Deterring Cheating in Online Environments

HENRY CORRIGAN-GIBBS, Stanford University
NAKULL GUPTA, Microsoft Research India
CURTIS NORTHCUTT, MIT EECS
EDWARD CUTRELL, Microsoft Research India
WILLIAM THIES, Microsoft Research India

Many Internet services depend on the integrity of their users, even when these users have strong incentives to behave dishonestly. Drawing on experiments in two different online contexts, this study measures the prevalence of cheating and evaluates two different methods for deterring it. Our first experiment investigates cheating behavior in a pair of online exams spanning 632 students in India. Our second experiment examines dishonest behavior on Mechanical Turk through an online task with 2,378 total participants. Using direct measurements that are not dependent on self-reports, we detect significant rates of cheating in both environments. We confirm that honor codes–despite frequent use in massive open online courses (MOOCs)–lead to only a small and insignificant reduction in online cheating behaviors. To overcome these challenges, we propose a new intervention: a stern warning that spells out the potential consequences of cheating. We show that the warning leads to a significant (about twofold) reduction in cheating, consistent across experiments. We also characterize the demographic correlates of cheating on Mechanical Turk. Our findings advance the understanding of cheating in online environments, and suggest that replacing traditional honor codes with warnings could be a simple and effective way to deter cheating in online courses and online labor marketplaces.

Categories and Subject Descriptors: K.3 [**Computers and Education**]; J.4 [**Social and Behavioral Sciences**]: Psychology

General Terms: Experimentation, Human Factors

Additional Key Words and Phrases: Cheating, honor code, massive open online course (MOOC), Mechanical Turk, honey pot, warning, crowdsourcing

ACM Reference Format:
Henry Corrigan-Gibbs, Nakull Gupta, Curtis Northcutt, Edward Cutrell, and William Thies. 2015. Deterring cheating in online environments. ACM Trans. Comput.-Hum. Interact. 22, 6, Article 28 (September 2015), 23 pages.
DOI: http://dx.doi.org/10.1145/2810239

1. INTRODUCTION

As everyday services migrate into online environments, it will be crucial to preserve the same levels of trust, honesty, and integrity online that people expect from

An NSF Graduate Research Fellowship under Grant No. DGE-114747, an NSDEG Fellowship, and the Catalyst Center for Sustainable Development at Microsoft Research India provided financial support for this work.
Author's addresses: H. Corrigan-Gibbs, Department of Computer Science, Stanford University, Stanford, CA 94305 USA; N. Gupta, E. Cutrell, and W. Thies, Microsoft Research India, "Vigyan," #9, Lavelle Road, Bangalore 560 001, India; C. Northcutt, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

DOI: http://dx.doi.org/10.1145/2810239

face-to-face interactions. For example, the increasing enrollment in online courses means that deterring cheating in such courses may quickly become as important as deterring cheating in brick-and-mortar classrooms. And as paid crowdsourcing platforms such as Amazon Mechanical Turk come to displace traditional colocated workplaces, it is important to ensure honest work and rewards, even when the identities of workers and employers are not known to each other.

When it comes to in-person interactions, societal institutions often appeal to honor and morality as a means to encourage rule-following behavior and to discourage dishonest conduct. In courtrooms, witnesses must "solemnly swear to tell the truth, the whole truth, and nothing but the truth" before giving legal testimony [Federal Rules of Evidence 1975]. In college, students must sign "honor statements" asserting that they did not cheat on their homework assignments before submitting them [Dirmeyer and Cartwright 2012; Hoover 2002]. In US medical schools, aspiring physicians must take an oath to practice their craft responsibly before receiving a diploma [Markel 2004]. This type of honor-focused appeal can be effective in physical encounters: study participants who sign honor codes are more likely to evaluate themselves honestly in self-graded tasks [Mazar et al. 2008; Shu et al. 2011a] and are more likely to answer questions honestly, even when given the opportunity to cheat [Jacquemet et al. 2013; Shu et al. 2011b].

There is a relative dearth of evidence, however, about how to promote honest behaviors in online environments. Techniques for preventing cheating in face-to-face interactions appear to be less effective online. For example, one study found that requiring participants to agree to an honor code had no effect on the rate of cheating in an incentivized online task [LoSchiavo and Shatz 2011] and other work comes to similar conclusions [Mastin et al. 2009]. Nonetheless, perhaps for a lack of better options, major open online education platforms are still relying on honor codes as the primary deterrent to cheating [Coursera 2014; edX 2014].

This paper describes the results of two experiments we conducted to better understand and deter cheating in online environments. The first experiment took place in the context of a pair of online exams, spanning 632 students in India. The second experiment studied the Mechanical Turk crowdsourcing platform and involved 2,378 workers in India and the United States.

In both contexts, we confirm that there are significant levels of cheating and that honor codes have little impact on rates of dishonesty. To detect cheating, we inspected short-answer responses for plagiarism and we deployed a "honey pot" (akin to those used in computer security [Provos 2004]) to measure how many participants surfed the web during each task, in violation of the instructions. Our baseline measurements are most novel in the case of online exams, in which we observed cheating by 26–34% of students. While the use of honor codes resulted in a small reduction in cheating in both experiments, the effects were not statistically significant.

In order to deter cheating more effectively, we propose and evaluate a new technique: a pre-task warning that highlights the potential consequences of being caught cheating. We ground our use of warnings in well-studied theories of misbehavior, and we discuss how the effects of honor codes and warnings differ in physical and online settings. In our experiments, the warning led to a significant (about twofold) reduction in the rate of cheating. The impact of the warning was consistent across all contexts and geographies studied, that is, the two online examinations and the Mechanical Turk experiments in India and the United States.

In addition, we use our experimental data to characterize the demographic correlates of cheating on Mechanical Turk. While prior work has investigated this question [Suri et al. 2011], our experiment offers more statistical power and reveals new and

significant trends. For example, we find that age correlates negatively with the likelihood of cheating and that the number of hours spent working on Mechanical Turk correlates positively with the likelihood of cheating. We discuss each of these phenomena in more detail.

Taken together, our results contribute to understanding cheating in online environments and suggest new measures to help control it. Our most specific recommendation is that a simple warning may be more effective at deterring cheating in certain contexts than a traditional honor code. In Section 7, we discuss the real-world implications of this finding and outline areas for future work. We conclude that with careful application of appropriate warnings, administrators of online courses and online labor marketplaces may be able to promote honesty more effectively and at negligible cost.

2. RELATED WORK

A series of studies has investigated how the design of a task affects participants' tendency to cheat in the course of completing it. In the work that is most relevant to our own, Mazar et al. demonstrated that having students recall the Ten Commandments or sign an honor code before completing a self-graded task (albeit an artificial one) reduced the rate of dishonesty evidenced [2008]. Follow-up work showed that making subtle modifications to the task could also *increase* the rates of cheating. In particular, participants who were more tired [Mead et al. 2009], who were in darkened rooms [Zhong et al. 2010], who were primed with a text about philosophical determinism [Vohs and Schooler 2008], who had less time to contemplate their actions [Shalvi et al. 2012], who had exercised some self-control before the task [Gino et al. 2011], or who felt like they had been treated unfairly [Houser et al. 2012] cheated at increased rates.

A parallel line of work has explored whether the results of Mazar et al. demonstrating the effectiveness of an honor code also hold in real-world settings. The findings are mixed: one experiment showed that an appeal to honor *was* effective at decreasing the rate of theft at a newspaper box [Pruckner and Sausgruber 2013], while another found that an honor-focused appeal *was not* an effective way to get tax evaders to pay a TV license fee by mail [Fellner et al. 2013].

Rosenbaum et al. survey these and other empirical studies from the economics and psychology literature on the topic of honesty [2014]. Synthesizing the results of 63 studies, the authors conclude that there "appear[s] to be a consistent proportion of unconditional cheaters and noncheaters [. . .], with the honesty of the remaining individuals being susceptible to a range of variables, most notably monitoring and intrinsic lying costs" [Rosenbaum et al. 2014, p. 194]. This hypothesis, which is worthy of further study, may explain why different study populations respond so differently to moral appeals not to cheat. For example, the lack of an observed effect in the Fellner et al. study mentioned previously may have been due to the fact that the proportion of "unconditional cheaters" is higher in a group of tax evaders than it is in the population at large (an issue that Fellner et al. raise in their analysis) [2013].

The computer science literature has typically taken a different approach to deterring cheating: rather than focusing on psychological techniques, computer scientists have focused on *algorithmic* approaches to detect cheating and filter out dishonestly generated or incorrect submissions. For example, the archetypical systems for crowdsourced work in the human–computer interaction (HCI) research corpus use IP address tracking in concert with task-specific heuristics to detect misbehavior [von Ahn and Dabbish 2004; von Ahn et al. 2006] and other crowdsourcing systems have adopted similar task-specific techniques to detect cheating [Walsh and Golbeck 2010]. More recent work uses a combination of cameras and body-mounted sensors to automatically detect cheating during a computerized exam [Li et al. 2015]. In this study, we adopt

techniques from both the psychology and HCI literature: we use psychological methods to deter cheating and use algorithmic methods to detect cheating.

