I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13689

The Effect of Observing Multiple Private Information Outcomes on the Inclination to Cheat

Sandro Casal
Antonio Filippin

SEPTEMBER 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13689

The Effect of Observing Multiple Private Information Outcomes on the Inclination to Cheat

Sandro Casal
University of Trento

Antonio Filippin
University of Milan and IZA

SEPTEMBER 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Effect of Observing Multiple Private Information Outcomes on the Inclination to Cheat*

This paper investigates experimentally how the inclination to cheat changes when agents report the result of multiple realizations of a (private information) stochastic event rather than a single outcome. Extreme outcomes clearly signal opportunistic behavior with multiple realizations. The consequent reputation concerns dramatically reduce cheating by large amounts. Multiple draws, however, erode the intrinsic cost of lying, inducing a widespread inclination to slightly misreport the outcomes in a plausible manner. These two opposite effects are similar in magnitude, on average, but show an interesting gender differentiation implying that multiple realizations can be effective with males but may backfire with females.

JEL Classification: C81, C91, D82
Keywords: cheating, reputation concerns, moral self-licensing

Corresponding author:
Antonio Filippin
Department of Economics
University of Milan
Via Conservatorio 7
20122 Milano
Italy
E-mail: antonio.filippin@unimi.it

* We are grateful to the University of Milan for financial and logistic support. We would like to thank Maria Bigoni, Martin Dufwenberg, Caterina Giannetti, Matteo Ploner, as well as the participants to the 1st and 2nd BEEN workshop (Bologna and Rome), the 9th SEET Conference (Lecce), the 2018 ESA World Meeting (Berlin), the 13th Nordic Conference on behavioral and Experimental Economics (Odense), and the Euregio Economics Workshop 2019 (Trento) for their useful suggestions. All remaining errors are ours.

1. Introduction

Asymmetric information is a crucial issue in economic relationships and has been extensively studied given the harmful consequences that opportunistic behavior may have on other parties. Theoretical contributions have investigated how to design incentives and institutions capable of reconciling the conflicting interests involved.

For instance, the literature in labor economics has shown that monitoring can be an effective device to reduce informative advantages (e.g., Sappington, 1991; Walker, 2000; Whynes, 1993). But, besides several limitations highlighted by the literature (e.g., Falk and Kosfeld, 2006; Ichino and Muehlheusser, 2008; Walker, 2000) and other pitfalls for legal or ethical reasons, monitoring is scarcely informative of the agent's behavior whenever his productivity is jointly determined by his (unobservable) effort and by a stochastic component. Imagine, for instance, a shop owner that is disappointed by the sales and wonders whether the employee bears responsibility. Even tracing the number of persons who enter the shop on a given day, the owner cannot link the unsatisfactory result to the employee's effort. The bad performance can reflect a 'bad day,' for instance because most of the persons were not potential customers but rather their friends. As long as the stochastic component substantially contributes toward the final outcome, a single observation of a worker's productivity is uninformative. However, if the distribution of the stochastic component is known, a combination of outcomes may become informative thanks to repeated observations. The range of plausible claims shrinks considerably, but the employee still retains his private information. Truthfully reporting the realization of the stochastic components depends on his monetary and non-monetary incentives.

The *employer-employee* is not the only relationship affected by such a problem. Financial intermediation also falls in the realm of the principal-agent problem.[1] In these examples, but also in many other real-world settings, truth-telling may be affected by both a stochastic component and the way in which it has been obtained: via single or repeated interactions. It is worth stressing that while most of the economic interactions entailing private information are repeated, the literature on lying, well summarized by Abeler et al. (2019), is based almost

[1] For a discussion on agents' accountability in financial markets see Pollmann et al. (2014).

entirely on one-shot interactions. Therefore, we believe that investigating what happens with multiple observations of a stochastic component constitutes an original exercise that complements the existing literature.

This paper investigates how reporting the result of *multiple* independent realizations of a stochastic component affects the inclination to exploit one's private information. We envisage two effects. By shaping the ex ante probability distribution, a sequence of several independent realizations reveals an opportunistic behavior more easily than a single realization. In the example of shop sales, if low productivity was indeed due to poor effort the employee could easily conceal his private information by blaming it on bad luck as long as his performance is measured over a short spell of time. In contrast, a larger time horizon would expose the worker to a costly loss of reputation if the bad luck necessary to justify the low productivity becomes sufficiently unlikely. We refer to this effect as *reputation concerns*. Reporting the result of multiple realizations may have a second effect, changing the inclination to cheat given the same ex ante probability distribution of the stochastic component. We identify *moral self-licensing* (Monin and Miller, 2001) as the most relevant explanation. With moral self-licensing, being honest most of the times could help the agent to preserve his self-image, thereby making an occasional opportunistic behavior morally more acceptable than cheating when reporting a single outcome.[2]

Studying agents' behavior when they possess private information is a fundamental research question, but getting evidence is a challenging task due to the intrinsic nature of the principal-agent problem, i.e., the incentive to exploit one's informative advantage. This endeavor becomes by construction prohibitive with naturally occurring data when the object of the analysis involves an exogenous change in the fundamentals shaping private information. A solution is provided by laboratory experiments which grant the existence of a much higher degree of control with respect to field data (Falk and Fehr, 2003) and in this

[2]There is indirect evidence suggesting that the repetition of a task may indeed induce dishonest behavior, although in experimental settings not meant to investigate this specific issue. For instance, Fischbacher and Föllmi-Heusi (2013) find that subjects participating for the second time in the same cheating task claim higher payoffs than subjects without previous experience. In situations where subjects are required to report only the first among multiple realizations Shalvi et al. (2011) and Gino and Ariely (2012) find that favorable results in non-relevant rounds make subjects feel justified for acting in a more opportunistic manner. Falk and Kosfeld (2006) posit that repeated monitoring can be perceived as a signal of the principals' distrust loosening the agent's inhibition to misconduct.

paper we follow this route. Our experimental design is equipped to disentangle the two effects mentioned above in a clean manner, by designing and exogenously manipulating a cheating task.

The recent strand of experimental literature on (dis)honest behavior has analyzed to what extent agents exploit their private information, showing that subjects do not maximize their monetary return when doing so entails opportunistic behavior. Analyzing more than 70 studies, Abeler et al. (2019) report that, on average, subjects do misreport their private information, but the realized gains are only approximately 30% of what is possible. Such a robust finding can only be rationalized by some non-monetary cost of cheating. Abeler et al. (2019) claim that the experimental evidence is better explained by a combination of two factors. The first is an intrinsic cost of lying: agents have a preference for being honest, and cheating negatively affects their self-image. The second are *reputation concerns*: agents hold a preference for being seen as honest, and therefore cheating threatens their social image. While the intrinsic cost of lying is suffered whenever the agent cheats, the reputation concerns entail a cost that is incurred only when someone else is aware of his dishonest behavior.

Three recent contributions have modeled the non-monetary costs of cheating: Dufwenberg and Dufwenberg (2018); Gneezy et al. (2018); Khalmetski and Sliwka (2019). With specific characteristics in their formalization, all models share the importance of reputation concerns predicting that individuals cheat less when their claim is more likely to be perceived as dishonest by an observer.[3] All three models therefore predict that repeated outcomes should reduce the amount of cheating via reputation concerns.

