

Honesty in an Online Labor Market

Siddharth Suri and Daniel G. Goldstein and Winter A. Mason

Yahoo! Research
New York, NY, USA

Abstract

The efficient functioning of markets and institutions assumes a certain degree of honesty from participants. In labor markets, for instance, employers benefit from employees who will render meaningful work, and employees benefit from employers who will pay the promised amount for services rendered. We use an established method for detecting dishonest behavior in a series of experiments conducted on Amazon's Mechanical Turk, a popular online labor market. Our first experiment estimates a baseline amount of dishonesty for this task in this population. The second experiment tests the hypothesis that the level of dishonesty in the population will be sensitive to the relative amount that can be gained by dishonest reporting, and the third experiment, manipulates the degree to which dishonest reporting can be detected at the individual level. We conclude with a demographic and cross-cultural analysis of the predictors of dishonest reporting in this market.

Introduction

Many institutions and social systems depend upon some degree of honesty to function as intended. The legal system, for example, is predicated on honest testimony, and oaths are used with the goal of promoting truth-telling. Moreover, many economic transactions assume a truthful description of what is being sold or a promise that an agreement will result in a payment.

For online labor markets like Amazon's Mechanical Turk, honesty between the employers and employees helps the market to be efficient. Employers who trust the work of the employees, and employees who trust that payment will be rendered by the employer, both benefit from an environment in which honest dealing is the norm. Newer uses of online labor markets, such as conducting demographic or opinion surveys, have components that are hard to verify, meaning that under prevalent dishonesty, such markets would be of limited interest to researchers.

In online labor markets there are often fewer guarantees that a commitment will be honored than in traditional employment contexts, which makes the honesty of the workers all the more important. For example, on Mechanical Turk there is very little recourse available to an aggrieved worker

or requester. The fact that workers are anonymous only exacerbates this problem because a requester might not even know who has deceived them.

Standard economic models capture the belief that people trade off the benefits of cheating with the costs of getting caught (Allingham and Sandmo 1972; Becker 1968). On Mechanical Turk the costs of getting caught are very low. A worker might get their submission rejected. If this happens over and over they might get banned from the site, in which case they could simply make a new account (doing so only requires an e-mail address) and repeat the process. Since the costs are so low to getting caught, one might expect to see a large amount of cheating on Mechanical Turk. On the other hand, psychological studies indicate that people have an intrinsic motivation to avoid feeling like they are dishonest (Mazar, Amir, and Ariely 2008). Additionally, people may want to maintain the appearance that they are being honest (Hao and Houser 2011). Thus, it is not *a priori* clear how much dishonesty would be exhibited by workers in an online labor market. The central focus of this work is measuring the degree to which workers on Mechanical Turk are honest and exploring which factors that affect their honesty.

Related Work

Fischbacher and Heusi (2008) conducted a study which is the inspiration for this work. In a series of offline laboratory experiments, the authors had participants roll a die *in private* and report their roll. Participants were paid 1, 2, 3, 4, and 5 CHF¹ for the corresponding die roll and 0 CHF for a roll of six. Since the experimenter could not see the roll, the participant could report any integer between one and six. While each number would be expected 17% of the time, the subjects reported a roll of four 27% of the time and reported a roll of five 35% of the time. A roll of six, the lowest paying roll, was only reported 6.4% of the time, suggesting dishonest reporting.

In addition to this baseline treatment, the authors conducted additional treatments where they increased the stakes (by a factor of three), ensured the anonymity of the participants, and changed the victim of the lie from the experi-

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹One swiss franc (CHF) was roughly equal to \$0.80 USD at the time of the study.

menter to another subject. These treatments did not have a large impact on the distribution of reported rolls. Having the participants do the experiment more than once did, however, have a large impact on the distribution of reported rolls. In this case, repeat players reported five, the highest paying roll, 52% of the time, some 35 percentage points over chance.

