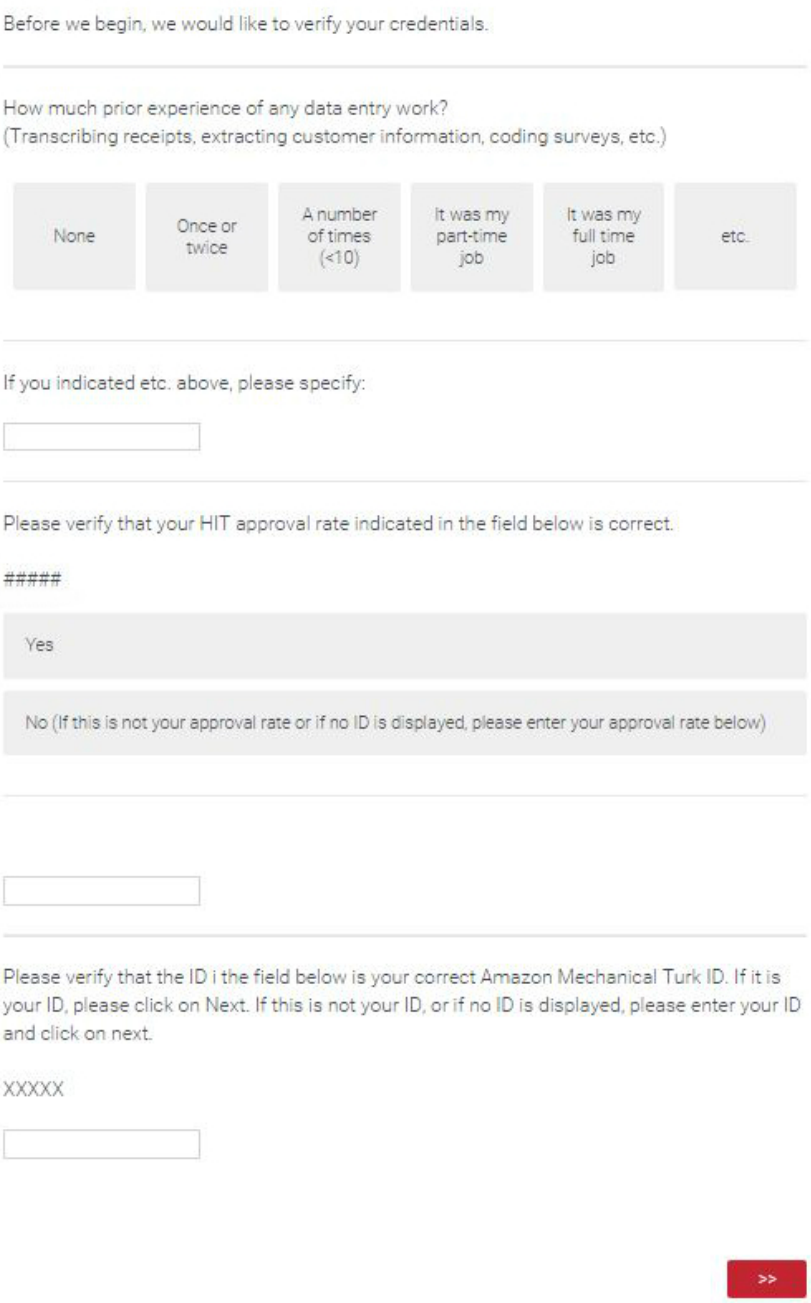
7.2 Sample Task Procedure

Following is a sample advertisement that the Turkers will see for our experiment:

Figure 4: HIT Advertisement



Once the Turker accepts the HIT, the following screen will appear:



If the Turker is verified as an eligible participant, he/she will be directed to one of the four the actual tasks. If not, he/she will be notified that they do not meet the requirements of the task and asked to return the HIT.

References

[1] Akerlof, George A. "Labor contracts as partial gift exchange." The Quarterly Journal of Economics (1982): 543-569.

[2] Al-Ubaydli, Omar, et al. "Carrots that look like sticks: Toward an understanding of multitasking incentive schemes." Southern Economic Journal (2014).