Differences in the theoretical models are more pronounced as far as the intrinsic cost of cheating is concerned. In Dufwenberg and Dufwenberg (2018) the true realization of the stochastic event does not play any role. Khalmetski and Sliwka (2019) posit that subjects pay a fixed cost of lying, while in Gneezy et al. (2018) the cost also depends on the distance between the report and the true state. Regardless of the specific formalization, the individual

[3]Gneezy et al. (2018) corroborate experimentally the role of reputation concerns by finding that partial lying occurs only when the subjects hold private information. By not claiming the highest possible payoff, the subjects who care about their social image try to leave their dishonest behavior undetected. In contrast, when the behavior is directly observed by the experimenter the fewer subjects who lie (and therefore who do not care about their social image) tend to do so to the maximum extent.

cost of lying is procedurally invariant to the number of realizations of the stochastic component in all these models. In other words, whether a given result to be reported comes from a single vs. a sum of multiple realizations should not affect the intrinsic cost of cheating.[4]

Reporting multiple outcomes may trigger opposite effects via self- and social image. As a result, the overall effect of repeating a cheating task is not obvious a priori. Surprisingly, there are no studies in the literature that directly compare reporting a single vs. multiple outcomes in a clean manner. In their meta-analysis Abeler et al. (2019) do not find a significant effect of the repetition of the task. Their evidence, however, relies upon different studies in which other features change together with the number of rounds, and in fact they also stress that an experimental test of the repetition of a cheating task is missing.[5]

As described in Section 2, we analyze the effect of observing multiple outcomes by administering a properly designed cheating task in which subjects hold private information about the realization of a stochastic event. The experiment consists of three conditions meant to identify the social and the intrinsic cost of cheating implied by reporting the result of multiple random realizations. The first treatment considers a single realization of a stochastic event changing the underlying probability distribution in order to make extreme claims implausible. The second treatment manipulates the number of realizations observed, keeping the underlying distribution unchanged. In doing so we pay particular attention in keeping constant the monetary incentive of misreporting the outcome across conditions.

Our findings, presented in Section 3, show that both mechanisms are at work. The first consequence of reporting the result of multiple outcomes is the change in the underlying distribution of the stochastic event. High claims become unlikely and we observe that, coherently, they tend to disappear. In other words, reputation concerns significantly decrease the amount of cheating. There is a second effect, however. The mere fact of observing multiple outcomes triggers a more pronounced inclination to cheat by small amounts. Reporting multiple outcomes seems to facilitate some cheating occasionally, as if the corresponding

[4]Only in case it is referred to the observation of each single realization of the stochastic component rather than to the report of the cumulative result, the intrinsic cost of cheating would be higher under multiple realizations because the fixed cost would be paid repeatedly.

[5]Abeler et al. (2014) find truthful reporting in a representative sample of the German population both when a cheating task is performed once and four times. Also in this case, however, different incentives do not render the two conditions directly comparable.

intrinsic cost is lowered by behaving honestly most of the time. The two mechanisms are similar in magnitude but operate in opposite directions. As a result, the amount of cheating observed reporting one-shot vs. repeated outcomes is similar, confirming the result of the meta-analysis by Abeler et al. (2019).

Identifying these two mechanisms in a clean experimental setting has the advantage of emphasizing that it is not just irrelevant how many outcomes are observed. A clear example is obtained along a gender perspective. While opportunistic behavior is almost entirely restricted among males when cheating does not imply reputation concerns, repeated observations induce partial lying across the board. As a result, males and females behave in a more similar manner when reporting the result of multiple realizations.

2. Experimental Design

The typical situation in the labor market is that the total compensation of the worker is based on his observable productivity, which is jointly determined by his unobservable effort and a private information stochastic component. The experimental design implements an isomorphic but cleaner setup, in which effort is observable while total productivity is not. The main reason for this choice is that heterogeneous levels of effort could act as a confounding factor in the subsequent cheating task, for instance, because the subjects exerting a higher level of effort may feel entitled to exploit their informative advantage more. By keeping effort observable we can check that the design has been successful in removing this source of heterogeneity, and we can use effort as an ex-post control if some differences survive.[6] In more detail, we frame the cheating task as the determination of the stochastic component of the compensation for a real effort task in which effort is observable. In the cheating task subjects are rewarded according to their claim about the realization of a random number. The treatments consist of reporting either the result of a *single* random outcome from different distributions or the result of *multiple* outcomes. In all cases the stochastic realizations are private information. In order to correlate the choices in the cheating task

[6]Our approach also has the advantage of not revealing to the participants that the primary goal of the experiment is to investigate their inclination to cheat. Seeing the cheating task as an end-in-itself would likely trigger experimenter demand effects (de Quidt et al., 2018; Zizzo, 2010).

with potentially relevant preferences, we also elicit subjects' degree of risk aversion as well as their trusting and trustworthiness.

Cheating Task. The main research goal of this paper is to identify whether reporting multiple private information outcomes affects the inclination to cheat. Disentangling this effect from potential confounding factor requires a careful design. Note first that generating multiple outcomes by repeating a classic die roll task à la Fischbacher and Föllmi-Heusi (2013) would not deliver a clean comparison for two reasons. First, avoiding wealth effects would require to pay the outcome of the one-shot task and the average in the repeated task.[7] Doing so, however, would induce a difference in the minimum amount of cheating allowed. In the repeated task this amount would be a fraction (1/number of repetitions) of its counterpart in the one-shot version. Consequently, a subject whose intrinsic cost allows him to cheat only marginally is more likely to misreport the outcome in the repeated task than in the one-shot version.[8] Second and foremost, the two effects induced by the change in the underlying probability distribution and by the number of repetitions would be confounded.

We implement a task that replicates rolling a die once vs. 10 times, avoiding such confounding factors. More precisely, the experiment manipulates separately i.) the number of draws necessary to generate the same stochastic component, and ii.) the underlying probability distribution of the stochastic component, resulting in the following three treatments:

1. **1-Uniform.** Participants draw once from a bag containing 51 tokens numbered from 10 to 60. The number reported represents the monetary equivalent in Experimental Currency Units (ECU).[9]
2. **10-Uniform.** Participants draw 10 times (with replacement) from a bag containing six tokens numbered from 1 to 6. They are asked to report the sum of the 10 draws, which represents the amount in ECU to be paid.

[7] The alternative of paying one outcome at random in the repeated task is not viable for two reasons. On the one hand, it would imply the comparison of a sure amount in the one-shot task with a lottery in the repeated version, with potential consequences driven by risk preferences. On the other hand, the pay-one-at-random protocol is suitable for independent decisions, while reporting repeated outcomes entails intertwined choices.

[8] This caveat further clarifies why the comparison of one-shot vs. repeated tasks presented in the meta-study of Abeler et al. (2019) cannot deliver conclusive evidence.

[9] The exchange rate is 20 ECU = 1 Euro.

3. **1-Normal.** Participants draw only once from a distribution generated in the same way as in *10-Uniform*.[10]

Note that the random variable in the three conditions is identical in terms of support (10 to 60) and expected outcome (35). Moreover, this setting equalizes the minimum amount of cheating across treatments (1 ECU) regardless of the number of draws. The three conditions are isomorphic in terms of monetary incentives, but imply different non-monetary costs of cheating.

Subjects claiming 48 or more in *1-Normal* and *10-Uniform* can be identified as liars with 99% confidence because such outcomes can actually be observed with a cumulative probability of 0.94%. In contrast, no claim could be labeled as mendacious in the *1-Uniform* condition, given that even 60 has about a 2% chance of occurring (see Figure 1). Using a 95% confidence level, the threshold would be 45 and 59, respectively.[11] Whatever the threshold chosen, *10-Uniform* and *1-Normal* entail higher reputation costs of reporting a high number, given the same monetary benefit as in the *1-Uniform*.[12]