Hao and Houser (2011) extended this work by using a two-stage experiment. In the first stage participants gave a probability distribution predicting the outcome of the roll of a four-sided die. In the second stage, participants either rolled the die publicly (with the experimenter watching) or privately (with the experimenter not watching). Participants knew which treatment they were in before the experiment started. They were paid more if they had placed higher weight on the actual outcome of the die. Participants reported similar distributions in both the public and private treatments indicating that they wanted to appear honest. But in the private treatment, significantly more rolls matched the most likely outcome in the predicted distributions than in the public treatment. Thus people lied more when given the opportunity.

Experiments

All three of the experiments described in this paper were conducted on Amazon's Mechanical Turk², a crowdsourcing website where requesters can post jobs, called human intelligence tasks (HITs), and workers can do those HITs for pay. After a worker submits their work for a HIT, requesters review the work and either accept it or reject it. There is a burgeoning literature that shows that the behavior of Mechanical Turk workers is comparable to laboratory studies (Buhrmester, Kwang, and Gosling 2011; Paolacci, Chandler, and Ipeirotis 2010; Horton, Rand, and Zeckhauser 2011; Suri and Watts 2011).

In this work, we present three experiments. Participants were only allowed to do each experiment one time. By tracking the IDs of the workers who accepted our HITs, we also restricted our analyses to include only participants who did one of the three experiments. Requesters can specify that workers who accept their HIT have certain qualifications, such as living in a certain country. In each of our experiments we used this qualification to obtain data from workers from the U.S. and India.

Single Roll: Baseline

Our first experiment sought to establish the baseline dishonesty level for this task on Mechanical Turk. Workers were instructed to answer a demographic questionnaire about themselves. They were told as payment they would receive a \$0.25 flat rate (as the HIT payment) plus \$0.25 times the roll of a die. The instructions indicated that if they did not have a die available they would be provided a link to a website that simulates die rolls. The website `random.org` provided the simulation of fair dice. The actual outcome of the die rolls were not recorded, only the reported rolls

that the participants made were recorded. If the participants reported the die roll honestly, the total expected payoff would be \$1.125. The minimum total payoff was \$0.50 and the maximum total payoff was \$1.75. The ratio of the maximum to minimum payoff is 3.5, thus there is a potentially large relative benefit to dishonesty.

We recruited 175 players from two HITs, 93 from the HIT that required workers to be from the U.S., and the remaining 83 from the HIT that required workers to be from India. The U.S. workers were 53.7% female, while the Indian workers were 29.3% female. Overall, 14.9% reported high school as the highest level of education attained, 17.1% reported having some college education, 46.3% reported having a bachelor's or associate's degree, and 21.7% reported attaining education higher than a bachelor's degree.

If the participants honestly reported their rolls, their reports would have followed a uniform distribution. To determine if the participants were dishonest, we compare the average reported roll to the mean of the uniform distribution which is 3.5. The average reported roll of the 175 participants was significantly higher than that which would be expected by chance, with a mean of 3.91 (Wilcoxon rank-sum test $V = 9870, p < 0.0005$). Moreover, the distribution of reported rolls is significantly different than uniform ($\chi^2 = 13.26, p < 0.05$). As shown in Figure 1, players reported more five and six outcomes and fewer one and two outcomes than would be expected. This also replicates a finding from Fischbacher and Heusi (2008), which is that individuals did not always cheat by choosing the highest-paying outcome. Instead, sometimes players reported a five, either because they wanted to maintain an appearance of being honest (Hao and Houser 2011) or because they wanted to avoid feeling dishonest (Fischbacher and Heusi 2008; Mazar, Amir, and Ariely 2008). Since Fischbacher and Heusi's and our study have slightly different incentive schemes it is not possible to directly compare the distribution of rolls reported by the participants in the two studies.

Single Roll: Low Variance

In the previous experiment, the ratio of the maximum to minimum payoffs was 3.5. It is possible that the incentive to be dishonest is so great that it is hard for the workers to resist. Conversely, it is possible that if the perceived gain one would obtain by being dishonest was smaller, participants might not deem it worthwhile to lie. This next experiment seeks to test this hypothesis.