Prior work investigated the quantitative and qualitative effects of honor systems in brick-and-mortar universities [McCabe and Trevino 1993; McCabe et al. 1999] and in online settings [LoSchiavo and Shatz 2011; Mastin et al. 2009]. Overt use of plagiarism detection software in the classroom, one study finds, may deter cheating [Braumoeller and Gaines 2001]. It is worth noting that this prior work on the effectiveness of honor codes and warnings in physical settings may not generalize immediately to online settings. In particular, a number of studies have found that participants are less honest when interacting via a computer than when they interact face-to-face [Rockmann and Northcraft 2008; Van Zant and Kray 2014]. Taking a class online may thus diminish the effectiveness of an honor code.

Our work complements the existing work on honor codes by investigating the effects of *warnings* in addition to those of honor codes. We also introduce a novel methodology for directly measuring cheating rates in online exams. In contrast, the bulk of honor-code studies in the prior work measure dishonesty after the fact by asking participants to anonymously report whether they cheated. As a result, with these studies it is difficult to determine whether an intervention (e.g., an honor code) decreased the rate of misbehavior or simply decreased the rate at which participants were *willing to admit* to having cheated.

Our second experiment explores the effect of an honor code and a warning on the behavior of workers on Amazon Mechanical Turk, an online labor platform. Related work has gathered demographic data on Mechanical Turk workers [Paolacci and Chandler 2014; Ross et al. 2010] and has discussed the challenges of using Mechanical Turk for user studies, psychology experiments, and for crowd work generally [Horton et al. 2011; Kittur et al. 2008, 2013]. Others used qualifying exams and "gold standard" tasks to detect careless workers [Downs et al. 2010; Kittur et al. 2008; Oleson et al. 2011; Gadiraju et al. 2015] and surveyed techniques for improving data quality on the platform [Difallah et al. 2012; Eickhoff and de Vries 2013].

Suri et al. measured baseline rates of dishonesty on Mechanical Turk and found levels of dishonesty consistent with levels measured in a laboratory setting [Fischbacher and Föllmi-Heusi 2013; Suri et al. 2011]. Goodman et al. found that asking Mechanical Turk workers not to search for answers to survey questions reduced the prevalence of this behavior [2013]. However, to the best of our knowledge, prior work has not considered whether targeted pretask messages (like the honor code and warning we use) affect cheating rates in the context of Mechanical Turk.

Our experiments within the Mechanical Turk environment loosely extend the experimental framework of Mazar et al. to online environments [2008]. In the Mazar et al. study, the experimenter gave participants a computational task to be completed within 5min and gave a monetary reward proportional to the number of correct answers. The researchers graded one group of participants and the other group graded themselves. The self-graded participants scored higher on average, which suggests that dishonesty was prevalent. Our Mechanical Turk experiment similarly measures the effects of dishonesty on an incentivized task. Unlike the prior work, we measure cheating on a per-user basis rather than relying on per-treatment-group aggregate statistics.

We detect cheating in part using a "honey pot" website. Honey pots are a celebrated technique in the computer security literature and have been used to defeat spam [Andreolini et al. 2005], to detect system vulnerabilities [Provos 2004], to find malware [Baecher et al. 2006], and even to challenge "419 Scammers" [Crampton 2007]. The use of a honey pot to detect cheating in an online exam is a novel aspect of this study.

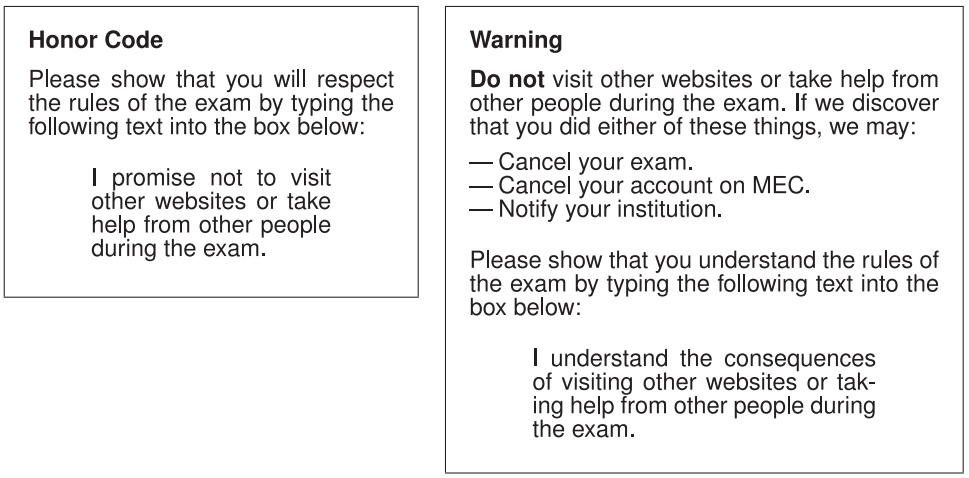

Fig. 1. Honor code (left) and warning (right) used in the online exam experiment.

3. RESEARCH QUESTIONS

Our work was guided by the following high-level research questions.

(1) Can we detect and measure cheating in different online contexts?
(2) Can the rate of cheating be reduced by usage of an honor code or warning?
(3) Is the rate of cheating correlated with demographic variables?

As an initial inquiry into these questions, we designed two separate experiments: one focusing on an online examination, and one focusing on Mechanical Turk. These different contexts shed light on different aspects of the research questions.

The first research question (measurement of cheating) is especially relevant for online courses, in which there is little prior data on the rate of cheating. A rigorous measurement of cheating on online exams could help course administrators to understand the risks and benefits of such evaluations and could underscore the importance of taking steps to reduce cheating. In the case of Mechanical Turk, the amount of cheating is likely to depend closely on the incentives offered, and thus we place less emphasis on the absolute rates of cheating observed. Nonetheless, the ability to construct a task that leads to consistent and measurable cheating on Turk is an important research tool that enables us to evaluate the efficacy of different cheating-deterrence strategies.

The second research question (impact of honor codes and warnings) is pertinent to both online education and Mechanical Turk. Since users in India and the United States make up a plurality of students in many online courses [Guo and Reinecke 2014] and comprise the bulk of workers on Mechanical Turk [Ross et al. 2010], we are also interested in if and how honor codes and warnings affect users in these two countries differently.

Our design of the honor code and warning messages is an important aspect of the experiments. In both experiments, the honor code followed the style of one-sentence honor codes used in US universities, although we modified the text slightly to directly address the sort of cheating behavior we wanted to discourage. For example, the honor code for the online examination (Figure 1) required participants to explicitly promise not to visit other websites while taking the examination. In contrast, the warning message listed three negative consequences that may befall participants who cheated. For example, the warning code used in the online exam (Figure 1) explained that

participants caught cheating may have their online course accounts suspended. We specifically chose the consequences to be ones that we could realistically enforce and that would be proportionate to the ethical violation entailed by cheating on the task. To ensure that participants actually read the pretask messages, we required participants to type out a one-sentence summary of the honor code or warning before proceeding.

The third research question (demographic correlates of cheating) is of broad interest across platforms; however, our data allow us to investigate it only in the context of Turk, in which we observed more diversity in users' age, education, gender, income, nationality, and other factors. Finding strong correlations between a particular demographic feature and tendency to cheat on online tasks could raise new research questions for further investigation and might provide the basis for future qualitative research into the drivers of cheating in online contexts.

4. THEORETICAL RATIONALE

Existing models of physical-world cheating underpin our proposed use of warnings as a means to deter misbehavior in online contexts. Mazar et al. theorize that there are two competing processes that influence a person's decision to cheat [2008].

—The first process consists of a rational cost-benefit analysis, in which the person evaluates the economic risks and rewards of cheating and the likelihood of getting caught. The more economically profitable cheating appears, the more likely one is to cheat.
—The second process is one in which the person evaluates the effect that cheating will have on their own self-image or "self-concept." If behaving dishonestly in a particular situation is clearly out of line with one's values (e.g., stealing $1 from a friend's wallet), this is more likely to harm one's self-concept than is behaving dishonestly in a more subtle way (e.g., taking a pen from a friend's house) [Mazar et al. 2008]. The more consistent cheating is with the person's self-concept, Mazar et al. theorize, the more likely they are to cheat.

The theory suggests that the final decision of whether to cheat is a function of the outcomes of these two competing processes—determining whether the economic incentives to cheat outweigh the detrimental effects of cheating to one's self-concept. Mazar et al. "suggest that people typically solve this motivational dilemma adaptively by finding a balance or equilibrium between the two motivating forces."