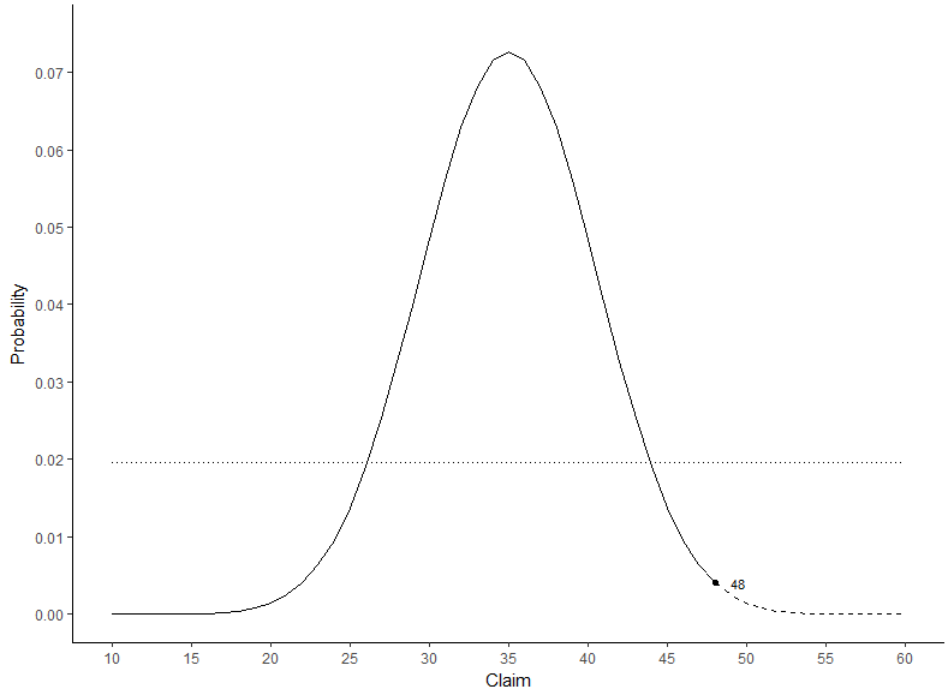
The *1-Normal* condition induces a dramatic change in the variance of the outcome with respect to the *1-Uniform*, while keeping constant the fact that subjects observe only a single outcome. The comparison of the results in these two conditions allows us to identify the effect of reputation concerns. The *10-Uniform* treatment manipulates the number of repetitions as compared to *1-Normal*, while keeping constant the underlying distribution of the event. Different choices across these two conditions can be attributed to the effect of observing multiple realizations of the stochastic component, where moral self-licensing can operate. The comparison between *1-Uniform* and *10-Uniform* represents the overall effect of observing multiple realizations, including both reputation concerns and moral self-licensing.[13]

[10]The distribution of the sum of 10 random draws from a uniform is not a normal distribution strictly speaking (e.g., it has zero density below 10 and above 60). Nevertheless, since it closely approximates a normal distribution we use this term for sake of simplicity.

[11]We disagree with the common practice in the literature of labeling as 'liars' those who report 6 when rolling a die, because a substantial fraction of them actually drew that number. As economists, we are ready to reject a null hypothesis (honest behavior in this case) only when the probability of type I error is sufficiently small.

[12]The thresholds that identify a dishonest behavior in *10-Uniform* and *1-Normal* depend on the confidence level, and in any case the formula to compute them is complex. Since we cannot expect the subjects to know it, in Section 3 we analyze lying behavior using a range of thresholds.

[13]In all the conditions subjects report only one outcome, either the single draw (*1-Uniform* and *1-Normal*) or the sum of the realizations *10-Uniform*. Geraldes et al. (2019) show that manipulating the number of outcomes



Note: The dotted line displays the uniform distribution in *1-Uniform*, in which each claim has a 1.96% chance of occurring. The solid and dashed line represents the (approximated) normal distribution of outcomes in *10-Uniform* and *1-Normal*. Claims on the dashed part of the line (i.e., starting from 48) can be classified as mendacious with at least 99% confidence.

Figure 1: Probability distribution of claims.

Real Effort Task. The real effort task chosen is the Coin Task (Gioia, 2016), which consists of recognizing the nominal value and the issuing country of a sequence of Euro coins. In every round subjects see a table describing 25 coins of different value and/or country. One of these coins is randomly selected and displayed to the participants, whose task is to identify it correctly (see Figure B.8 in the Appendix for an example). The deterministic compensation is 2 ECU for each coin correctly identified. Participants in our experiment have 15 minutes to correctly identify a maximum of 20 coins, with the possibility of making mistakes. Such a long time interval was purposely chosen to facilitate the successful completion of the task in order to avoid the introduction of confounding factors via wealth effects in the cheating task.[14]

reported, given the same underlying random event, does not alter the choices significantly.

[14] As shown in Section 3 all subjects indeed successfully completed the task.

Bomb Risk Elicitation Task (BRET). We elicit subjects' risk preferences using the BRET proposed by Crosetto and Filippin (2013). Participants are presented with a 10×10 square in which each cell represents a box: 99 boxes contain 1 ECU, while one contains a bomb. Participants choose how many boxes to collect $k_i \in \{1, 100\}$ knowing that if the bomb is collected earnings will be zero. The position of the bomb, $b_i \in \{1, 100\}$, is randomly determined after the participant's choice. If $k_i \geq b_i$, it means that the subject collected the bomb, which by exploding wipes out the earnings. In contrast, if $k_i < b_i$ the subject receives 1 ECU for every box collected. The number of boxes chosen provides a measure of risk attitude: the lower the number, the more risk-averse the subject. $k_i = 50$ represents a risk-neutral choice.

Mini Trust Game. We propose the mini-trust game reported in Figure 2 to measure participants' trusting and trustworthiness. Player A chooses between keeping 10 ECU or passing the money to Player B. In the second case the 10 ECU are multiplied by four, and it is then Player B who decides whether to keep the money or split it in a roughly equal way.[15] Player A is defined as trusting when passing the 10 ECU, while Player B is defined as trustworthy when returning 22 ECU back. The game is played only once in strategy method, so that every subject makes a choice as both Player A and Player B. At the end of the experiment participants are randomly matched in pairs and roles are randomly assigned. Individual payoffs are computed by combining participants' decisions in the assigned role.

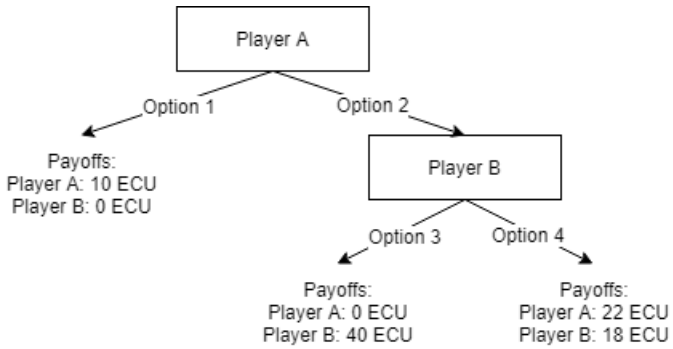


Figure 2: Mini Trust Game

[15] As noted by Ermisch and Gambetta (2006), this structure of the game without equal payoffs after Option 4 prevents choices from being driven by reasons of fairness. See Ermisch et al. (2009) for a detailed description of why this structure of the game is particularly appropriate for capturing trust and trustworthiness.

2.1. Experimental Procedures

The experiment was run between January 2018 and February 2019 at MiLab, the laboratory for experimental economics of the University of Milan. Subjects were recruited with ORSEE (Greiner, 2015) among university students, and the experiment was programmed and conducted using z-Tree (Fischbacher, 2007).

Upon their arrival, participants were randomly assigned to one of the seats, which were all separated by partitions. Participants were informed that the experiment was composed by three independent phases, with the respective instructions (see Appendix) read aloud at the beginning of each phase:

1. BRET;
2. Mini Trust Game;
3. Real Effort Task (with the Coin Task and the Cheating Task defining the deterministic and the stochastic part of the compensation, respectively.)

The three phases were preceded by a short questionnaire and followed by the resolution of uncertainty of the first two phases.

The experimental protocol is identical in all sessions as far as the questionnaire, the BRET, the Mini Trust Game, and the Coin Task are concerned. The only difference is the between-subjects manipulation of the Cheating Task (*1-Uniform*, *10-Uniform*, *1-Normal*). In *10-Uniform* subjects have to report the sum of 10 draws. In order to avoid a potential confounding factor due to a higher cognitive load, they are told that they can write down the outcomes using their smartphones or a piece of paper. In *1-Normal* physically performing the random draw from the underlying distribution is not feasible. Therefore, subjects draw the number accessing an external website with their smartphone. Before seeing the outcome, they are reminded that the number they are going to see is the sum obtained by drawing 10 times a number from 1 to 6 with replacements. Crucially, the different implementation grants the same level of anonimity as in *1-Uniform* and *10-Uniform*. The random number generator was programmed to replace the numbers actually displayed to the participants with a string of zeros after 10 seconds, thereby making impossible any (ex-post) comparison between claims and visualized numbers. Subjects see the numbers disappearing, but even in case they do not believe that the numbers shown are not kept in the external server,

it is clear that lies cannot be detected. In fact, the subjects access the website at the same time without leaving their cubicles, using the same QR code from their private smartphones about which no information is available to the experimenters.