The instructions and the general setup of this experiment were the same as the previous experiment. Again, the HIT had a \$0.25 flat rate (to avoid differential participant selection effects). In contrast to the first experiment, participants were promised an extra \$0.75 bonus, plus \$0.05 times the roll of a die. The expected total payoff was \$1.175, which was intentionally kept as close as possible to the mean total payoff of the previous experiments. Crucially, the minimum total payoff was \$1.05 and the maximum total payoff was \$1.30, so the ratio of the maximum to minimum total payoff was merely 1.24 (compared to 3.5 in the single roll, baseline experiment). The participants in this experiment can therefore maintain the appearance of honesty as they did in

²<http://www.mturk.com>

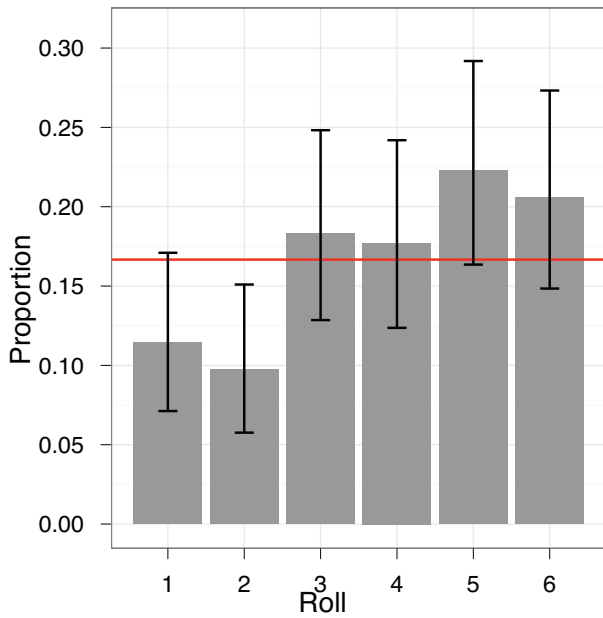


Figure 1: The distribution of rolls in the single roll, baseline experiment. Error bars are confidence intervals for two-sided binomial tests relative to chance ($p = 0.167$) with Bonferroni-corrected $\alpha = 0.05$.

the first study because they are only rolling one die, but they also have less incentive to cheat, as the proportional increase in payoff from the lowest roll to the highest roll was much lower.

In this experiment we recruited 267 players from two HITs: 140 from the U.S., and 127 from India. Comparable to the first study, the U.S. workers were 59.3% female, while the Indian workers were only 30.7% female. Overall, 13.1% reported high school as the highest level of education attained, 22.1% reported having some college education, 46.8% reported having a bachelor's or associate's degree, and 17.9% reported attaining education higher than a bachelor's degree.

A test of the mean shows that even with the reduced financial incentive, the average roll value reported was higher than that which one would expect to see by chance at 3.77 (Wilcoxon $V = 21046, p < 0.01$). Although this mean is not as extreme as that in the first study, there was no significant difference in means between the two studies (Mann-Whitney $W = 24346, ns$). Moreover, as can be seen in Figure 2, the distribution of reported outcomes is also fairly similar to that in the first study. That is, there is marginally significant under-reporting of twos and (non-significant) over-reporting of fives. Together, these results suggest that changing the relative benefit of being dishonest did not strongly affect the incidence of dishonesty.

