Journal of Behavioral Economics for Policy, Vol. 8, Issue 2, 55-60, 2024

Auditing with Rewards for Honesty and Punishments for Cheating

Austin Henderson[1]*

Abstract
Policies to deter misbehavior often use audits, which probabilistically detect misbehavior, and punishments. While punishments may be undesirable due to practical or ethical concerns, policies of auditing with rewards for detected good behavior are seldom used or studied. In an online experiment, I tested the effects of rewards for honest behavior, as well as combinations of punishments and rewards, on cheating. I found that small probabilities of large rewards were effective when combined with punishments, but small rewards were ineffective or backfired. These findings provide guidance for implementing rewards and contribute to knowledge of how auditing schemes affect behavior.

JEL Classification: D91, C91, D81

Keywords
auditing — deception — deterrence — risk — unethical behavior

[1] *Washington State University, Elson S. Floyd College of Medicine, Institute for Research and Education to Advance Community Health*
***Corresponding author**: austin.m.henderson@wsu.edu

Introduction

Cheating is an important behavior in situations wherein a party can exploit an information asymmetry to their advantage, such as in reporting income to a tax authority or hours worked to an off-site manager. Strategies to deter cheating have frequently involved the use of audits, which with some probability detect cheating and punish the cheater. Economics research has typically examined the impact of auditing on cheating through the lens of the economics-of-crime approach, in which the gains from cheating are weighed against the probability and cost of detection (Becker, 1968). This logic suggests, for example, that if the probability of detection is low and costly to increase, then the size of the punishment should be increased, an insight that has been influential in shaping real policing approaches.

In this study, I experimentally tested two questions that emerged from the incentives' perspective of auditing. The first and central question is: how do rewards for good behavior affect cheating? The second: what effect do negligible changes in expected value from audits, as produced from a low detection probability and small incentive, have on cheating?

The first motivation for studying these questions is straightforward: cheating can be costly, and identifying incentive schemes that can reduce cheating, particularly cost-effective ones, has clear policy value. Importantly, an optimal incentive structure is not necessarily one that maximally deters cheating but that balances the costs of deterrence against the value of the misbehavior deterred. Consider the example of public transport fare evasion: if more is spent on enforcing fare payment than the enforcement induces riders to pay, there is over-enforcement.

The second motivation to study rewards is that, in some circumstances, punishments are not attractive for moral reasons, and so even if rewards are less cost-effective than punishments they could still be an attractive tool. Continuing the fare evasion example, those who have the strongest financial incentives to free-ride are in poverty and could be greatly harmed by a fine. Rewarding good behavior attenuates this dilemma. While this could also mean rewards going to those not in poverty, or people who would have paid the fare regardless, these considerations should be weighed against the ethical imperative to minimize harm in pursuit of public policy goals. Individuals may also prefer to face incentive structures framed as rewards rather than punishments, even if the expected payoffs of actions are the same (e.g. Brink and Rankin, 2013). Potentially, this could improve the subjective welfare of those subject to audits or enable the implementation of auditing incentive systems that might otherwise engender political push-back.

I examined cheating behavior in two complementary online studies. In the first and primary study, participants were incentivized to cheat, faced known audit probabilities (of either 0, 1%, or 10%), and rewards and/or punishments for audited actions. The rewards for detected honesty varied between a small reward of \$1.00 or a large reward of \$10.00, and punishments were always for \$1.00, all of considerable magnitude given the default payoff for honest behavior in the study was around \$0.50. To study negligible changes in expected value, I tested the effect of a low audit probability, 1%, with both rewards and punishments of \$1.00 such that they only change the expected values of cheating/behaving honestly by \$0.01. Notably, in addition to the small size of the punishment/reward, this probability is smaller than previous auditing probabilities studied in experiments, though there are numerous real-world examples where the true probability of detection is that small. In the second study, presented and considered mainly in the supplemental materials of this manuscript, I investigated the role of social norms in driving the results found in the first experiment.