Traditional honor codes target the second of these two process: by reminding users of their ethical standards, it becomes more difficult to cheat while maintaining a positive self-concept. Thus, the theory predicts that participants subjected to an honor code are less willing to cheat than they would otherwise be. A rich set of studies conducted in in-person settings provide experimental support for this view [Jacquemet et al. 2013; Shu et al. 2011b; Mazar et al. 2008; Shu et al. 2011a]. In essence, once a study participant has put their commitment not to cheat on paper, it is more difficult to violate that promise while still maintaining a positive self-concept. By making cheating more harmful to one's self-concept, an honor code makes cheating less attractive for many users.

We conjecture that this logic does not translate directly to online settings. Indeed, while many people might feel that it is unethical to violate a promise written out on paper, the average Internet user may have no ethical problem with violating a promise spelled out on a website. That is, it may be less harmful to one's self-concept to violate a "digital" honor code, signed on a computer, than it is to violate an in-person honor code, signed on paper in a classroom. The common experience with online "terms of service" gives credence to this conjecture: one study found that only 5% of participants reported reading terms of service when downloading software online [Stark and Choplin 2009]

and the prevalence of software and media piracy indicates that precious few users actually heed these terms. If our conjecture holds, it would be more compatible with the average Internet user's self-concept to violate an honor code signed online than it would be to violate an honor code signed in person. This would, in turn, lead the honor code to have a weaker effect online.

Furthermore, we reason that Internet users are still sensitive to cost-benefit arguments (the first of Mazar et al.'s two processes previously), even in an online context. That is, if an online service makes a credible threat of imposing a cost on misbehaving Internet users, users will incorporate that cost into their economic reasoning. If the cost is high enough and the threat is credible enough (i.e., the probability of being caught is high), we expect that users will refrain from cheating for purely economic reasons. This reasoning leads us to expect that a stern warning will be effective in an online setting, even when an honor code is not. With a warning, we attempt to shift the users' cost-benefit calculus, by invoking the possibility of punishment and thus making cheating less attractive.

5. EXPERIMENT I: ONLINE EXAM

Our first experiment investigated the effect of an honor code and a warning message on students taking an online exam in India. Though honor codes have been explored widely in brick-and-mortar classroom scenarios, until now their ability to promote honesty in online courses has been largely untested.

5.1. Methods

Background. In the spring of 2014, and again in the winter of 2015, Microsoft Research India offered a free online course on "The Design and Analysis of Algorithms." Unlike other massive open online courses (MOOCs), the course targeted undergraduate engineering students in India who were taking the same subject in their local college. Students who completed the online lectures and activities were given the option of taking a final exam to qualify for a certificate. As an incentive to participate, the top-scoring students were interviewed for an internship position at Microsoft Research India.

We conducted our study in the context of the online final examination for this algorithms course. We ran the study twice: once in August 2014[1] (after the spring session of the course) and once in February 2015 (after the winter session). We also offered a proctored in-person version of the August exam in five different cities, for students who lived near urban centers.

As cheating is a serious threat to the viability of online examinations, we used this opportunity to explore different potential interventions to reduce cheating on the online exam. Similar to other MOOCs, we also offered slightly different certificates to offline and online test takers. Since the student body was not very demographically diverse—most students were in their second or third year of an engineering college—we defer analysis of demographic effects to the next experiment.

Design of Exam. The exam consisted of fifteen multiple-choice questions and one free-response question. (Due to an ambiguity in one question, we graded only 14 of the multiple-choice questions on the August 2014 exam.) All questions were original and answering them required critical thinking—students would likely not benefit from having access to third-party reference materials during the exam. As an added barrier to plagiarism, we created 15 different versions of the exam; each student received one at random. We randomized the order of a subset of the questions and answers in each version of the exam. Though we anticipated that the exam could be completed in 1hour,

[1]We published the results of the August 2014 trial as a short paper [Corrigan-Gibbs et al. 2015].

we kept the exam open for 2.5hours to allow for scheduling conflicts, interruptions, or technical difficulties.

We required that students taking the exam: (1) did not consult other materials (books, notes or other websites) while taking the test, and (2) neither gave nor received aid from other people during the exam. To explore the effect of honor codes and warnings on students' compliance to these rules, we randomly assigned all students taking the online exam to one of the following three conditions.

—*No Additional Instructions*. This was identical to exams provided to students in the proctored setting.
—*Honor Code*. For this condition, students were asked to read and type out an honor code at the beginning of the exam (Figure 1, left). This code was designed to appeal to their sense of integrity.
—*Warning*. For this condition, students were asked to read and type out a warning statement at the beginning of the exam (Figure 1, right). This statement emphasized the negative consequences of breaking the rules by cheating.

To ensure that students in the latter two conditions actually read the honor code or warning, we asked students to type a one-sentence version of the honor code or warning into a text box. We displayed the relevant text as an embedded image to prevent students from copying-and-pasting the sentence.

Measuring Prevalence of Cheating. One approach to measuring cheating would be to compare scores on the proctored in-person exam to scores on the online exam. Under the assumption that the rate of cheating on the in-person exam was negligible, this would give us some indication of the benefit reaped by taking the exam online, in which it would be easy violate the rules of the exam (e.g., by searching for answers online). Unfortunately, this comparison is fraught with several uncontrolled variables. For example, the population of students taking the online exam was likely quite different from the population taking the proctored exam, and this difference could lead to a difference in scores between the proctored and online versions of the test. For this reason, we focus only on the online exam.

We employed two techniques to measure the prevalence of cheating.

(1) *Examination of "Free Response" Questions*. The free response question was one of the harder questions on the exam. It required students to design a graph algorithm and describe it informally in a few sentences (our solution was 38-words long). We checked these responses for plagiarism using two heuristics.

First, to catch students who copied their answer directly from an Internet website, we identified responses with idiosyncratic language or symbols and entered them into an Internet search engine. Responses that exactly matched the text on a returned page were labeled as cheaters.

Second, to identify students who copied their answers from each other, we performed manual comparison of responses. We ranked all response pairs by longest common substring, and also by longest common subsequence, and inspected pairs with high scores. In addition, we sorted all responses alphabetically, and also by total length, and examined adjacent entries. Three of the authors, all blinded to treatment condition, examined similar answers and reached a consensus as to whether cheating had likely occurred. For each pair that evidenced cheating, we labeled both students as cheaters.

At a later date, we sought to measure the inter-rater reliability of our plagiarism judgements (something that we did not record originally). To assess this, two of the authors independently relabeled each response as "honest" or "cheating" according to the criteria outlined previously. This exercise took place ten months after the

initial inspection of the exams, when neither author could recall the labels that were previously assigned by the group. The two graders agreed with each other on 95.0% of the 603 responses. To assess inter-rater reliability, we used Cohen's Kappa statistic, which takes into account the fact that some agreement between graders may occur by chance [Cohen 1960]. We found $\kappa = 0.83$, which corresponds to "almost perfect or perfect agreement" [Hallgren 2012]. This exercise demonstrates that there was some noise in our grading procedure, but that the classification of exams was consistent on the whole.

(2) *"Honey Pot" Website*. To check if students consulted the web to seek help during the exam, we placed all of the exam questions on a public-facing website (the "honey pot") that was indexed by Google. If students searched for the exact text of any exam question, our website was the first hit returned. The website did not include the answers to the questions, but it did include a button ("Click to show answer") for each question on the test; when that button was pressed, the website paused and simulated a timeout (without showing the answer). We instrumented the honey pot website to check for a cookie set by our online exam platform.

The cookie allowed us to recover the unique identifier of students who tried to download the exam answers from the honey pot page. We also used a tracking `iframe` from our learning platform on the honey pot page as a fall-back tracking mechanism when we were not able to set a third-party cookie. We could not detect students who used a second device (e.g., a mobile phone) to access the honey pot site, so the rates of cheating reported reflect lower bounds on the true rates.

We filtered hits to the honey-pot website to exclude hits occurring after the end of the exam. Honest-but-curious students may have looked for exam answers online after the conclusion of the test, so we were careful not to label these students as cheaters.

Statistical Techniques. To evaluate the effect of the treatments (honor code and warning) in each of our studies, we first constructed a 2×2 contingency table for the treatment condition in question. Each of the four cells of the table contained the number of students who fell into one of the four condition–outcome pairs, in which the conditions were "Control" or "Treatment," and the outcomes were either "Honest" or "Cheated." Thus, the first cell contained the number of students in the control group who were honest, the second cell contained the number of students in the control group who cheated, and so on. We then applied Pearson's chi-squared (χ^2) test of independence to the table as our primary statistical test.

We report the results of all of our chi-squared tests as adjusted with the Yates continuity correction. The Yates correction reduces the rate of Type I errors (false positives) that occur when using the chi-squared tests with small sample sizes. Applying the Yates correction is standard for the sort of 2×2 contingency table analysis we perform and is the default for many statistical packages, including `R`. For details on the history of the Yates correction and related statistical issues, consult Richardson [1994].