In addition to a show-up fee of 2.50€, participants received the payoff of each phase, with that in the real effort task composed by two parts: a deterministic part linked to the performance in the Coin Task, and a stochastic component determined by the participant's claim in the Cheating Task. The pay-all-tasks protocol is the most appropriate in this experiment because the compensation for the overall productivity needs to encompass both the effort exerted and the stochastic component. Moreover, the outcome in the real effort task always needs to be paid to make sure that all the underlying incentives are salient.[16] For the sake of salience of the non-monetary incentives, payments are made privately but directly by the experimenters. On average, sessions lasted 50 minutes and the average earnings amounted to 8.15€.

3. Results

Table 1 presents the descriptive statistics of the sample across treatments. Three hundred and sixty participants took part in the experiment, 182 in the *1-Uniform*, 88 in the *10-Uniform*, and 90 in the *1-Normal* treatment. The larger sample in *1-Uniform* is due to the higher variance of the outcomes.

The groups are balanced from a gender perspective. Trust and trustworthiness are also indistinguishable. The randomization instead failed to generate equality across treatments in terms of risk aversion since participants in *1-Normal* turned out to be significantly more risk-averse. As shown below, however, this variable does not play a significant role in explaining the inclination to cheat and therefore it does not constitute a confounding factor. All participants achieve the maximum score in the Coin Task, ensuring that claims in the Cheating Task are not influenced by a different level of effort exerted or by the corresponding monetary rewards.

[16] At the same time, wealth effects are not an issue in our setting. On the one hand, earnings in the Coin Task are equal for all the participants. On the other hand, the resolution of uncertainty in the first two phases occurs only at the end of the whole experiment and subjects cannot secure any positive amount with their choices.

Table 1: Descriptive statistics

	Treatment			p-value (Test)		
	1-U	10-U	1-N	1-U vs. 10-U	1-U vs. 1-N	10-U vs. 1-N
N. of Subjects	182	88	90			
% of Females	53.30	55.68	60	0.795 (FE)	0.303 (FE)	0.650 (FE)
BRET	42.63	41.59	34.70	0.623 (MW)	< 0.001 (MW)	0.007 (MW)
% Trusting	64.29	60.23	62.22	0.591 (FE)	0.789 (FE)	0.878 (FE)
% Trustworthy	53.30	48.86	56.67	0.518 (FE)	0.608 (FE)	0.367 (FE)
# of Coins	20	20	20	.	.	.
Cheating Task	39.76	39.27	35.99	0.244 (MW)	0.006 (MW)	0.002 (MW)

Note: 1-U = 1-Uniform; 10-U = 10-Uniform; 1-N = 1-Normal; FE = Fisher's Exact Test; MW = Mann–Whitney U test

The average claim in the Cheating Task, our main variable of interest, is 38.7 on average. This amount is significantly higher than the expected outcome (35). Hence, subjects lie on average. However, as commonly found in the literature, they exploit the opportunity to cheat only to a low extent. The extra claim, computed as the difference between participants' claim and the theoretical prediction without cheating, is only 3.70 ECU on average, i.e. about 15% of the possible maximum (25 ECU). The intrinsic cost of cheating, which is constant across all the experimental conditions, seems to be the most important bulwark against opportunistic behavior.

Table 1 shows that the average claims in *1-Uniform* and *10-Uniform* are virtually identical. This result corroborates the evidence provided by Abeler et al. (2019), confirming that what they find comparing different papers also holds within a homogeneous subject pool and removing possible confounds. Nevertheless, this finding is counterintuitive because it suggests that reporting the sum of several random draws does not affect the average degree of cheating despite reputation concerns only characterizing the *10-Uniform* treatment.

The value added of our experiment is the possibility to explain what drives this 'null' result, excluding that reputation concerns play no role. The similar outcome in these two conditions is in fact due to the composition of the effects of both reputation concerns and moral self-licensing, which operate in opposite directions. The *1-Normal* treatment allows us to disentangle these two effects. We can see in Table 1 that claims in *1-Normal* are indeed significantly lower than in *1-Uniform* and *10-Uniform*, and fairly close to the theoretical prediction without cheating.

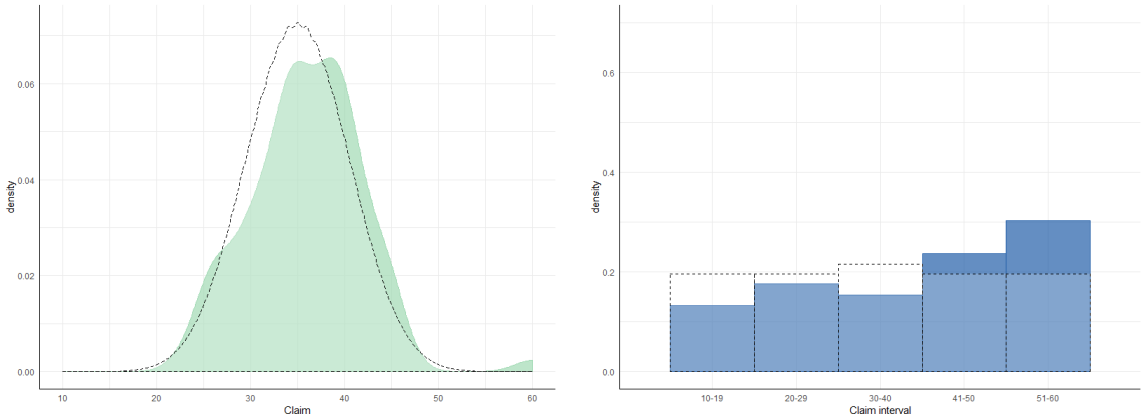
3.1. Reputation concerns

We focus first on reputation concerns, which in our setting can be identified by comparing the *1-Uniform* and the *1-Normal* treatments. In both conditions subjects perform a single random draw. However, by reducing the range of outcomes that can be reasonably expected, the *1-Normal* condition makes an opportunistic behavior detectable also at the individual level. The disutility suffered when others (the experimenters in this case) believe that their behavior is dishonest should affect subjects' claims only in *1-Normal*. In contrast, reputation concerns are virtually absent in the *1-Uniform* condition. Since all the other determinants (monetary incentives, intrinsic cost of lying, number of random draws observed) are constant, the difference in the claims in these two treatments measures the intensity of reputation concerns. Claims are indeed significantly higher in *1-Uniform* than in *1-Normal* (39.76 Vs. 35.99). We therefore find that reputation concerns matter.

The average claim is a useful statistic, but it overlooks a lot of relevant information. Saying that reputation concerns matter, on average, is not informative about the underlying pattern. For instance, is this result driven by a slight but widespread change in the behavior, or does it follow from an effect concentrated on the highest claims?

A first answer to this question is provided by Figure 3, which describes the observed distribution of claims compared with their theoretical counterparts (dotted lines). The observed claims in *1-Normal* (Panel 3a) follow rather closely the underlying theoretical distribution. As regards *1-Uniform* (Panel 3b), the distribution is markedly skewed to the left. The absence of reputation concerns facilitate an opportunistic behavior that is more evident among the highest claims. At the same time we still observe a substantial fraction of low (and presumably truthful) claims.

In order to assess how widespread cheating is with and without reputation concerns, we first need to clarify what we deem as opportunistic behavior at the individual level. Such a definition is important because reputation concerns are suffered only if others believe that the claim is dishonest. Imagine a subject observing a realization of the stochastic component equal to 25 in *1-Normal*. While this subject will suffer the intrinsic cost of cheating whenever reporting more than 25, he will suffer reputation concerns only if the claim is sufficiently unlikely to be meaningfully labeled as mendacious. Since any rule to classify a claim as



(a) 1-Normal (b) 1-Uniform

Note: For graphical convenience, we report the kernel density of observed claims in *1-Normal* (Panel 3a) and histograms (collapsed in 5 bins) for *1-Uniform* (Panel 3b). Dotted lines represent the expected frequencies.