Methods

Theory

The primary experiment aimed to evaluate the efficacy of rewards and of small incentives, and to compare the efficacy of rewards and punishments. To motivate the analysis and facilitate a theoretically grounded comparison of rewards and punishments, I rely upon the workhorse Becker (1968) model, which uses a standard utility function that is concerned with outcomes, with the additional inclusion of an intrinsic cost of cheating, which facilitated a power analysis. Full utility models are presented in **Supplemental Materials S1**, and the power analysis is presented in Supplemental Materials S2.

The experiment was designed such that the expected value of cheating or honesty was equal across various combinations of incentives and audit probabilities. The equivalency of expected values provides clear comparisons of efficacy and it is theoretically important, as the model predicts that the effect of rewarding honesty will be equivalent to punishing cheating so long as the expected utilities of each action are identical. This model, in conjunction with plausible estimates of negligible risk preferences over small stakes (Rabin, 2000), also provides the baseline hypothesis that small rewards and punishments will have a negligible effect in comparison to no auditing at all because the changes to expected utility from changes in expected outcome alone should not cause honesty to become more preferable in any case.

The hypotheses I tested in the first experiment are as follows.

> H1: The rate of cheating will be significantly lower when the gap in expected value between cheating and honesty is substantially reduced.
>
> H2: Very small changes in the expected value of cheating will not lead to a significant difference in the observed rate of cheating.
>
> H3: Holding expected values constant, rewards for honest behavior will yield the same proportion of subjects who cheat as punishments for cheating.

Experimental Design

Participants were recruited via Amazon MTurk to participate in a 2-minute task, with a posted pay rate of $0.10 for completion. Amazon MTurk is one of the largest online platforms for task-based work, in which individuals choose to accept posted tasks for some promised payment upon satisfactory completion (as determined by the task lister). MTurk has recently become widely used in the social sciences to study a variety of phenomena, such as preferences for redistribution (Almås, Cappelen, and Tungodden, 2016) and the responsiveness of deception to detection probabilities and the size of the fine (Laske, Saccardo, and Gneezy, 2018). Upon accepting the task on MTurk, participants were directed to the online survey platform Qualtrics where they participated in the experiment. After completing the task, participants were given a matching code to enter into Qualtrics and thus receive payment according to their actions and whether or not they had been audited. Further details on the experimental procedure and full instructions are presented in the **Supplemental Materials S3**.

Participants played a one-shot cheating game. In the game, the decision maker observed the outcome of what essentially was a coin flip, with outcome [H,T] equally likely with p=0.5. The decision maker then faced a choice: whether to report H, which has a low payoff, and T, which has a high payoff. In either case, they face some probability (p), (e.g. p = 0.1), of being audited, in which the true flip is revealed to the auditor and they are either rewarded or punished according to the treatment.

The exact potential payoffs were calibrated such that the expected values of cheating and honesty were the same across treatment group pairs. For example, in the 10% chance of reward treatment, honestly reporting an H earned $0.50 plus the chance of reward, and reporting a T earned $1.00 guaranteed. The expected value of being honest with a $1.00 reward was $0.60, so the difference in expected value was $0.40 in favor of misreporting. In the equivalent punishment treatment, honestly reporting an H has a guaranteed value of $0.60, and cheating earned $1.10, minus the chance of being punished by one dollar. Thus, in both treatments, the expected value of honesty was $0.60 and dishonesty $1.00, with a gap of $0.40. The expected values of honesty and cheating in each treatment are summarized in Table 1.