Unless otherwise noted, we present p-values for the chi-squared tests as adjusted using the Bonferroni–Holm correction to account for multiple hypotheses testing [Holm 1979]. The Bonferroni–Holm correction adjusts for the fact that when a single experiment tests many hypotheses at once, it is increasingly likely that some individual test statistics will be outliers, even when there is not an underlying effect. The Bonferroni–Holm correction reduces the probability of such a "false discovery" and is thus a conservative practice.

5.2. Ethical Considerations

Prior to the online exam, we informed all students that we would be conducting a research study as part of the test. Although we did not explain the exact research

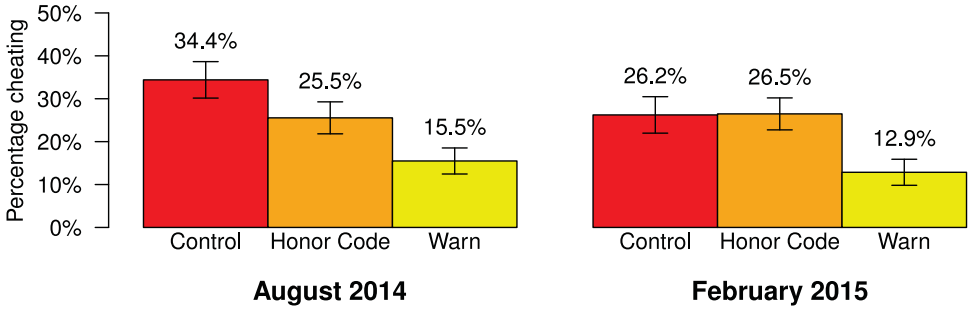

Fig. 2. Rates of cheating in the two online exams. Error bars indicate the standard error of the proportion.

hypothesis (to avoid biasing the results), we did say that we would be testing some features of the exam platform that might improve fairness or reliability of online exams. *We gave all students the opportunity to opt out of these experiments*, in which case they received a default version of the exam (following other MOOCs, the default version used the honor code). When we sent the final test scores to the students, we also debriefed them regarding the purpose of the experiment and the high-level approach we used to monitor potential cheating behavior during the exam. *We did not take any disciplinary actions against students who we suspected had cheated on the exam.*

When grading the exam, we wanted to be fair to honest students across all experimental conditions. We set the "pass/fail" threshold to a relatively low score that was informed by student performance on the proctored exam (in which cheating was very difficult). Since the pass threshold was held constant across conditions, our experiment may have offered different advantages to different cheaters, if their cheating behavior depended on the honor code or warning received. Even so, the benefits (and costs) of honesty and cheating on our exam were likely very close to the benefits and costs of these behaviors on a typical university examination.

Institutional Review Boards at Microsoft Research and Stanford University reviewed and approved our experimental protocol.

5.3. Results: August 2014 Exam

Participants. There were 409 students who took the online exam in August 2014 (and 674 students who took a proctored exam). We exclude from our analysis two students who opted out of the experiment and three students who did not advance past the exam's instructions page. Thus, we analyze data for 404 students.

Effect of Honor Code and Warning. Overall, we classified 24% of students taking the exam as cheaters. The breakdown of cheaters by experimental condition is shown in Figure 2. Cheating was highest in the baseline condition (34% of students), followed by the honor code (25% of students), followed by the warning (15% of students).

The difference between the baseline condition and the warning was significant ($\chi^2(1, N = 267) = 11.9$, $p < 0.001$). The difference between the honor code and the warning has borderline significance without correction ($\chi^2(1, N = 279) = 3.74$, without Bonferroni–Holm correction $p = 0.053$) but is not significant with correction ($p = 0.11$).

Cheating Behavior. We labeled students as cheaters if they either submitted a plagiarized response to our free response question or if they visited our honey pot website. In total, we classified 100 students as cheaters. Of these, 84% plagiarized on the free response question, 18% visited the honey pot, and 2% plagiarized *and* visited the honey pot. Of the 23 unique IP addresses that visited the honey pot site during the exam, there were 5 that we could not associate with any student. These may represent

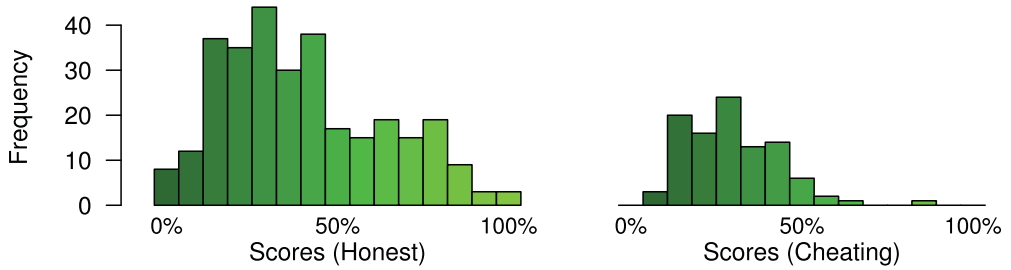

Fig. 3. Histogram showing the number of correct exam answers for honest and cheating participants in the online exam.

visitors who were not taking the exam, or may correspond to additional devices that exam participants used to access the honey pot but not to access the exam site.

Among the plagiarized responses, most (80%) showed similarity to another student's response, 42% showed similarity to an Internet website, and 21% fell into both categories. The vast majority of repeated answers were submitted by exactly two students, though six responses were submitted by 3–5 students and one answer (copied from Wikipedia) had overlap between eight students. Often the responses were identical, though sometimes they were reworded slightly, suggesting that students may have intentionally tried to avoid detection by a plagiarism check. For example, one student wrote "find the path though all safe edges using dijikshraw's algorithm" while another wrote "find the path using safe edges only by dijikshraw's algorithm." In addition to similarity in sentence structure, both of these responses had a unique misspelling of "Dijkstra's algorithm."

The set of online resources copied by students included Wikipedia, a tutorial, a course book, and peer-reviewed publications. None of the plagiarized responses were correct, and many were nonsensical. For example, two students submitted a response copied from a paper published in the *Wilson Journal of Ornithology*. The response begins, "edges are often associated with a high risk of brood parasitism by Brown-headed Cowbirds." We later discovered that this paper is the top search hit for "risky edges," a phrase that appeared in the question.

Figure 3 illustrates the distribution of exam scores for cheaters and noncheaters. Despite their efforts to cheat, the average score for cheaters (30% correct, $SD = 14\%$) was lower than the average score for non-cheaters (40% correct, $SD = 23\%$). This difference is statistically significant ($t(402) = 4.18$, $p < 0.0001$). We conjecture that the weaker students felt more inclined to cheat, but due to the questions (and random variations) on the exam, cheating did not offer large benefits to their scores.

5.4. Results: February 2015 Exam

Participants. There were 223 students who took the February 2015 version of the examination. One student opted out of the study and 23 students did not advance past the instructions page, so we analyze results for 199 students.

Effect of Honor Code and Warning. Unfortunately, since only 199 students took the February examination, compared with 404 in August, our tests have less statistical power at their disposal. The warning did not have a statistically significant effect at the $p < 0.05$ level but it was statistically significant at the $p < 0.1$ level ($\chi^2(1, N = 131) = 3.24$, $p = 0.0854$). Students in the control condition cheated at a rate of 26.2% and students in the warning condition cheated at a rate of 12.9% (Figure 2). Consistent with the prior exam, we did not find that the honor code had any effect on the rate of cheating ($\chi^2(1, N = 129) = 1$, $p = 1$).

The baseline rate of cheating on the February exam was much lower than the rate of cheating on the earlier exam (26.2% versus 34.4%). One factor that may have contributed to this is that the February exam was more difficult, causing fewer people to consider and attempt the free response question (which appeared last on the test). Because more students left the question blank (21% of test takers on the February exam, versus 9% of test takers on the earlier exam), there were fewer responses that we could evaluate for plagiarism. A second potential contributor to the difference in baseline cheating is that we offered an in-person version of the August 2014 exam but did not offer an in-person version of the February 2015 exam. We conjecture that some of the students who had no plans to cheat on the exam (i.e., the well-prepared ones) may have opted to take the in-person exam in August, and thus the fraction of dishonest students online could have been higher. Since there was not an in-person exam in February, more well-prepared students (who would have taken the exam in person) took it online, reducing the baseline rate of cheating.

Cheating Behavior. Out of the 43 students we classified as cheating, we detected that 30 students visited the honey pot, 16 submitted plagiarized responses, and 3 students both submitted plagiarized responses and visited the honey pot site. The nature of plagiarized responses and the distribution of exam scores was similar in the two exams, so we omit these details for the February 2015 exam.