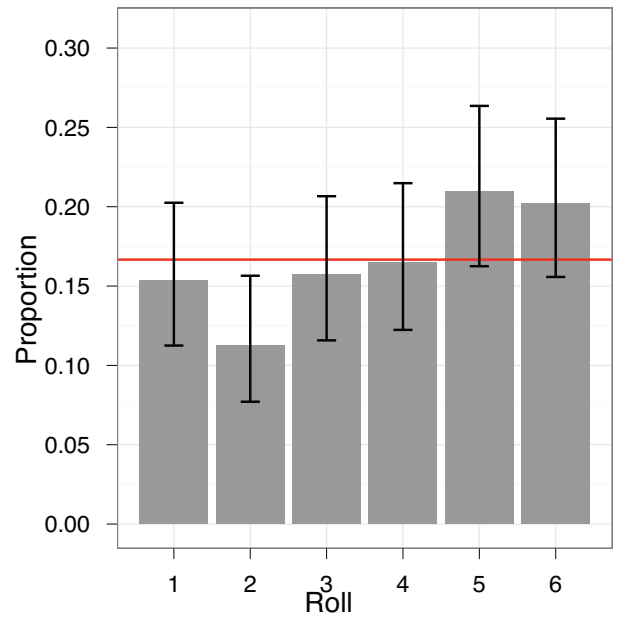


Figure 2: The distribution of rolls in the single roll, low variance experiment. Error bars are confidence intervals for two-sided binomial tests relative to chance ($p = 0.167$) with $\alpha = 0.05$.

Multiple Rolls

The data from the first two experiments are consistent with inflated reporting of the actual die rolls. However, with the design of the previous experiments one cannot tell if an *individual* is reporting their roll honestly or not. One can only infer the amount of dishonesty in the population. Participants may have an intuitive (or explicit) notion of this when they report their rolls. To investigate whether increasing the ability to detect cheating will lead to more honest reporting, in this next experiment, we had 232 participants roll a die 30 times³. The demographic questionnaire was the same as before, and the HIT payment was also kept consistent at \$0.25. This time, however, subjects were paid the sum of the 30 die rolls in cents in addition to the HIT payment. To avoid arithmetic errors, participants were asked to enter all their rolls in a 5 x 6 grid of input fields, and we calculated the sum and resulting bonus.

For this experiment we recruited a total of 233 participants, 108 from the U.S. and 125 from India. This experiment was somewhat more heavily male than the previous two, with the Indian participants 72% male and the U.S. workers 54.6% male. The distribution of education was comparable, with 15% reporting high school or lower, 20.1% reporting some college, 39.5% reporting associate's or bachelor's, and 25.3% reporting educational attainment higher than bachelor's.

³To speed up this process the link to `random.org` simulated five fair dice at a time.

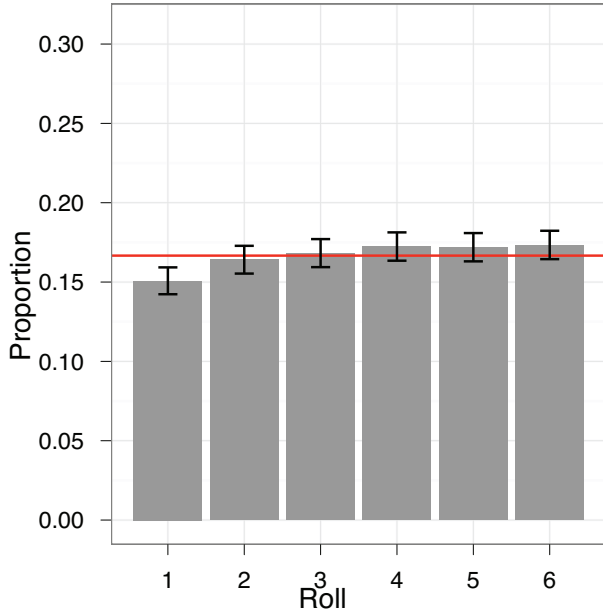


Figure 3: The distribution of rolls in the multiple roll experiment (when each participant rolled 30 dice). Error bars are confidence intervals for two-sided binomial tests relative to chance ($p = 0.167$) with $\alpha = 0.05$.

If the participants reported their score honestly, the expected payoff would be \$1.30. The minimum payoff was \$0.55 and the maximum payoff was \$2.05. Here the ratio of the maximum to minimum payoff was 3.73. All of these values were intentionally kept as close as possible to the corresponding values of the single roll, baseline experiment in order to keep as much constant between the two setups as possible. It is somewhat surprising, therefore, that participants reporting multiple dice rolls were *substantially* less dishonest than those reporting a single roll. On average, participants’ mean outcome was 3.57. While this is still significantly different from the expected mean of 3.5 (Wilcoxon $V = 12794556, p < 0.0005$), it also represents significantly less dishonesty than in the single-roll, baseline study (Mann-Whitney $W = 540850, p < 0.01$).