Figure 3: Reputation concerns: Density of claims by treatment.

dishonest intrinsically contains a degree of arbitrariness, we decide to rely upon the conventional 95% confidence level. This approach identifies the range 45–58 as that in which claims are unlikely to occur only in *1-Normal*. In fact, claims up to 44 leave more than a 5% probability that a greater or equal outcome is actually also observed in *1-Normal*. Consequently, the probability of a type I error is too high to reject the null assumption of honest reporting. Conversely, 59 and 60 constitute unlikely claims in *1-Uniform*, too. Hence, the 45–58 range is where the effect of reputation concerns should be reasonably expected.

The next step is to estimate the fraction of cheaters along the distribution of claims. In doing so we have to take into account the different objective probabilities of each outcome occurring in the two treatments (see Figure 1). A simple statistic is provided by the percentage of subjects reporting more than a *threshold* divided by the percentage that should not have observed it given the objective probability (Abeler et al., 2019). For instance, in *10-Uniform* 112 subjects out of 182, i.e., 61.5%, claim strictly above 35. Only about 49% should have actually observed a higher value. Hence, we observe 12.5% of claims higher than 35 coming from the 51% of the population that should have observed an outcome up to 35. Therefore, 24.5% ($^{12.5\%}/_{51\%}$) is the estimated percentage of cheaters. Garbarino et al. (2018) develop a method to estimate the fraction of cheaters weighting all the possible realizations of the random variables according to their objective probability rather than imputing the expected outcome. In the example above, the method iterates the computation of the

percentage of cheaters assuming that any number of subjects, from 0 to 112, have actually observed an outcome higher than 35.[17] Another advantage of this procedure is that it also provides an estimate of the probability distribution of the fraction of cheaters.

Table 2 reports for every threshold between 45 and 58:

1. the estimated fraction of cheaters by treatment;
2. the p-value of a chi-square test for the difference in the proportions.[18]

Table 2: Reputation concerns

	Threshold													
	45	46	47	48	49	50	51	52	53	54	55	56	57	58
% of cheaters														
1-Normal	0.8	0.45	0.6	0.74	0.88	0.98	1.04	1.08	1.09	1.1	1.1	1.11	1.11	1.11
1-Uniform	17.3	18.79	18.73	19.38	17.91	13.08	13.16	12.58	12.03	12.13	10.39	10.52	10.06	8.47
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	0.003	0.003	0.004	0.005	0.005	0.012	0.011	0.015	0.034

Note: Every column represents a binary partition of the claims between those lower than or equal to the indicated threshold from those strictly higher. The fraction of subjects lying for every threshold is estimated following Garbarino et al. (2018). The p-value refers to a chi-square test for the difference in the proportions across treatments.

In *1-Uniform*, where each claim is equally likely, claims in the range 45–58 are observed with a frequency in the order of 10–19% higher than expected. Interestingly, this fraction decreases as we approach the upper bound of the range where reputation concerns should also matter in this treatment according to our approach. In *1-Normal* claims over the whole interval signal opportunistic behavior and we observe that claims higher than expected tend to disappear. As a result, for every threshold the estimated fraction of cheaters is significantly higher in *1-Uniform* than in *1-Normal*. Summarizing, we find compelling evidence that the likelihood of being detected as a liar triggers reputation concerns:

Result 1. When indicative of opportunistic behavior the occurrence of high claims tends to converge to its expected frequency.

[17] The upper bound is 112 because of the assumption that subjects do not lie downward.

[18] We apply the Yates's continuity correction because the frequencies are lower than 5 for some thresholds. See for instance Hitchcock (2009) for a discussion of the pros and cons of this type of correction.

3.2. Multiple observations

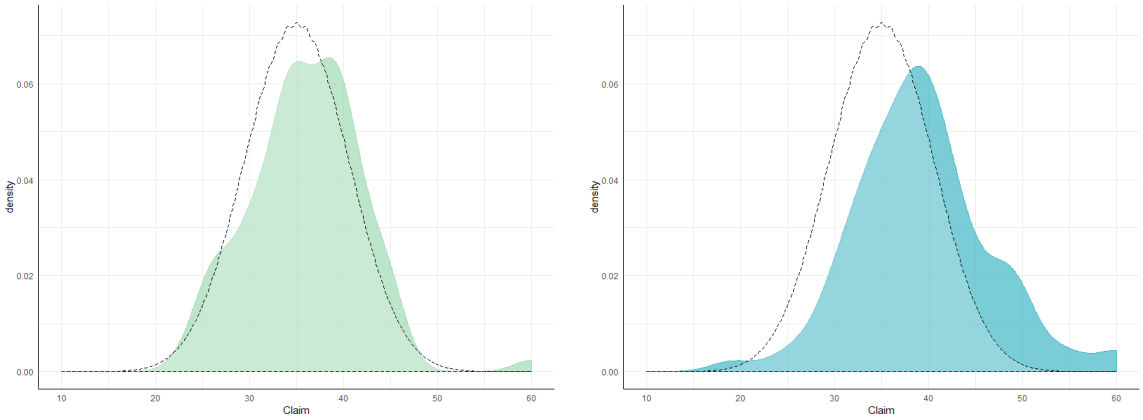
We now get to the main research question of the paper, i.e., whether the mere observation of multiple stochastic outcomes plays an independent role. This effect can be identified in our experimental setting by comparing *10-Uniform* and *1-Normal*. In these two treatments subjects report only one number, whose objective distribution is identical. The distribution is obtained in both cases drawing 10 times randomly – with replacement – a number from 1 to 6. Extreme claims have the same (low) probability of occurring and reputation concerns are therefore constant. Any difference in the choices must depend on the way the final outcome is generated. In *10-Uniform* the subjects perform the 10 random draws, while in *1-Normal* they draw only once from the final distribution generated with the same procedure. As already shown in Table 1, claims are indeed significantly higher in *10-Uniform* than in *1-Normal* (39.27 vs. 35.99). We therefore find that observing multiple outcomes increases the inclination to cheat. This effect is of the opposite sign but is comparable in size to reputation concerns.

Also, in this case we try to shed some light on the underlying mechanism going beyond the average claim. Figure 4 reports the observed claims compared with the same theoretical distribution (dotted lines) in the two conditions. The vertical distance between the solid and dotted lines provides an indication of the incidence of aggregate cheating along the distribution. It is immediately evident that the observed distribution of claims in *10-Uniform* (Panel 4b) appears to shift entirely to the right as compared to *1-Normal* (Panel 4a). The implication is that while reputation concerns are still somehow effective in hindering very high claims, observing multiple outcomes seems to induce a substantial misreporting toward medium-high values.

This speculation is confirmed by looking at the proportion of participants claiming values strictly higher than the average theoretical claim (35). Such a fraction is 50% in *1-Normal* and 68.18% in *10-Uniform*.[19] Once again, we implement the Garbarino et al. (2018) method to estimate the fraction of cheaters along the distribution.[20]

[19] A chi-Square test confirms that the figure in *10-Uniform* is significantly higher than the corresponding theoretical frequency ($p = 0.005$).

[20] Thresholds below 30 are not reported because the number of potential cheaters is too low to make any meaningful inference.



(a) 1-Normal (b) 10-Uniform

Note: Kernel density of observed claims in both conditions: *1-Normal* (Panel 4a) and *10-Uniform* (Panel 4b). Dotted lines represent the expected frequency.

Figure 4: Multiple observations: Density of claims by treatment.

We find that multiple observations of the stochastic event induce opportunistic behavior. Up to the 49 threshold, the percentage of cheaters is significantly higher in *10-Uniform* as compared to *1-Normal*. No differences are observed for higher thresholds, suggesting that reputation concerns are still effective in preventing very high claims. Repeated observations seem therefore to erode the intrinsic moral of the subjects, and they therefore foster partial lying. Subjects tend to readjust their outcome slightly upward, while still preserving their social image.