6. EXPERIMENT II: MECHANICAL TURK

Our second experiment investigated the impact of an honor code and warning message in the context of the Amazon Mechanical Turk platform. By measuring the effect of honor codes in a second context, we explore whether the results of Experiment I generalize to another online environment.

6.1. Methods

Background. For some tasks on Mechanical Turk (e.g., "Transcribe this audio clip."), a requester can check for high-quality work by using techniques such as occasional "gold standard" tasks with known responses or by crosschecking a worker's answers with that of a second or third worker who completes the same task [Kittur et al. 2008; Oleson et al. 2011]. Other types of tasks—like opinion or psychological surveys—typically have no single correct response, so requesters posting these tasks must rely on the honesty and thoughtfulness of workers.

In particular, those who administer survey on Mechanical Turk must rely on participants to carefully read and diligently follow the survey instructions. Surveys are very common on the Mechanical Turk platform—on a recent day, 423 out of 1,945 task groups posted on Mechanical Turk contained the word "survey" as a keyword.

We used a five-page survey as our vehicle for measuring the effect of an honor code and warning on Mechanical Turk. The challenge of designing the survey task was to balance the conflicting requirements of: (1) being able to monitor cheating behavior on a per-workers basis, and (2) not making it obvious to workers that we could detect when they cheated. The second requirement was important for verisimilitude with other Mechanical Turk surveys: if workers knew that they were being monitored, they might alter their cheating behavior.

Survey Design. We mimicked the interaction between a requester and worker on Mechanical Turk by creating a survey asking a series of focus-group-style questions about a particular domain name that we owned ("pualikoa.com"). The introductory text of the survey explained that "[W]e want to understand how different people view a given domain name" and that we would be asking the participant to guess the characteristics of the site hosted at pualikoa.com.

The survey instructions repeatedly asked workers to *not* visit pualikoa.com while completing the survey, to avoid biasing their impressions of the domain name. Workers who visit the site, the instructions said, "will learn about the project and will be unable to answer honestly."

After workers consented to participate in our study, they were randomized into one of three groups, as in Experiment I. The three groups were: (1) control, (2) honor code, and (3) warning. The first group saw no honor code or warning, the second group had to type out a one-sentence honor code, and the third group had to type out a one-sentence summary of the warning.

We first launched the survey with a fourth "audio honor code" group, in which participants had to record themselves reading the one-sentence honor code. After running the survey for 9 days, we found that participants randomized into the audio condition aborted the survey mid-way at more than three times the rate of participants in other conditions (79.4% abort rate versus 22.4% abort rate). Since the supply of Mechanical Turk workers is limited, and since the participants who aborted the audio version of the survey were excluded from taking any other version of the survey, we decided to remove the audio condition from the study to maximize the number of workers available for other conditions.

The honor code closely followed that of Experiment I (Figure 1), and read

> I promise that I will not visit other websites or get help from other people while taking this survey.

The warning also followed that of Experiment I, except that the three consequences listed were adapted to the Mechanical Turk platform. The warning text explained to each worker that if we discovered that a worker visited other websites or took help from other people while taking the survey, we could: (1) reject the survey without payment, (2) "block" the worker's Turk account (preventing the worker from completing our tasks in the future), and (3) notify other requesters on Mechanical Turk.

The survey consisted of five pages.

(1) *Informed Consent*. The first page contained a standard informed consent form, explaining that workers had the option to participate in a research study involving "a 5–10min survey about a domain name."
(2) *Honor Code or Warning*. For workers randomized into one of the treatment groups, the second survey page contained the honor code or warning text. Workers had to type out the honor code or warning before continuing to the third page of the survey. Workers in the control group skipped directly to the third page of the survey.
(3) *Demographics*. The third page of the survey asked five demographic questions about the worker's sex, age, level of education, frequency of Turk usage (in hours per day), and income.
(4) *Survey Questions*. The fourth page of the survey contained five multiple-choice questions asking the worker to guess the purpose and target audience of the website hosted at pualikoa.com. The instructions (Figure 4) specifically asked workers *not to visit* pualikoa.com while completing the survey. For example, one question stated:

> Pualikoa.com targets users in a particular country.
> Please guess which country this is.

At pualikoa.com, we posted a "Coming soon" landing page (Figure 5), which revealed the correct answers to all of the survey questions.
(5) *Follow-up Question*. The last page of the survey asked the worker if she visited pualikoa.com while completing the survey (i.e., to admit to cheating). The ostensible reason for asking this question was to enable us to "filter out biased" survey results.

We want to understand how different people view a given domain name.

IMPORTANT! DO NOT VISIT pualikoa.com while completing the survey! If you visit the site, you will learn about the project and will be unable to answer honestly.

Fig. 4. Survey preamble.

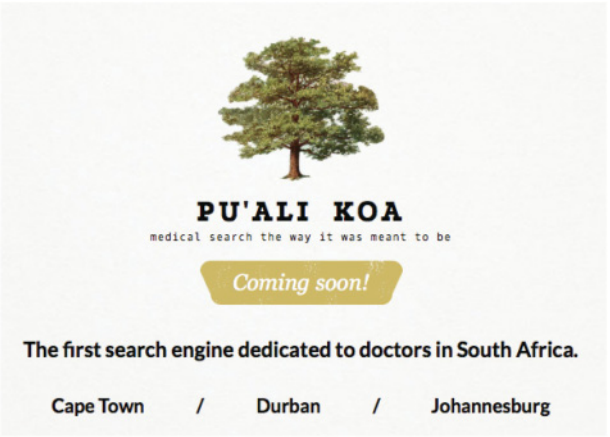

Fig. 5. The pualikoa.com home page.

We paid each worker who submitted the survey USD 0.30 for participating. At the start of the survey, we informed participants that we would award a USD 1.00 bonus to participants whose five answers matched the content at pualikoa.com (i.e., who guessed the correct answers to *all five* survey questions). The stated goal of the bonus was to "encourage thoughtful responses" to the survey questions. There was no bonus advertised or awarded for guessing correct answers to some but not all of the five survey questions. In this way, there was an incentive for participants to violate the survey instructions by visiting pualikoa.com. The answers to all of the survey questions were apparent from the landing page hosted at pualikoa.com.

Measuring the Prevalence of Cheating. For the purposes of this study, we defined "cheating" participants as those who visited pualikoa.com, in direct violation of the survey instructions, while taking the survey. Using pualikoa.com's web server logs, we could identify the IP address of every visitor to the website and could correlate these with the IP addresses and survey completion times of workers taking our Mechanical Turk survey.

We lumped together participants who knowingly violated the instructions of the survey with those who mistakenly violated the instructions (e.g., by not reading them carefully). Since, in the context of Mechanical Turk surveys, the latter form of negligence is as damaging as the former, we considered both of these violations as cheating.

We anticipated three methods of cheating.

—*Same-IP Cheating*. The survey-taker visited pualikoa.com from the same IP address used to visit our Mechanical Turk survey. In these cases, we could conclude with near certainty that the survey taker cheated. We filtered out multiple surveys from

Table I. Participant Population Size
(Including Participants Later Excluded)

	Phase I	Phase II	Phase III	Total
India	231	138	137	506
US	486	231	1,155	1,872

the same IP address, since it would not be possible to accurately assign blame for cheating in these cases.

—*Second-IP Cheating*. The survey taker used a second device (e.g., a mobile phone) to visit pualikoa.com while taking our Mechanical Turk survey. To identify second-IP cheaters, we correlated the time of the request to pualikoa.com with the survey submission time to identify which participant cheated.

—*Social Cheating*. The survey taker asked a friend or coworker—who had taken our survey previously—for the correct answers to the survey questions. Since this sort of social cheating does not generate traffic to pualikoa.com, we could not track it.

Identifying second-IP cheaters required a subtle analysis. We tried to assign responsibility for each visit to pualikoa.com to a known user. We first identified visits by search engine crawlers and same-IP cheaters. For the remaining visits, we first looked for cases in which a worker was on page four of our survey (which contained the first mention of pualikoa.com) at the same time as a new IP address visited pualikoa.com. We marked these as cases of second-IP cheating.

In 18 cases (out of a total of 1,712), there were multiple workers taking the survey at the same time as a new IP address visited pualikoa.com. In these cases, we did not mark any of the workers as cheaters, though we realize that this conservative approach might have led us to slightly undercount the number of second-IP cheaters. Even in the most pessimistic scenario, this undercounting would only have caused us to misclassify at most 1% of workers. Further, even if we misclassified ALL of the second-IP cheaters, the statistical significance of our results would not change.

We recorded the time of each request to our servers down to the second to ensure that we could detect workers who visited pualikoa.com immediately *after* submitting their survey results. There were a number of curious workers who visited pualikoa.com at some point after completing the survey and we did not label these workers as cheaters.

Participants. We ran two instances of the survey task: one only open to workers in the United States and one only open to workers in India. The tasks ran from August 12 until September 8, 2014, and we did not limit the number of respondents.