The distribution of rolls also showed a different pattern than in the single roll studies, as shown in Figure 3. Here, it appears that the participants were averse to reporting the worst outcome, significantly under-reporting the actual distribution of ones (two-sided binomial test, $p < 0.0005$). However, because none of the other outcomes were so clearly impacted, it suggests the participants were not simply changing the ones to, say, sixes.

Because every participant submitted 30 rolls, it was possible to test whether the distributions submitted by each participant was unlikely to have arisen by chance. Of course, given the very slight differences observed in the overall distribution, it is unlikely that it would be possible to detect any obvious dishonesty. Indeed, conducting a Wilcoxon rank-

Table 1: Mean Outcome by Qualifications

	0-89%	90-94%	95-97%	98-99%	100%
U.S.	3.71	3.57	3.60	3.63	3.46
India	3.48	3.56	3.65	3.55	3.54

sum test on each individual’s set of rolls and using a Bonferroni correction with $n = 232$, only three individuals’ averages were significantly greater than the expected outcome of 3.5. Surprisingly, only one of the 232 participants cheated to the fullest extent, reporting all sixes and earning the maximum, \$1.80.

Qualifications on Mechanical Turk As mentioned previously, HITs on Mechanical Turk can have a qualification associated with them restricting who can work on the HIT. For example, a requester on Mechanical Turk can specify which country a worker must be from in order to work a HIT. Also, Amazon keeps track of the fraction of HITs that each worker submitted that were accepted by the requesters. This is called a worker’s approval rating and a requester can require a worker to have an approval rating in a certain range to work on a HIT. In this study, we created 10 different HITs, splitting the worker population by country of origin (India vs. U.S.A.) and 5 levels of prior approval rate on Mechanical Turk: 0-89%, 90-94%, 95-97%, 98-99%, and 100%. Focusing on the approval rate, one might expect that the workers with a low approval rate would be more dishonest than workers with a high approval rate—although again, with the small incidence of cheating overall, this might be hard to detect. In fact, we found no significant differences in the means reported from the different approval rates (Kruskal-Wallis $\chi^2 = 4.4, ns$). Moreover, we found no differences between any of the qualifications. When comparing each subset against the fair mean with a Bonferroni-corrected ($n = 10$) Wilcoxon rank-sum test, only the group with the lowest approval rate (0-89%) in the U.S. and (marginally) the middle approval rate group (95-97%) in India were reliably different from the expected mean.

Dishonesty and demographics

Is the kind of dishonesty observed in these experiments predictable from demographics provided by the experimental participants? Fischbacher and Heusi (2008) posited that a cross-cultural comparison of dishonesty would be very interesting. Although we cannot be sure that the participants were honestly reporting their demographics, there is at least some prior evidence of consistency in reporting this type of information in online settings (Voracek, Stieger, and Gindl 2001). Moreover, two previous studies (Mason and Suri 2011; Rand 2011) found roughly 90% consistency of self-reported demographics taken from multiple studies on Mechanical Turk. For these analyses, because the configurations of the two single roll studies were comparable, we combine the single-roll baseline and single-roll low variance datasets to increase statistical power.