Table 1. Expected values of honesty and cheating in each treatment

Treatment	EV of Honesty	EV of Cheating
Control	$0.50	$1.00
1% Punishment	$0.50	$0.99
1% Reward	$0.51	$1.00
10% Punishment	$0.60	$1.00
10% Reward	$0.60	$1.00
1% Reward $10	$0.60	$1.00
1% Reward $10 or Punishment	$0.60	$0.99

Results

A total of N = 815 participants faced an incentive to cheat in this experiment. Of those, 118 faced a 1% chance of reward and 32 (27.12%) cheated, 126 faced a 10% chance of reward and 43 (34.13%) cheated, 122 faced no chance of audit or consequences (control) and 29 (23.77%) cheated, 123 faced a 1% chance of punishment and 19 (15.45%) cheated, and 124 faced a 10% chance of punishment of whom 13 (10.48%) cheated. Facing a 1% chance of a $10 reward, 16 of 102 cheated (15.69%), and facing a 1% chance of a $10 reward as well as a $1 punishment 7 of 103 cheated (6.80%).

Figure 1 displays the proportion of individuals who cheated, separated by treatment, with the control treatment setting a baseline. Supplemental Materials Table 2 presents all pairwise comparisons of the proportion of participants who cheated in each treatment. Achieved power for all estimates is presented in **Supplemental Materials S4**.

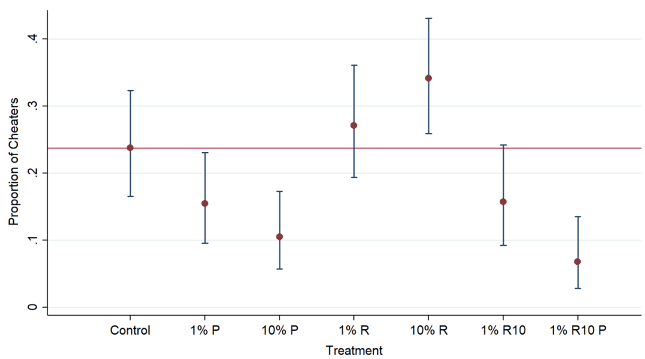

Figure 1. Proportion of Participants who Cheated by Treatment

Note: Error bars represent 95% binomial exact confidence interval. % is audit probability, P indicates punishment for cheating, R indicates reward for honesty, R10 indicates $10 reward. Horizontal line is rate of cheating in control (no audit).

Hypothesis 1

H1: The rate of cheating will be significantly lower when the gap in expected value between cheating and honesty is substantially reduced.

Result: mixed. The rate of cheating was lower with a 10% chance of audit and punishment, but higher with a 10% chance of audit and reward. The rate of cheating was also lower with a 1% chance of a $10 reward, and lowest when a $10 reward was combined with a punishment.

When the chance of punishment was 10%, the rate of cheating was 10.48%, with the difference in rates of cheating between this and control (23.77%) significantly different at the 0.05 level ($\chi^2(1) = 7.67$, $p < 0.01$). The achieved power for a difference between a 10% chance of punishment and control significant at the 0.05 level was 0.84, indicating that this finding was sufficiently powered. This finding is potentially economically significant as well, as the reduction to roughly 40% of the control rate of cheating could be meaningful in many contexts.

However, with a 10% chance of reward, 34.13% of participants cheated, which was significantly different at the 0.10 level from the control ($\chi^2(1) = 3.23$, p = 0.07) but not a 1% chance of reward ($\chi^2(1) = 1.41$, p = 0.24). This was not only not a reduction of cheating, as predicted, but an increase, and similarly a potentially economically meaningful one.

This failure did not carry over to alternative reward structures. With a 1% chance of audit and a $10 reward, 15.69% of participants cheated, which although not significantly below the rate of cheating in the control ($\chi^2(1) = 2.26$, p = 0.13), was in the predicted direction. However, this result was not highly powered and it may be that with a larger sample size, it could be a significant effect.

Lastly, a 1% chance of audit and a $10 reward combined with a punishment led to 6.80% of participants cheating, which was significantly below the control ($\chi^2(1) = 11.97$, $p < 0.01$). This was the lowest observed rate of cheating, which should engender future research on mixed incentive structures.

Overall, 2 of the 4 treatments in which the expected value of cheating was reduced by $0.10 or $0.11 produced rates of cheating significantly lower than control, and a 10% chance of punishment and 1% chance of large reward and punishment both achieved power greater than 0.8. The outlier, a 10% chance of audit and a small reward went significantly in the opposite direction of the incentive and prediction.