We conducted the survey in three phases. In each phase, we loosened the qualifications necessary for workers to participate in the study so that we could increase the size of our study population.

In the first phase, we followed the standard recommendations for conducting surveys on Mechanical Turk [Mechanical Turk Blog 2012]: workers had to have a 95% approval rating on prior tasks and have completed at least 1000 prior tasks. The first phase also included the audio-recorded honor code (as explained previously) as a fourth condition, so we required all participating workers to have a microphone attached to their computer. In the second phase, we removed the audio condition and the requirement for a microphone, since a small fraction of Mechanical Turk workers had audio recording capabilities and was willing to make a recording of the honor code. In the third phase, we removed the requirements of 95% approval and 1000 completed tasks. Table I indicates the number of workers who completed the survey in each phase of the study.

We received a total of 2,378 completed surveys, of which we excluded 254. We excluded surveys in which the worker:

—completed the survey from the same IP address using different Mechanical Turk accounts (57 workers),
—were assigned to the "audio" condition (59 workers),
—left one or more survey questions blank (15 workers),
—did not type out the honor code or warning (3 workers),
—aborted the survey and restarted it later (87 workers), or
—participated in one of our pilot studies (33 workers).

Of the 2,124 workers included in our final data analysis, 394 were from India (age: $M = 31.0$ years, $SD = 9.1$ years; 70% male) and 1,730 were from the United States ($M = 31.2$ years, $SD = 10.0$ years; 56% male).

Indian workers self-reported a mean total monthly income of USD 391 ($SD = 658$) and reported spending an average of 4.5hours ($SD = 3.9$hours) completing Mechanical Turk tasks daily. Workers in the United States reported a mean total monthly income of USD 2,626 ($SD = 3,608$) for an average of 3.5hours spent on Mechanical Turk daily ($SD = 3.0$hours).

6.2. Ethical Considerations

Many workers on Mechanical Turk heavily rely on income from the platform [Ross et al. 2010], so we were very careful to ensure that our design would not put workers at risk in any way, or even give them the impression of being at risk. The net effect on a worker's Mechanical Turk rating from participating in our study would have been *positive*: we "accepted" completed tasks from all workers within 48hours of their submission and we paid bonuses to workers within the same time period. Workers could abort the task at any time and we did not take any disciplinary measures against workers who cheated on the task.

All workers in the experiment knew that they would be participating in a research study, since we displayed a short consent form to workers before the task, explaining that it would be a study "about a domain name." We designed the consent text to be as concise possible to avoid taking up workers' time.

We collected workers' IP addresses as part of the task, since having IP addresses was necessary to detect second-IP cheating. Collecting workers' IP addresses is an accepted practice on Mechanical Turk: whenever requestors host their surveys externally (e.g., on Survey Monkey), the third-party server collects IPs by default and many requestors also collect IP addresses to detect duplicate submissions.

Finally, we did not debrief participants after the task. We reasoned that the risks of a debriefing would outweigh the benefits. If participants learned that we knew that they had cheated or lied on the task, the shame associated with that realization would likely be more detrimental than would the lack of a debrief. Institutional review boards at Microsoft Research and at Stanford University approved our experimental protocol.

6.3. Results

Effect of Honor Code and Warning. Our primary metric for evaluating the effect of the honor code and warning was the fraction of workers in each treatment group who "cheated" on the task (i.e., visited pualikoa.com while taking the survey). In both India and in the United States, we found that the honor code *did not* have a discernible effect on the rate of cheating and that the warning *did* have a statistically significant effect.

Figure 6 summarizes the rates of cheating in the United States and in India across each of the three experimental conditions. In both India and in the United States, we found that the warning caused a significant decrease in the rate of cheating. In India,

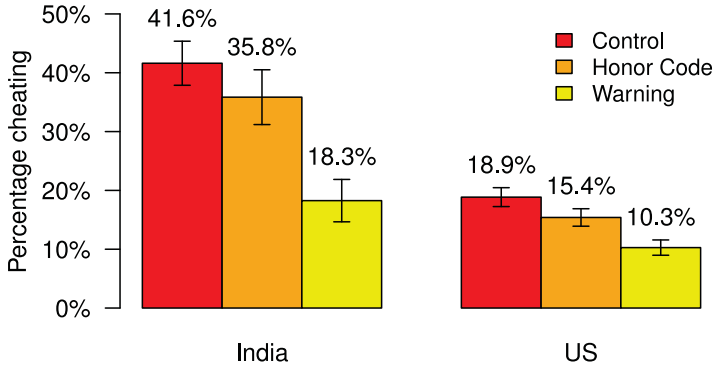

Fig. 6. Rates of cheating for each experimental condition, disaggregated by country. Error bars indicate the standard error of the proportion.

the rate of cheating fell to 18.3% from 41.6% ($\chi^2(1, N = 221) = 7.84$, $p < 0.001$) with application of the warning. In the US, the rate of cheating fell to 10.3% from 18.9% ($\chi^2(1, N = 1139) = 16.2$, $p < 0.001$) when workers were shown the warning before taking the survey. The honor code had no discernible effect on the rate of cheating in India ($\chi^2(1, N = 279) = 0.69$, $p = 0.41$) or in the United States ($\chi^2(1, N = 1185) = 2.26$, $p = 0.27$). As Figure 6 shows, the rate of cheating in the honor code condition was lower than in the control condition in both countries, but these differences were not statistically significant.

We did find, however, that the honor code had an effect on the fraction of cheating workers that *admitted* to cheating on the survey. As described earlier, the last page of the survey asked users to admit if they cheated on the task so that we could remove "biased" responses to the tasks. In India, 38.0% of cheating workers in the control group admitted to visiting pualikoa.com while taking the survey. This rate decreased to 10.5% when participants were shown the honor code before taking the task ($\chi^2(1, 110) = 8.37$, $p = 0.023$). In the US, the rate also decreased, to 6.6% from 19.6%, but this difference was not statistically significant ($\chi^2(1, 203) = 6.13$, $p = 0.066$). These results suggest that self-reported statistics on cheating are not always an accurate proxy metric for actual cheating. The warning had no significant effect on the rate at which cheaters admitted to visiting pualikoa.com.

When shown the warning message, workers aborted the survey (i.e., closed it without completing it) at a rate of 13.6%, compared with an abort rate of 7.4% for the honor code. This effect, which was statistically significant ($\chi^2(1, N = 1643) = 15.81$, $p < 0.001$), may have indicated that some participants were "scared off" by the warning. That is, the expected payoff of completing the survey for these users was not worth the potential risks of, for example, being falsely accused of cheating. However, even under the pessimistic assumption that the dropouts would have cheated at the base rate (i.e., the warning would have had no effect on them), the statistical significance of our results would not change if all of the dropouts completed the task.

Demographic Effects. We investigated demographic effects by comparing the results of the Mechanical Turk studies from India and the United States and by analyzing the workers' self-reported demographic data.

The baseline rates of cheating in the India and U.S. control groups were significantly different ($\chi^2(1, N = 767) = 36.8$, $p < 0.001$). Workers in the India control group cheated at a rate of 41.6%, compared with 18.9% in the US control group. While the difference in baseline cheating rates is large, any interpretation of this result should account for the fact that the value of the USD 1.00 payment was much greater in India than it was

Table II. Logistic Regression Analysis, Where the Binary-Dependent Variable is "Did Not Cheat"/"Did Cheat" (0/1)

	India			United States		
	Coefficient	Std. Error	z value	Coefficient	Std. Error	z value
(Intercept)	8.37×10^{-1}	1.03×10^{-0}	0.81	-1.11×10^{-1}	4.79×10^{-1}	-0.23
Age	$-7.16 \times 10^{-2***}$	1.72×10^{-2}	-4.16	$-4.96 \times 10^{-2***}$	9.44×10^{-3}	-5.26
Education	1.44×10^{-2}	5.30×10^{-2}	0.27	-2.60×10^{-2}	3.06×10^{-2}	-0.85
Female	-1.19×10^{-1}	2.56×10^{-1}	-0.47	-1.48×10^{-1}	1.43×10^{-1}	-1.03
Income	6.44×10^{-6}	6.76×10^{-6}	0.95	4.51×10^{-6}	3.33×10^{-6}	1.35
Qualified	-7.22×10^{-2}	2.47×10^{-1}	-0.29	1.12×10^{-2}	1.45×10^{-1}	0.08
Turk Hours	$7.25 \times 10^{-2*}$	2.86×10^{-2}	2.53	$4.38 \times 10^{-2*}$	2.11×10^{-2}	2.07

Significance values: $p < 0.05(^{*})$, $p < 0.01(^{**})$, $p < 0.001(^{***})$.

in the US. Workers in the US reported an average income seven times that of Indian workers, so the one-dollar bonus payment likely constituted a much larger incentive in India than in the US. Rather than looking at the absolute values of the baseline, it may be more useful to look at the relative effects of the honor code and warning on the two countries. Here, the pattern is almost identical: the rate of cheating in the warning code condition was roughly half of the rate of cheating in the control group, for both Indian and US workers.