To begin with, we compare the reported outcomes of males and females. For both genders, the reported means are

Table 2: Generalized linear model predicting the reported roll value in the single-roll Experiment based on all reported demographics

Variable	Estimate	Std.		P-val
		Error	t-val	
Intercept	3.50	0.25	13.76	<2e-16***
Ed: Pre-College	0.11	0.16	0.67	0.50
Race:Other	0.37	0.29	1.28	0.20
Race:White	0.24	0.21	1.15	0.25
Income:5-20k	0.33	0.22	1.49	0.14
Income:20k-40k	-0.10	0.26	-0.40	0.69
Income:40k+	0.12	0.25	0.50	0.62
Age:25-30	-0.22	0.22	-1.02	0.31
Age: 31-40	-0.22	0.24	-0.89	0.37
Age: 40+	-0.04	0.25	-0.17	0.86
Sex: Male	0.27	0.17	1.60	0.11

significantly greater than the expected mean under honest reporting (Wilcox rank-sum test, one-tailed; Men: 3.91, $V = 19603$, $p < 0.0005$; Women: 3.71, $V = 11196$, $p < 0.05$). Although the average for men is higher than that for women, the difference in the means is not significant (Mann-Whitney $W = 22612$, ns).

Dividing education attainment into “Pre-college degree” (no college degree attained) and “College degree” (college degree and beyond) gives groups of 238 and 209, respectively. While both groups reported values that were significantly higher than what would be expected by chance (Wilcox rank-sum test, one-tailed; Pre-college: 3.87, $V = 17719$, $p < 0.0005$; College: 3.77, $V = 12281$, $p < 0.05$), the means also do not differ significantly (Mann-Whitney $W = 25646$, ns).

Income was collected as a continuous variable, which did not correlate significantly with the roll value reported (Pearson correlation .008, Spearman correlation .02, both p values $> .6$). When this variable was divided into four levels: \$0 to \$5k ($n=137$), \$5k to \$20k ($n=112$), \$20k to \$40k ($n=94$), and greater than \$40k ($n=104$), the respective mean roll values of these groups were 3.69, 4.04, 3.66, 3.90. The \$0 to \$5k and \$20k to \$40k groups did not differ significantly from 3.5 (p values .11 and .18) while the remaining two did ($p < .01$), though even the most extreme of these four means do not differ significantly from one another (Mann-Whitney $W = 6283$, ns).

Ethnicity was collected over a range of categories listed on the US census. However, due in part to the national targeting our the experiment “Asian Indian” ($n=210$) and “White” ($n=192$) received far more responses than the other categories (combined $n=45$). Both those declaring themselves Asian Indian and White reported values so high as to be unlikely to occur by chance (Wilcox rank-sum test, one-tailed; Asian Indian: 3.72, $V = 12647$, $p < 0.05$; White: 3.88, $V = 11690$, $p < 0.001$), but as with the other examples, the difference between groups was not significant (Mann-Whitney $W = 19190$, ns).

We used all these predictors in a generalized linear model to predict the reported roll value in the single roll studies.

Table 3: Linear mixed model predicting the reported roll value base on all reported demographics for the multi-roll Experiment

Variable	Estimate	Std.		t-val
		Error	t-val	
Intercept	3.62	0.07	51.14	
Edu: Pre-College	0.02	0.07	0.22	
Race:Other	-0.03	0.09	-0.27	
Race:White	-0.06	0.06	-1.03	
Income:5-20k	0.01	0.06	0.16	
Income:20k-40k	-0.02	0.08	-0.3	
Income:40k+	0.07	0.08	0.87	
Age:25-30	-0.11	0.06	-1.77	
Age: 31-40	0.00	0.07	-0.01	
Age: 40+	0.04	0.08	0.49	
Sex: Male	-0.02	0.05	-0.37	

As Table 2 indicates, no demographic predictor is significant in the regression and furthermore, in refining this model with the Akaike information criterion-based stepAIC routine (Venables and Ripley 2002), all of demographic predictors are eliminated, leaving only the intercept.

As an additional check, we used the same set of demographics to predict reported roll values in the multi-roll experiment. With 30 observations per individual, these data lent themselves to a linear mixed-effect model analysis. We modeled the reported roll value as a function of all observed demographics with participant as a random effect using the lme4 package in R (Bates, Maechler, and Bolker 2011). Consistent with the single roll analysis, Table 3 shows that no strong demographic predictors arise as significant in this model of the reported multi-roll values. In sum, there is reasonable support for the idea that low-level dishonest reporting spans our observed demographic categories and that the substantial differences between groups are not apparent.