[3] Al-Ubaydli, Omar, Andersen, Steffen, Gneezy, Uri, and John List. (2008) "For Love or Money? Comparing the Effects of Non-pecuniary and Pecuniary Incentive Schemes in the Workplace." George Mason University.

[4] Behrend, Tara S., et al. "The viability of crowdsourcing for survey research."Behavior research methods 43.3 (2011): 800-813.

[5] Bellemare, Charles, and Bruce Shearer. "Gift exchange within a firm: Evidence from a field experiment." (2007).

[6] Bénabou, Roland, and Jean Tirole. Incentives and prosocial behavior. No. w11535. National Bureau of Economic Research, 2005.

[7] Ben-Ner, Avner, et al. "Reciprocity in a two-part dictator game." Journal of Economic Behavior & Organization 53.3 (2004): 333-352.

[8] Charness, Gary, and Ernan Haruvy. "Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach." Games and Economic Behavior 40.2 (2002): 203-231.

[9] Dellavigna, Stefano and John List, Ulrike Malmendier, and Gautam Rao. "Voting to Tell Others". Working paper. January, 2015

[10] Falk, Armin, and Urs Fischbacher. "A theory of reciprocity." Games and Economic Behavior 54.2 (2006): 293-315.

[11] Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. "Reciprocity as a contract enforcement device: Experimental evidence." Econometrica: journal of the Econometric Society (1997): 833-860.

[12] Fishbach, Ayelet, and Jinhee Choi. "When thinking about goals undermines goal pursuit." Organizational Behavior and Human Decision Processes 118.2 (2012): 99-107.

[13] Gächter, Simon, and Armin Falk. "Reputation and reciprocity: Consequences for the labour relation." The Scandinavian Journal of Economics 104.1 (2002): 1-26.

[14] Gneezy, Uri, Werner Güth, and Frank Verboven. "Presents or investments? An experimental analysis." Journal of Economic Psychology 21.5 (2000): 481-493.

[15] Gneezy, Uri, and John A. List. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." Econometrica 74.5 (2006): 1365-384.

[16] Haley, Kevin J., and Daniel MT Fessler. "Nobody's watching?: Subtle cues affect generosity in an anonymous economic game." Evolution and Human behavior 26.3 (2005): 245-256.

[17] Heyman, James, and Dan Ariely. "Effort for payment a tale of two markets." Psychological science 15.11 (2004): 787-793.

[18] Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith. "Social distance and other-regarding behavior in dictator games." The American Economic Review(1996): 653-660.

[19] Kube, Sebastian, Michel André Maréchal, and Clemens Puppe. "The currency of reciprocity: Gift exchange in the workplace." The American Economic Review 102.4 (2012): 1644-1662.

[20] Kube, Sebastian, Michel André Maréchal, and Clemens Puppe. "Do wage cuts damage work morale? Evidence from a natural field experiment." Journal of the European Economic Association 11.4 (2013): 853-870.

[21] Lepper, Mark R., David Greene, and Richard E. Nisbett. "Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis." Journal of Personality and social Psychology 28.1 (1973): 129.

[22] List, John A. The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. No. w11616. National Bureau of Economic Research, 2005.

[23] List, John A., Sally Sadoff, and Mathis Wagner. "So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design." Experimental Economics 14.4 (2011): 439-457.

[24] Mellström, Carl, and Magnus Johannesson. "Crowding out in blood donation: was Titmuss right?." Journal of the European Economic Association 6.4 (2008): 845-863.

[25] Mifune, Nobuhiro, Hirofumi Hashimoto, and Toshio Yamagishi. "Altruism toward in-group members as a reputation mechanism." Evolution and Human Behavior31.2 (2010): 109-117.

[26] Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. "Running experiments on amazon mechanical turk." Judgment and Decision making 5.5 (2010): 411-419.

[27] Paolacci, G., Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. Current Directions in Psychological Science, 23(3), 184-188.

[28] Piazza, Jared, and Jesse M. Bering. "Concerns about reputation via gossip promote generous allocations in an economic game." Evolution and Human Behavior 29.3 (2008): 172-178.

[29] Rand, David G. "The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments." Journal of theoretical biology 299 (2012): 172-179.