Hypothesis 2

H2: Very small changes in the expected value of cheating will not lead to a significant difference in the observed rate of cheating.

Result: mixed. A 1% chance of audit and reward led to an insignificantly higher rate of cheating than control. However, a 1% chance of audit and punishment led to a marginally significantly lower rate of cheating.

With a 1% chance of detection and punishment, the rate of cheating was 15.45%, which was lower than the rate of cheating in control, 23.77%. This decrease was statistically significant at the 0.10 level ($\chi^2(1) = 2.7$, p = 0.10). This result is notable because it suggests that a small audit probability and punishment led to a 60% rate of cheating compared to control. However, larger studies will be necessary to confirm this result with more precision.

With a 1% chance of reward, the rate of cheating (27.12%) was higher than in control, but this was not statistically significant ($\chi^2(1) = 0.65$, p = 0.42).

Hypothesis 3

H3: Equal differences in expected values between honesty and cheating will lead to equal rates of cheating between treatment groups who face punishment and reward incentives.

Result: negative. There was no equivalence in the rates of cheating between punishment and reward treatments with equal expected values of actions, or even between reward treatments with different incentive structures.

The rate of cheating in the 1% reward treatment was significantly higher than in the 1% punishment treatment at the 0.05

level ($\chi^2(1) = 4.92$, p = 0.03). This finding was reasonably powered, with an achieved power for a difference significant at the 0.05 level of 0.60.

More distinctly, the rate of cheating in the 10% reward treatment was higher than in the 10% punishment treatment at the 0.01 level ($\chi^2(1) = 20.10$, $p < 0.01$). This result achieved power for a difference significant at the 0.05 level of 0.99.

The rates of cheating between the 1% chance of a $10 reward and a 10% chance of punishment was not significant ($\chi^2(1) < 0.01$, p = 0.96). However, the rate of cheating with a 1% chance of audit, a $10 reward and punishment, and a 10% chance of punishment was significant ($\chi^2(1) = 4.12$, p= 0.04), but achieved a low power equal to 0.16.

Overall, these results suggest that gaps in expected value were a poor predictor of the effect of auditing systems on cheating behavior. Mirrored treatments (1% and 10% chance of audit and small rewards or punishments) led to particularly different behavior. Implications of this finding are considered further in the discussion.

Discussion

The questions this paper aimed to answer are on what effects rewards and negligible incentives have on cheating. The answer regarding rewards is complex. Small rewards were either ineffective or backfired. However, large rewards showed more promise, as there was suggestive evidence a large reward by itself could reduce cheating, and a large reward combined with a punishment drastically reduced cheating. Punishments, on the other hand, were generally effective, even when they had a negligible effect on the expected values of choices.

While the standard economic perspective of auditing posits that it deters cheating by changing the material incentives, this experiment (and the secondary experiment presented in the supplemental materials) contributes to research that shows auditing influences behavior through other channels. The results showed that an incentives-based model failed in several dimensions to predict behavior: cheating rates changed by magnitudes or directions outside of the bounds of what the model can reasonably accommodate, although large changes in incentives did decrease cheating compared to control in 3 of 4 treatments, significantly so in 2, which is in line with incentive-based predictions.

Whese findings are potentially useful for designing incentive structures to reduce cheating or other misbehaviors, but an explanation fully consistent with the data is not obvious. To investigate one potential mechanism, I conducted a follow-up experiment to test whether changes in perceived social norms could explain the pattern of cheating behavior, based on the work of Krupka and Weber (2013). I found that while social norms did shift according to the treatment, it was not entirely consistent with the cheating behavior (**Supplemental Materials S5**).

This leaves an important question: what drove the observed patterns of cheating behavior? Leading alternative behavioral theories do not seem to fully explain the data. Prospect theory (Kahneman and Tversky, 1979) suggests that loss aversion and the overweighting of small probabilities could explain the lower rate of cheating from a 1% chance of punishment, but it fails to account for the lack of similar effects from a 1% chance of a small reward.