Result 2. Multiple observations of a random event increase the inclination to cheat, thus eroding the intrinsic cost of cheating.

We found three explanations in the literature that are broadly consistent with this interpretation that multiple observations may reduce the intrinsic cost of cheating. The first is that subjects tend to indulge in opportunistic behavior as long as their lies are justifiable both in the eyes of others and to themselves (Shalvi et al., 2011). This self-justification is intended as '*observed* desired counterfactuals' and has been shown by Shalvi et al. (2011) to operate in repeated tasks. For instance, subjects might feel justified when changing some unlucky draws with higher values that they have actually observed. Although appealing, we believe that this explanation does not fit well to our setting for two reasons. First, in *10-Uniform* each draw cannot be meant as a counterfactual strictly speaking, because all the

Table 3: Observing multiple outcomes

	Threshold													
	30	31	32	33	34	35	36	37	38	39	40	41	42	43
% of cheaters														
1-Normal	15.25	15.35	24.10	22.99	16.55	9.59	14.23	17.51	11.39	9.36	5.92	3.46	3.74	5.68
10-Uniform	53.55	50.49	53.25	52.77	47.82	40.07	43.49	42.53	35.14	31.18	24.39	24.03	19.32	17.90
p-value	0.037	0.025	0.045	0.020	0.005	0.001	0.002	0.005	0.003	0.002	0.003	$<$0.001	0.004	0.026

	Threshold												
	44	45	46	47	48	49	50	51	52	53	54	55	56-59
% of cheaters													
1-Normal	1.56	0.80	0.45	0.60	0.74	0.88	0.98	1.04	1.08	1.09	1.10	1.10	1.11
10-Uniform	16.00	16.03	14.55	12.8	12.02	7.68	4.40	4.47	4.51	4.53	3.40	2.27	2.27
p-value	0.002	0.001	0.001	0.003	0.006	0.061	0.343	0.346	0.349	0.348	0.596	0.981	0.985

Note: Every column represents a binary partition of the claims between those lower than or equal to the indicated threshold from those strictly higher. The fraction of subjects lying for every threshold is estimated following Garbarino et al. (2018). The p-value refers to a chi-square test for the difference in the proportions across treatments.

outcomes count toward the final payoff. Second, it is difficult to argue that lies are more justifiable to others in *10-Uniform* than in *1-Normal* because the underlying distribution is the same.

The second explanation refers to ethical blind spots, which could rationalize a more pronounced cheating in *10-Uniform* with an attentional bias (Pittarello et al., 2015). According to this justification subjects' attention may focus on pleasant information (high numbers) during the repetition of the task, while overlooking bad outcomes (low numbers). Although this mechanism can potentially operate only when multiple outcomes are observed, i.e., in *10-Uniform*, we do not find it fully convincing either. An upward adjustment of the sum requires an active effort to substitute low numbers with higher ones. An attentional bias that simply disregards unpleasant information would not be enough.

The third explanation refers to the so-called moral self-licensing effect (Monin and Miller, 2001) and is in our opinion the most convincing one. This effect entails that subjects are more likely to cheat when their misconduct can be compensated by behaving in a moral way in other circumstances, something that in our framework can only occur when observing mul-

tiple outcomes.[21] A possible pattern in *10-Uniform* is that after taking truthful note of some outcomes, subjects may feel entitled to adjust upwards the result in a subsequent round. Mendacious behavior may then induce them to return to an honest attitude in order to restore their self-image (moral cleansing effect, Blanken et al., 2014).[22] The exact sequence of decisions is not important and the implication is that subjects are honest most of the times while occasionally indulging in opportunistic behavior. The final claim should then exceed the actual sum only by few units, because in a single round the upward adjustment is limited by the maximum number being 6. The implications of this mechanism are therefore consistent with the increase of partial lying that we observe.

3.3. The overall effect

Summarizing, the similar degree of cheating in *10-Uniform* and *1-Uniform* is due to the composition of two opposite effects. On the one hand, the multiple observations decrease (likely via moral self-licensing) the intrinsic cost of cheating, leading to a more pervasive inclination to readjust the outcome slightly upward. On the other hand, the repetition of the task introduces the social cost of cheating (via reputation concerns) that prevents very high claims that would otherwise be observed. The two effects are similar in magnitude leading to the 'null' result across *1-Uniform* and *10-Uniform*.

The overall effect stemming from Results 1 and 2 is summarized in Figure 5. Figure 5 displays for each treatment the difference between the observed and the expected fraction of claims in seven intervals of equal size between 36 and 60. A negative bar implies that fewer subjects than expected claim in that interval, while a positive value represents claims in excess. This difference is not directly informative of the percentage of cheaters, because it is compatible with different underlying dynamics at the individual level. Nevertheless, it has the merit to visualize immediately where anomalous behavior is concentrated at the aggregate level.

[21] See Blanken et al. (2015) and Effron and Conway (2015) for comprehensive reviews of the argument or Ploner and Regner (2013) and Clot et al. (2014) for examples of moral self-licensing in cheating games. Although evidence of such an effect comes primarily from the experimental literature, moral self-licensing is a robust phenomenon that also occurs outside the laboratory (Hofmann et al., 2014).

[22] Barron (2019) finds that subjects lie downward in low-stakes situations in order to signal honesty and to mitigate the repercussions of upward lying in high-stakes contexts.

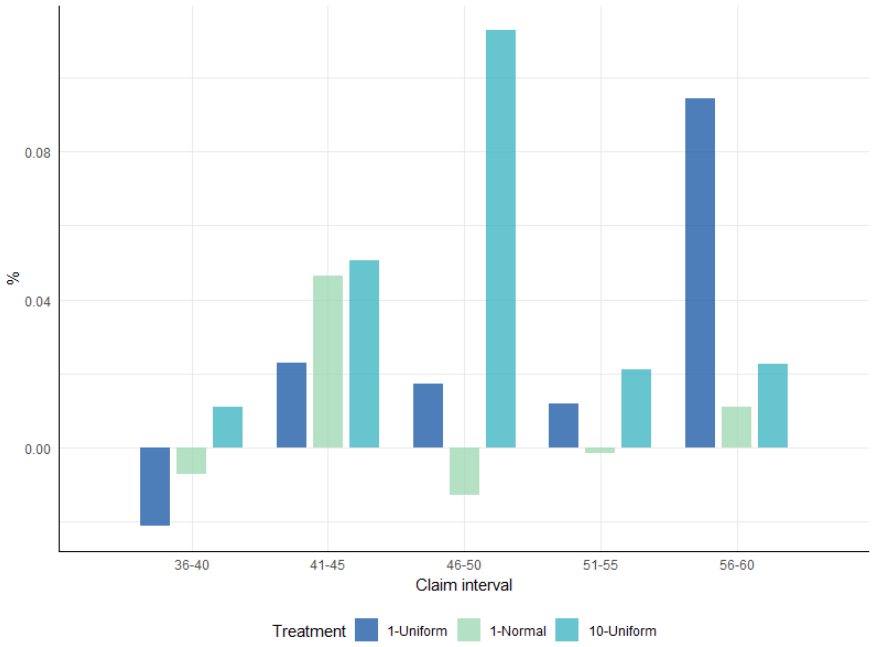


Figure 5: Difference between observed and expected fraction of claims in the range of value 36–60.

Without reputation concerns and with a single random draw (*1-Uniform*) the large majority of cheating occurs by claiming very high payoffs. The introduction of reputation concerns (*1-Normal*) dramatically reduces the inclination to cheat. We only detect a small incidence of partial lying in the interval 41–45. Therefore, to some extent our results support the finding of Gneezy et al. (2018), that a lower probability of the maximum payoff induces partial lies.[23]

What substantially increases the occurrence of partial lying is observing multiple random events (*10-Uniform*). The repetition of the task seems to lower the intrinsic cost of cheating, inducing a more widespread inclination to misreport outcomes, as long as they are plausible and do not signal opportunistic behavior.[24]

Result 3. Observing multiple random events has two opposite effects at the same time:

i) it prevents cheating by very large amounts thanks to reputation concerns;

ii) it increases partial lying because of a lower intrinsic cost of cheating.