Conclusion

Since reporting dishonestly on this task has such a low cost, it is surprising that there was not more dishonesty in these experiments, especially in the single roll experiments in which one cannot infer whether a specific individual reported dishonestly. One possible explanation for this is that workers were afraid that their work would be rejected if they lied. Future research might investigate the effectiveness of messages that assure that work will be accepted regardless of the content. This might make participants feel more free to deceive. Conversely, it might also turn a financial decision regarding what number to report into an ethical decision about taking advantage of the experimenter. Thus, it is not *a priori* clear how participants would behave this type of experiment.

While it is possible that some participants may not have visited the die-rolling website at all and instead reported numbers that appear to be random, we observe that this would nonetheless be a form of dishonesty given the experimental instructions, and furthermore be a form of dishonesty.

esty that happened to increase the participants' payments in all studies.

The results of the single-roll, baseline experiment and the single-roll low-variance experiment show that changing the percentage gain that people could earn by lying did not have a large impact on the levels of dishonesty. This leaves several possible explanations for why there was such a drastic reduction in the amount of lying in the multiple roll experiment. One possibility is that people did not want to risk being individually detected as lying in the multiple roll experiment. Another plausible explanation could be that people are more comfortable with the idea of lying once in the single roll setting than they are lying many times in the multiple roll setting. A third explanation could be that the multiple roll experiment put people in the frame of mind of a data entry task. If this were the case, it is easy to imagine them carefully copying the rolls and lying less.

In the multiple roll experiment, participants did not lie egregiously. Only one person reported 30 rolls of six, and only three reported sets of rolls that were unlikely to have come from a fair die. That said, the overall mean of 3.57 in the multiple roll experiment, and the lack of differences in mean between demographic groups suggests that while few cheated a lot, many cheated a little.

Acknowledgments

We would like to thank our colleagues at Yahoo! Research for useful feedback. All plots in this paper were made using `ggplot2` (Wickham 2009).

References

- Allingham, M. G., and Sandmo, A. 1972. Income tax evasion: A theoretical analysis. *Journal of public economics* 1:323–338.
- Bates, D.; Maechler, M.; and Bolker, B. 2011. *lme4: Linear mixed-effects models using Eigen and S4 classes*. R package version 0.999375-39.
- Becker, G. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76(2):169–217.
- Buhrmester, M. D.; Kwang, T.; and Gosling, S. D. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* . . .
- Fischbacher, U., and Heusi, F. 2008. Lies in disguise: An experimental study on cheating. Research Paper Series 40, Thurgau Institute of Economics and Department of Economics at the University of Konstanz.
- Hao, L., and Houser, D. 2011. Honest lies. Discussion paper, Interdisciplinary Center for Economic Science, George Mason University.
- Horton, J. J.; Rand, D. G.; and Zeckhauser, R. J. 2011. The online laboratory. *Experimental Economics*. To appear.
- Mason, W., and Suri, S. 2011. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*. Under Review.
- Mazar, N.; Amir, O.; and Ariely, D. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research* 45(6):633–644.
- Paolacci, G.; Chandler, J.; and Ipeirotis, P. G. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5:411–419.
- Rand, D. G. 2011. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*. To appear.
- Suri, S., and Watts, D. J. 2011. Cooperation and contagion in web-based, networked public goods experiments. *PLoS One* 6(3).
- Venables, W. N., and Ripley, B. D. 2002. *Modern Applied Statistics with S. Fourth Edition*. Springer. ISBN 0-387-95457-0.
- Voracek, M.; Stieger, S.; and Gindl, A. 2001. Online replication of evolutionary psychology evidence: Sex differences in sexual jealousy in imagined scenarios of mate's sexual versus emotional infidelity. In Reips, U.-D., and Bosnjak, M., eds., *Dimensions of Internet Science*. Lengerich, Germany: Pabst Science. 91–112.
- Wickham, H. 2009. *ggplot2: elegant graphics for data analysis*. Springer New York.