Another potential, but incomplete, explanation is that aspects of the different incentive structures differed in how salient they were to the cheating decision. Punishments may be a particularly salient feature of decisions with an ethical choice, compared to the less common feature of a reward for good behavior. Due to their experiences of being punished for unethical actions, participants may place a higher weight on the negative feelings associated with a punishment than the unfamiliar positive feelings of a reward, and in anticipating these feelings, be more motivated by punishments. This would be consistent with punishments being generally more effective than similarly structured rewards, and punishments combined with large rewards being most effective. Similarly, the reward size may be a more salient feature of the decision-making process than the audit probability (see Laske, Saccardo, and Gneezy 2018 for an example of the punishment size being more salient than the probability). Salience alone however does not seem to conveniently explain why crowding out was observed with small rewards but not large rewards.

Emotional factors, such as fear induced by the potential for punishment, may play a role, as suggested by the ”risk-as-feelings” hypothesis. When punishment is anticipated, fear or anxiety might drive behavior away from maximizing expected utility, leading to alternative decision-making processes. This theory has some direct empirical support from laboratory cheating studies, which have found that emotional arousal and stress markers correlate with misbehavior (Coricelli et al., 2010; Dulleck et al. 2016). In this framework, fear not only influences anticipated outcomes but also alters decision-making mechanisms, leading individuals to avoid fear-inducing outcomes, regardless of their actual probability. This explains the observed insensitivity to differences between a 1% and 10% chance of punishment. However, this dual-process explanation is complex and requires further evidence. Future research could explore this mechanism using physiological tools like functional magnetic resonance imaging to pinpoint differences in brain activity, or by testing whether emotional models outperform expected utility models in predicting behavior under varying punishment types or emotionally charged environments. Such studies could illuminate the limits of standard economic reasoning and provide deeper insights into the psychological underpinnings of cheating behavior.

This study builds upon work by Andreoni et al. (2003) on a proposer-responder game in which responders, depending on treatment, could punish or reward the proposer. Andreoni et al. (2003) found that the combination of possible rewards and punishments was most effective in inducing larger offers, and that relative to a dictator game the possibility of punishments alone induced larger offers, and the possibility of rewards alone was ineffective. This work extends upon that of Andreoni et al. (2003) first by demonstrating an additional domain in which a combination of rewards and punishments produced the largest behavioral effect. Second, incentive probabilities in this study were fixed and known, making it possible to examine the relative effects of the size and probability of incentives.

There is a need for more laboratory experiments to better understand the dynamics of rewards, which can both aid in theory development as well as provide more pragmatic guidance for field implementation. The straightforward systematic laboratory exploration of combinations of rewards and punishments might provide more evidence useful to developing theory, as well as cost-effective practical guidance for field implementation. Using a non-monetary reward (which previous experiments have suggested may not induce crowding out) or an experimental paradigm with a greater ethical component (e.g. the deception game (Gneezy, 2005)) could also disentangle the motivation of participants. Similarly, using an unobservable cheating game such as the die-roll task (Fischbacher and Föllmi-Heusi, 2013) could rule out effects from participants knowing their actions could be observed, even if they believe the stated auditing probability and consequences. Examining rewards in a repeated game context may be particularly useful in determining the long-term efficacy of such incentives before field implementation. Lastly, the importance of emotions (e.g. fear) is not sufficiently understood in auditing and may provide an explanation for the results of these experiments.

One motivation for this paper was that populations may prefer or accept incentive systems that offer the possibility of a reward where they may dislike or reject systems that only offer punishment for misbehavior. The acceptance of reward-based systems could open up new areas for policymakers to change behavior where punishment-based auditing systems are not viable because they are unpopular. Conjecturally, populations might dislike oversight mechanisms that threaten them with negative outcomes, even if they support the behavior changes such systems are meant to induce. Future experimental research should examine whether it is actually the case that reward systems are more palatable, for example by allowing participants to choose which incentive structure they, or other participants, should face.