[23] In Section 3.5 below we show that this effect is concentrated among males.

[24] It is worth noting that the subjects claiming above 55 in this treatment are all unfamiliar with joint probabilities according to their answers in the pre-experimental questionnaire. See AppendixA for the details.

3.4. Determinants of cheating behavior

In this section we investigate the determinants of cheating behavior using subjects' observable characteristics and behavioral traits as explanatory variables. Table 4 presents the results of different linear regression models in which the dependent variable is the subject's claim. The treatment effects are captured by the *1-Uniform* and *10-Uniform* dummies. The omitted condition is *1-Normal*, so that the coefficients of *1-Uniform* and *10-Uniform* identify the absence of reputation concerns and observing multiple random draws, respectively. Column 1 trivially confirms that both effects are strongly significant.

Table 4: Linear regression models

	Dependent variable:			
	Claim in the cheating task			
	(1)	(2)	(3)	(4)
1-Uniform	3.775***	3.361**	3.572***	3.209***
	(1.519)	(1.534)	(1.511)	(1.527)
10-Uniform	3.284***	2.790***	3.153***	2.712***
	(1.767)	(1.770)	(1.756)	(1.760)
Trusting		−3.696***		−3.676***
		(1.409)		(1.401)
Trustworthy		0.105		0.333
		(1.369)		(1.365)
Risk		0.062		0.058
		(0.039)		(0.039)
Female			−3.022**	−2.815**
			(1.244)	(1.237)
Constant	35.989***	36.070***	37.802***	37.747***
	(1.242)	(2.045)	(1.442)	(2.163)
Observations	360	360	360	360
Adjusted R^2	0.012	0.033	0.026	0.044
F Statistic	3.226** (df = 2; 357)	3.450*** (df = 5; 354)	4.149*** (df = 3; 356)	3.772*** (df = 6; 353)

Note:
1-Uniform and 10-Uniform are treatment dummies.
Trusting is a dummy equal to 1 when the subject passes the 10 ECU in the Mini Trust Game.
Trustworthy is a dummy equal to 1 when the subject returns the 22 ECU in the Mini Trust Game.
Risk Tolerance is the number of boxes collected in the BRET. Female is a gender dummy.
Robust standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01.

Column 2 adds as explanatory variables the elicited behavioral traits. In more detail,

Trusting is equal to 1 when the participant passes the 10 ECU as Player A in the Mini Trust Game. *Trustworthy* is equal to 1 when the participant returns 22 ECU as Player B. *Risk Tolerance* is a continuous variable capturing the choice in the BRET. Only *Trusting* displays a significant correlation, with more trusting subjects claiming lower outcomes. In contrast, trustworthiness does not affect participants' claim. These results look counterintuitive as we would have expected the trustworthy subjects to cheat less. Rather than the irrelevance of trustworthiness, this finding possibly signals that trust to other participants and trust to the experimenters do not necessarily coincide. Risk aversion negatively correlates with the level of the claims, although not significantly so. The treatment effects are robust to the inclusion of these controls. This fact is particularly reassuring because we have seen in Table 1 that the groups are not balanced ex ante in terms of risk aversion. Indeed, Column 2 shows that the treatment effects are genuine and that risk aversion does not act as a confound.

Columns 3 and 4 replicate the analysis while also including a gender dummy. The results above hold virtually unchanged in terms of size and significance while females claim significantly less than males on average, in line with what was previously found in the literature (Abeler et al., 2019). In the next section we analyze in more detail the treatment effects along a gender dimension.

Result 4. *i)* Behavioral traits do not play a major role: only trusting leads to lower claims;
ii) Females claim less than males on average.

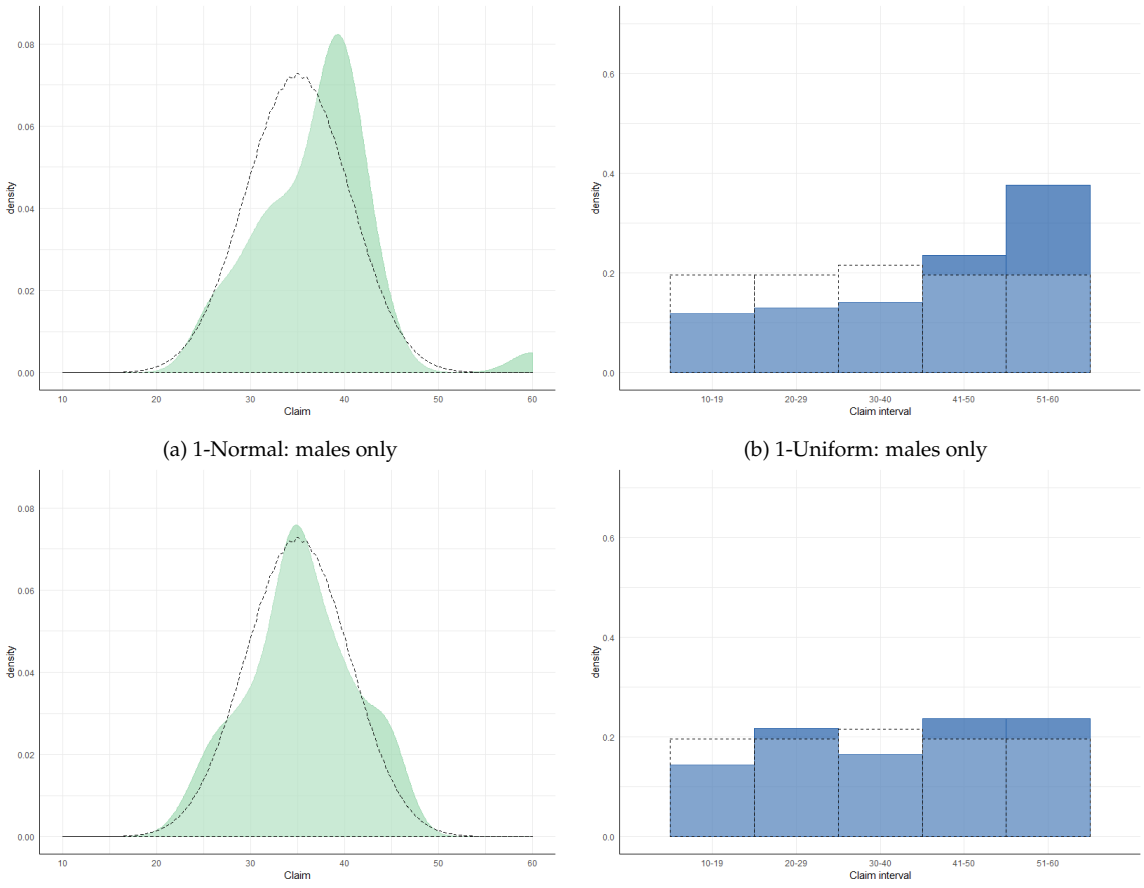
3.5. Gender comparison

In this section we analyse the treatment effects along a gender dimension, showing that the two opposite effects highlighted above have a clear gender characterization.

Reputation concerns. In *1-Uniform* the average claims are equal to 42.14 and to 37.68 for males and females, respectively, and they differ significantly (MW, $p = 0.038$). In contrast, no significant difference is detected in *1-Normal* (MW, $p = 0.239$), where means amount to 36.97 (males) and 35.33 (females).

Figure 6 reports the distribution of claims by gender in the two conditions. It is immediately evident that in these two treatments cheating is a male phenomenon. Very high claims

in *1-Uniform* as well as partial lying around 40 in *1-Normal* pertain almost entirely to male participants.