Using the self-reported demographic data from the Mechanical Turk survey, we ran a logistic regression to determine the extent to which demographic features correlated with cheating behavior (Table II). In the regression analysis, we coded ages in years, education in years of school, and income as INR/month (India) or USD/year (US). The "Qualified" variable in the regression is a binary variable which takes on the value "1" if the worker was recruited in Phase I of our study, when we restricted participation to workers who had at least a 95% task approval rating and who had completed at least 1,000 tasks.

The logistic regression analysis rejects the null hypothesis that age is uncorrelated with cheating behavior ($p < 0.001$ for both India and the US). The negative coefficient on the age term in the regression reflects that older workers are *less* likely to cheat on the task than younger workers. The analysis also indicates that there is a positive correlation between the number of hours a worker reports spending on Mechanical Turk and the worker's likelihood of cheating ($p < 0.05$ for both countries). That is, the regression model predicts that workers who spend a greater number of hours on Mechanical Turk each day were *more* likely to cheat on our task.

One potential explanation for this finding is that workers who spend more hours on Mechanical Turk each day are more reliant on Mechanical Turk for income and thus have a stronger incentive to cheat. Alternately, it may be that these workers have had more extensive interactions with prior requesters, including some negative interactions that lead to more adversarial behaviors. Either hypothesis would yield the counter-intuitive finding that more experienced Mechanical Turk workers (who presumably work more hours per day) are *less* likely to complete tasks honestly. This would contradict the prevalent notion, also given on the official Mechanical Turk blog [Mechanical Turk Blog 2012], that more experienced workers will produce better work. We leave investigation of this phenomenon to future work.

The other terms in the analysis (education level, "qualified" status, and sex) do not have statistically significant predictive power ($p > 0.05$).

Cheating Behavior. We investigate cheating behavior by analyzing the method of cheating (same-IP versus second-IP), the effectiveness of cheating (distribution of correct answers), and other deviant behavior we observed during the course of the experiment. As expected, we observed a handful of workers using a second device (e.g., a

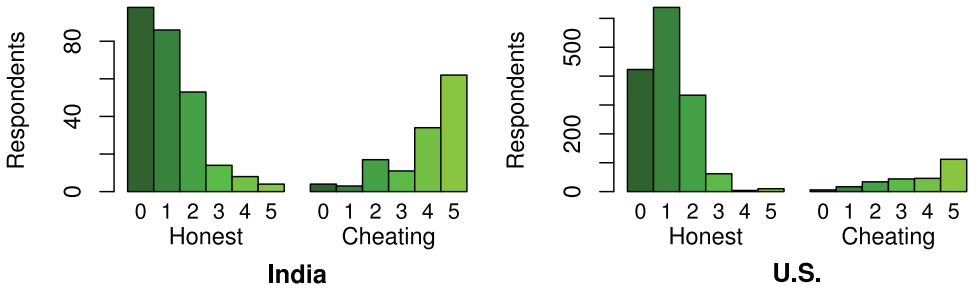

Fig. 7. Histogram showing the number of correct survey answers for honest and cheating participants in India and the U.S.

mobile phone or tablet) to visit pualikoa.com while taking our survey on their primary computer. Out of the 131 detected cheaters from India, 7 (5.3%) visited pualikoa.com from a different IP than the IP address they were using to complete the survey. Of the 259 detected cheaters from the United States, 32 (12.4%) visited pualikoa.com from a second IP address.

We observe that even workers who cheated on the survey often did not answer all five survey questions correctly and thus did not receive the USD 1.00 bonus. Figure 7 shows histograms of the survey scores, broken down by country. Of the 131 cheating workers in India, only 62 (47%) answered all five questions correctly. Of the 259 cheating workers in the United States, only 112 (43%) answered all five questions correctly.

We offer three potential explanations for this phenomenon.

—The first potential explanation is that workers intended to cheat but were not able to cheat effectively. We observed this phenomenon in Experiment I, in which cheaters, on the whole, had lower scores than honest students (Figure 3). These workers may have misread or misunderstood the survey questions, or they might not have taken the time to find the correct answer on the pualikoa.com landing page.
—The second potential explanation is that workers intended to cheat "partially"—they visited pualikoa.com to boost their score but they did not intend to get a perfect score. This partial-cheating behavior parallels prior findings that study participants often cheat in ways that do not maximize their pay-off as a way to maintain a positive self-concept [Fischbacher and Föllmi-Heusi 2013]. However, partial cheating would not have helped workers in our study, since we only awarded the USD 1.00 bonus if participants answered *all five* questions correctly.
—The third potential explanation is that workers were simply curious about pualikoa.com, so they visited the site while taking the survey, without intending to win the USD 1.00 bonus by answering all of the survey questions correctly.

Regardless of the reason for this partial cheating behavior, it indicates that high scores are a poor measure of cheating behavior. Although most participants with perfect scores did cheat on the task, we found that many cheaters did not receive perfect scores.

There were fifteen workers (0.7%) who admitted to cheating on the fifth page of the survey but who did not appear to cheat on the task. Out of five possible correct answers, these users had fewer than two correct answers on average ($M = 1.27$, $SD = 1.67$). These users may have been answering the survey questions carelessly or they may not have understood the text of the questions.

A total of 284 workers visited pualikoa.com after submitting the survey. Workers who visited the site after sending in their survey responses waited between 3s and two weeks to visit pualikoa.com, and the median time-until-visit for these workers was

15s. We did *not* label these "curious" workers as cheaters, since visiting the site after submitting the survey did not violate the rules of the task.

We found 57 workers (2.4%) who submitted surveys from the same IP address using different worker IDs. This is precisely the same rate of duplicate responses observed in prior work on Mechanical Turk [Berinsky et al. 2012].

7. DISCUSSION

The results of our two experiments are strikingly consistent. In both experiments, we measured significant baseline rates of cheating. In addition, we found that priming participants with a stern warning significantly decreased the rate of cheating by as much as 56%, while an honor code did not have a significant effect. In our second experiment, we found that these results were consistent in both India and the United States.

Our work shows that warnings can be a powerful tool for reducing the rates of cheating online. MOOC administrators, Mechanical Turk requesters, and other online community organizers might benefit from using pretask warnings. Two prominent MOOC platforms [Coursera 2014; edX 2014] currently use honor codes in the style of the one we tested in this paper (Figure 1). Our results suggest that a warning of negative outcomes may be a more effective deterrent to cheating.

Although our work finds that warnings can deter cheating online, there are a number of aspects of online settings that may complicate our results. One open question is how the effectiveness of a warning changes over time. If users see the same pretask warning repeatedly, they may become numb to the warning's threat of negative consequences. In addition, if the consequences are difficult to impose or if it is difficult to detect cheating behaviors, users may cheat in spite of the warning. Studying how the effect of online warnings changes over time would be an interesting challenge for future work.

For the case of crowdwork (e.g., Experiment II), there are two interesting considerations. First, warnings may have the potential to "scare off" users who have the option to skip a task. As noted in Experiment II, participants shown the warning were more likely to abort the task than those shown the honor code and those in the control group. Second, we have anecdotal evidence from conversations with crowd workers that experienced Mechanical Turk workers are fairly sophisticated at sniffing out poorly constructed or monitored tasks, and may be more likely to dismiss them as "nonserious." Since our task was somewhat odd and clearly easy to "game," experienced workers may have ignored the warnings and been more tempted to cheat. This is an alternative explanation for the findings related to the positive correlation between hours working on Mechanical Turk and likelihood to cheat, and would be an interesting topic to explore further.

Finally, we would like to investigate whether honor codes or warnings can have an effect on other types of antisocial behavior online. For example: do stern warnings deter online forum users from bullying other users? Since bullying and trolling can be extremely harmful, finding ways to reduce the prevalence of these behaviors would be valuable to many online communities.

8. CONCLUSION

We found that displaying a pretask warning that focused users on potential negative outcomes of dishonesty deterred cheating in the context of both an online exam and a Mechanical Turk task, typically reducing the rate of cheating by about 50% from the baseline. Further, we replicated previous findings, showing that priming participants with an honor code (instead of a warning) had no significant effect. Our results indicate that pretask warnings may be an effective and easy-to-implement alternative to honor codes in two distinct online contexts. Providers of online courses and crowdsourcing

platforms should take these findings into account as they design terms of service, user agreements, and honor codes for online use.

ACKNOWLEDGMENTS

We would like to thank Srinath Bala, Andrew Cross, Viraj Kumar, Madhusudan Parthasarathy, Shambwaditya Saha, Sumit Gulwani, and the whole MEC team for help creating and administering the online exam. We would also like to thank David Nemer for helpful conversations about the study design.