Testing rewards in field settings is ideal for evaluating another key motivating feature: whether incentives to reduce misbehavior are cost-effective. Promising environments include fare evasion on public transport, rewarding those who can prove they have purchased a ticket or rewarding safe driving behavior. Several studies have found rewards can be effective when combined with continuous monitoring equipment in reducing bad driving (Mazureck and van Hattem, 2006; Lahrmann et al., 2012). An extension of this would be to incentivize good driving when continuous monitoring is not feasible. For instance, police who observe good driving could send rewards to drivers. In these field contexts, it is possible to estimate the value of reduced misbehavior (e.g. foregone car accidents) and compare it to the expenditure on rewards.

Experimental methods are an important tool in developing a better understanding of how to structure incentives in order to reduce cheating. This study provides novel contributions to the literature on cheating by providing evidence that negligible punishments may be effective as a deterrent, that small rewards may backfire, and that rewards can be effective in reducing cheating when large and combined with punishments.

Declarations

Funding

Funding for this research was provided by the University of California San Diego

Conflicts of interest/Competing interests

The author declares no conflict of interests.

Availability of data and material

All experimental materials and data are available upon request.

Code availability

All code is available upon request.

References

Almås, Ingvild, Alexander W. Cappelen, and Bertil Tungodden (2016). Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians?

Andreoni, James, William Harbaugh, and Lise Vesterlund (2003). The carrot or the stick: Rewards, punishments, and cooperation. *American Economic Review*, 93(3): 893-902.

Becker, Gary S. (1968). *"Crime and punishment: An economic approach." The economic dimensions of crime*. Palgrave Macmillan UK, pp. 13-68.

Brink, Alisa G., and Frederick W. Rankin (2013). The effects of risk preference and loss aversion on individual behavior under bonus, penalty, and combined contract frames. *Behavioral Research in Accounting*, 25(2): 145-170.

Coricelli, Giorgio, et al. (2010). Cheating, emotions, and rationality: an experiment on tax evasion. *Experimental Economics*, 13(2): 226-247.

Dulleck, Uwe, et al. (2016). Tax compliance and psychic costs: behavioral experimental evidence using a physiological marker. *Journal of Public Economics*, 134: 9-18.

Fischbacher, Urs, and Franziska Föllmi-Heusi (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3): 525-547.

Gneezy, Uri. The W effect of incentives (2003). *University of Chicago working paper*.

Gneezy, Uri (2005). Deception: The role of consequences. *The American Economic Review*, 95(1): 384-394.

Internal Revenue Service (2019). *Data Book, 2018.* Publication 55B.

Kahneman, Daniel, and Amos Tversky (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2): 263-292.

Krupka, Erin L., and Roberto A. Weber (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3): 495-524.

Lahrmann, Harry, Niels Agerholm, Nerius Tradisauskas, Kasper K. Berthelsen, and Lisbeth Harms (2012). Pay as You Speed, ISA with incentives for not speeding: Results and interpretation of speed data. *Accident Analysis & Prevention*, 48: 17-28.

Laske, Katharina and Saccardo, Silvia and Gneezy, Uri (2018). Do Fines Deter Unethical Behavior? The Effect of Systematically Varying the Size and Probability of Punishment (April 5, 2018). Available at SSRN: https://ssrn.com/abstract=3157387 or http://dx.doi.org/10.2139/ssrn.3157387

Lowenstein, George F., et al. (2001). Risk as feelings. *Psychological bulletin*, 127(2): 267-286.

Mazureck, Undine, and Jan van Hattem (2006). Rewards for safe driving behavior: Influence on following distance and speed. *Transportation research record 1980*, no. 1: 31-38.

Prendergast, Daniel and Donohue, Pete (2011). Subway turnstile jumpers saving loads of money, even when they are caught, study shows. *New York Daily News.*