REFERENCES

Mauro Andreolini, Alessandro Bulgarelli, Michele Colajanni, and Francesca Mazzoni. 2005. Honeyspam: Honeypots fighting spam at the source. In *Proceedings of the Steps to Reducing Unwanted Traffic on the Internet on Steps to Reducing Unwanted Traffic on the Internet Workshop*. USENIX Association, 77–83.

Paul Baecher, Markus Koetter, Thorsten Holz, Maximillian Dornseif, and Felix Freiling. 2006. The Nepenthes platform: An efficient approach to collect malware. In *Recent Advances in Intrusion Detection*. Springer, New York, 165–184.

Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Anal.* 20, 3, 351–368.

Bear F. Braumoeller and Brian J. Gaines. 2001. Actions do speak louder than words: Deterring plagiarism with the use of plagiarism-detection software. *Political Sci. Politics* 34, 04, 835–839.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Edu. Psychol. Meas.* 20, 1, 37–46.

Henry Corrigan-Gibbs, Nakull Gupta, Curtis Northcutt, Edward Cutrell, and William Thies. 2015. Measuring and maximizing the effectiveness of honor codes in online courses. In *Learning @ Scale*. ACM, 223–228.

2014. Coursera Student Support Center: What is the Honor Code? Retrieved 20 September 2014 from http://help.coursera.org/customer/portal/articles/1164381-what-is-the-h onor-code-.

Thomas Crampton. 2007. Scamming the e-mail scammers. *New York Times* (Jul. 2007).

Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2012. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*. 26–30.

Jennifer Dirmeyer and Alexander C. Cartwright. 2012. Commentary: Honor codes work where honesty has already taken root. *The Chronicle of Higher Education* (Sep. 2012).

Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are your participants gaming the system?: Screening Mechanical Turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2399–2402.

edX 2014. edX Honor Code Pledge. Retrieved 20 September 2014 from https://www.edx.org/edx-terms-service.

Carsten Eickhoff and Arjen P. de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Inf. Retrieval* 16, 2, 121–137.

Federal Rules of Evidence. 1975. Rule 603. Oath or Affirmation to Testify Truthfully. Public Law 93-595, §1, 88 Stat. 1934. (Jan. 1975).

Gerlinde Fellner, Rupert Sausgruber, and Christian Traxler. 2013. Testing enforcement strategies in the field: Threat, moral appeal and social information. *J. Eur. Economic Assoc.* 11, 3, 634–660.

Urs Fischbacher and Franziska Föllmi-Heusi. 2013. Lies in disguise—an experimental study on cheating. *J. Eur. Economic Assoc.* 11, 3, 525–547.

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings ACM Conference on Human Factors in Computing Systems*. ACM, 1631–1640.

Francesca Gino, Maurice E. Schweitzer, Nicole L. Mead, and Dan Ariely. 2011. Unable to resist temptation: How self-control depletion promotes unethical behavior. *Org. Behav. Hum. Decision Processes* 115, 2, 191–203.

Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. 2013. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *J. Behav. Decision Making* 26, 3, 213–224.

Philip J. Guo and Katharina Reinecke. 2014. Demographic differences in how students navigate through MOOCs. In *Learning @ Scale*. ACM, 21–30.

Kevin A. Hallgren. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor. Quant. Methods for Psychol.* 8, 1, 23–34.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian J. Stat.* 6, 2, 65–70.

Eric Hoover. 2002. Honor for Honor's Sake? *The Chronicle of Higher Education* (May 2002).

John J. Horton, David G. Rand, and Richard J. Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Exp. Economics* 14, 3, 399–425.

Daniel Houser, Stefan Vetter, and Joachim Winter. 2012. Fairness and cheating. *Eur. Economic Rev.* 56, 8, 1645–1655.

Nicolas Jacquemet, Robert-Vincent Joule, Stéphane Luchini, and Jason F. Shogren. 2013. Preference elicitation under oath. *J. Environ. Economics Manage.* 65, 1, 110–132.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 453–456.

Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the Conference on Computer Supported Cooperative Work*. ACM, 1301–1318.

Xuanchong Li, Kai-min Chang, Yueran Yuan, and Alexander Hauptmann. 2015. Massive open online proctor: Protecting the credibility of MOOCs certificates. In *Proceedings of the Conference on Computer Supported Cooperative Work*. ACM, 1129–1137.

Frank M. LoSchiavo and Mark A. Shatz. 2011. The impact of an honor code on cheating in online courses. *MERLOT Journal of Online Learning and Teaching* 7, 2, 179–184.

Howard Markel. 2004. "I Swear by Apollo"–on Taking the Hippocratic Oath. *New Eng. J. Med.* 350, 20, 2026–2029.

David F. Mastin, Jennifer Peszka, and Deborah R. Lilly. 2009. Online academic integrity. *Teach. Psychol.* 36, 3, 174–178.

Nina Mazar, On Amir, and Dan Ariely. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *J. Marketing Res.* 45, 6, 633–644.

Donald L. McCabe and Linda Klebe Trevino. 1993. Academic dishonesty: Honor codes and other contextual influences. *J. Higher Educ.* 64, 5, 522–538.

Donald L. McCabe, Linda Klebe Trevino, and Kenneth D. Butterfield. 1999. Academic integrity in honor code and non-honor code environments: A qualitative investigation. *J. Higher Educ.* 70, 2, 211–234.

Nicole L. Mead, Roy F. Baumeister, Francesca Gino, Maurice E. Schweitzer, and Dan Ariely. 2009. Too tired to tell the truth: Self-control resource depletion and dishonesty. *J. Exp. Soc. Psychol.* 45, 3, 594–597.

Mechanical Turk Blog. 2012. Mechanical Turk Blog: Improving Quality with Qualifications—Tips for API Requesters. Retrieved from: http://mechanicalturk.typepad.com/blog/2012/08/. (Aug. 2012).

David Oleson, Alexander Sorokin, Greg P. Laughlin, Vaughn Hester, John Le, and Lukas Biewald. 2011. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Human Computation*. 43–48.

Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions Psychol. Sci.* 23, 3, 184–188.

Niels Provos. 2004. A virtual honeypot framework. In *USENIX Security Symposium*, Vol. 173. 1–14.

Gerald J. Pruckner and Rupert Sausgruber. 2013. Honesty on the streets: A field study on newspaper purchasing. *J. Eur. Economic Assoc.* 11, 3, 661–679.

John T. E. Richardson. 1994. The analysis of 2×1 and 2×2 contingency tables: An historical review. *Stat. Methods Med. Res.* 3, 2, 107–133.

Kevin W. Rockmann and Gregory B. Northcraft. 2008. To be or not to be trusted: The influence of media richness on defection and deception. *Org. Behav. Hum. Decision Processes* 107, 2, 106–122.

Stephen Mark Rosenbaum, Stephan Billinger, and Nils Stieglitz. 2014. Let's be honest: A review of experimental evidence of honesty and truth-telling. *J. Economic Psychol.* 45, 181–196.

Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In *Proceedings of the Extended Abstracts on Human Factors in Computing Systems*. 2863–2872.

Shaul Shalvi, Ori Eldar, and Yoella Bereby-Meyer. 2012. Honesty requires time (and lack of justifications). *Psychol. Sci.* 23, 10, 1264–1270.

Lisa L. Shu, Francesca Gino, and Max H. Bazerman. 2011a. Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality Soc. Psychol. Bull.* 37, 3, 330–349.

Lisa L. Shu, Nina Mazar, Francesca Gino, Dan Ariely, and Max H. Bazerman. 2011b. *When to Sign on the Dotted Line?: Signing First Makes Ethics Salient and Decreases Dishonest Self-reports*. Technical Report 11-117. Harvard Business School.

Debra Pogrund Stark and Jessica M. Choplin. 2009. A license to deceive: Enforcing contractual myths despite consumer psychological realities. *New York Univ. J. Law Business* 5, 2, 617–744.

Siddharth Suri, Daniel G Goldstein, and Winter A Mason. 2011. Honesty in an online labor market. In *Human Computation*. 61–66.

Alex B. Van Zant and Laura J. Kray. 2014. "I can't lie to your face": Minimal face-to-face interaction promotes honesty. *J. Exp. Soc. Psychol.* 55, 1, 234–238.

Kathleen D. Vohs and Jonathan W. Schooler. 2008. The value of believing in free will encouraging a belief in determinism increases cheating. *Psychol. Sci.* 19, 1, 49–54.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 319–326.

Luis von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: A game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 55–64.

Greg Walsh and Jennifer Golbeck. 2010. Curator: A game with a purpose for collection recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2079–2082.

Chen-Bo Zhong, Vanessa K. Bohns, and Francesca Gino. 2010. Good lamps are the best police darkness increases dishonesty and self-interested behavior. *Psychol. Sci.* 21, 3, 311–314.

Received May 2015; revised July 2015; accepted July 2015