(a) 1-Normal: males only

(b) 1-Uniform: males only

(c) 1-Normal: females only

(d) 1-Uniform: females only

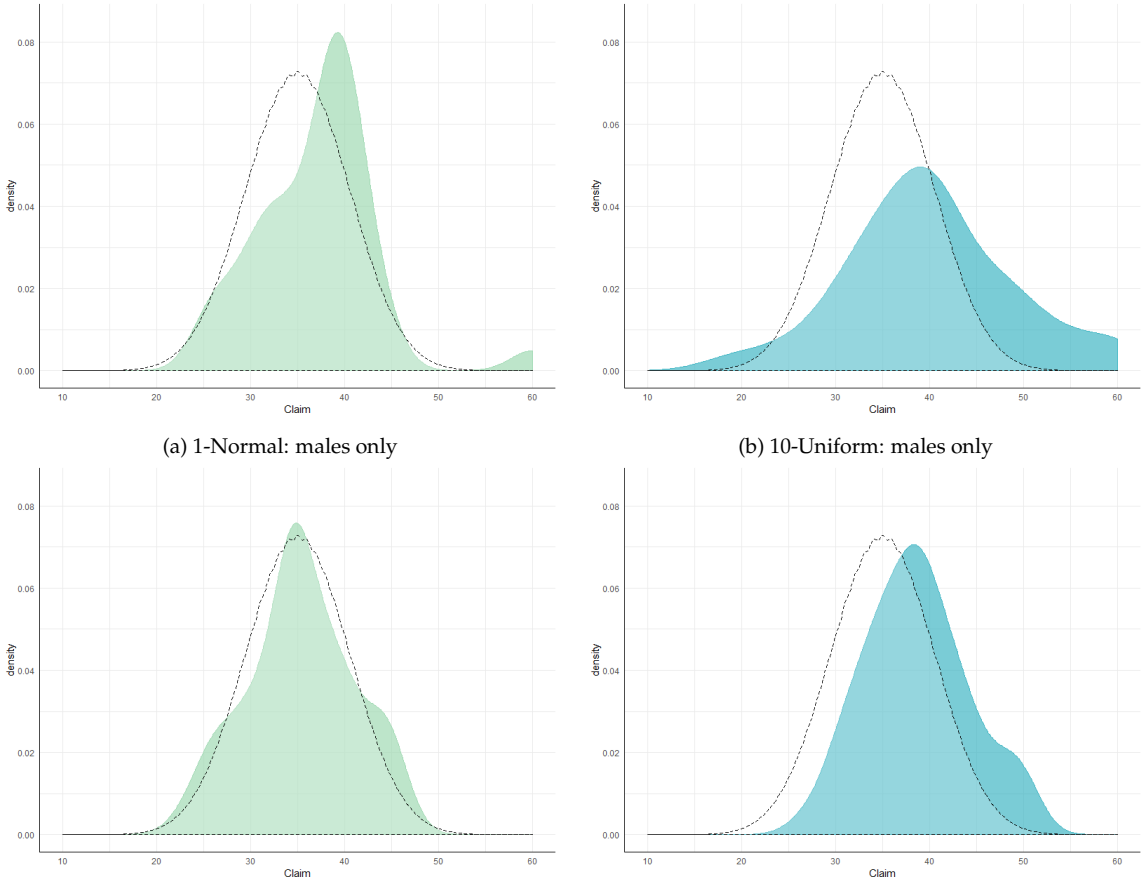
Note: we report kernel density estimates for *1-Normal* (Panels 6a and 6c) and histograms (collapsed in 5 bins) for *1-Uniform* (Panels 6b and 6d). Dotted lines represent the expected frequencies.

Figure 6: Reputation concerns: Frequencies of claims by gender.

The observed distribution of females' claims is instead fairly close to the expected one even in *1-Uniform*. This evidence can be rationalized by the fact that females exhibit a sufficiently high intrinsic cost of cheating to inhibit misreporting even when reputation concerns are less relevant. Consequently, the introduction of reputation concerns in *1-Normal* has no room by construction to substantially decrease their reported outcome. In contrast, males' claims significantly decrease when reputation concerns matter (MW, $p = 0.011$). High claims in excess are widespread in *1-Uniform* and they are substituted by partial lying in *1-Normal*. These results are confirmed analyzing the fraction of cheaters in the range 45–58

(Table C.5 in AppendixC).

Multiple outcomes. The average claim is 40.51 for males and 38.65 for females in the *10-Uniform* treatment (MW, $p = 0.461$). The average reported outcomes significantly increase by a similar amount (about 3.1 and 3.3 ECU, respectively) as compared to *1-Normal*. Generating the outcome through multiple random draws induces partial lying. Figure 7a shows that the distribution of observed claims shifts to the right in both cases, although the female distribution exhibits a smaller variance.



(a) 1-Normal: males only

(b) 10-Uniform: males only

(c) 1-Normal: females only

(d) 10-Uniform: females only

Note: we report kernel density estimates. Dotted lines represent the expected frequencies and therefore distance between solid and dotted lines measures the density of cheaters.

Figure 7: Frequencies of claims by gender and treatment.

Table C.6 in AppendixC emphasizes a significantly higher fraction of cheaters among females in the interval 30–41. Males display a pronounced inclination to cheat for a wider range of outcomes. The difference across conditions is significant, only starting from values

around 40, because partial lying is also present in *1-Normal*. Among males we also detect a substantial fraction of cheaters for (unlikely) claims between 40 and 47.

Result 5. *i)* Females exhibit a high intrinsic costs of cheating that makes reputation concerns irrelevant;

ii) Males respond to reputation concerns with partial lying;

iii) Multiple outcomes increase the incidence of partial lying for both genders.

Our experimental setting helps to provide a unifying framework for the results in the literature about gender differences in cheating. As already mentioned, Abeler et al. (2019) emphasize as a robust finding that females report less than males, but their evidence is mainly based on studies comparable with our *1-Uniform* treatment. Restricting the analysis on studies entailing repeated cheating tasks we find that gender differences are way less pronounced. Table C.7 in AppendixC shows that only two studies out of 12 (Abeler et al., 2014; Cohn and Maréchal, 2015) report clear gender differences in a repeated framework.[25] Another two works (Barfort et al., 2015; Halevy et al., 2014) report weak evidence supporting the claim that females have a lower inclination to cheat. In most cases (7 out of 12) the authors report no significant differences, while in Fosgaard (2013) females claim significantly more than males.

4. Conclusion

When workers' productivity is jointly determined by effort and by a stochastic component, observing a single outcome can be uninformative. The monetary incentives of a misconduct must be weighted only against the self-image cost of dishonest behavior. As long as the stochastic component may account for the final outcome over and above the worker's effort there is much room for shirking. Multiple outcomes are instead more informative thanks to the repeated observations, provided that the distribution of the stochastic component is known both to the principal and to the agent. In this case, a misconduct also implies the additional reputation cost of being detected as a liar. The literature suggests that

[25] For the sake of comparability with our results we exclude the studies requiring the subjects to report only a single outcome out of a number of repetitions.

observing multiple outcomes can also backfire, however, by lowering the intrinsic cost of cheating.

This paper investigates experimentally how the agents' inclination to cheat is affected by reporting a single vs. multiple private information outcomes. We analyze these two alternatives by administering a properly designed cheating task in which subjects hold private information about the realization of a stochastic event. Two treatments identify the social and the intrinsic cost of cheating implied by reporting multiple outcomes, while keeping the monetary incentives constant. Reputation concerns are measured changing the underlying probability distribution, while the intrinsic cost is manipulated through the number of realizations observed *ceteris paribus*.

We find that reporting the sum of multiple outcomes rather than a single realization induces two opposite effects that cancel out on average. On the one hand, it induces reputation concerns that dramatically reduce the occurrence of severe misreporting. Knowing that extreme outcomes are unlikely, subjects tend to avoid the social-image cost of making implausible claims. A parallel example in the labor market predicts that workers should be less likely to fully exploit their private information when observed over a sufficiently long period. The reason is that systematically claiming that the stochastic component negatively affects their productivity is not credible and would expose them as liars.

On the other hand, reporting the sum of multiple outcomes seems to decrease the intrinsic cost of cheating, leading to a more pervasive inclination to adjust slightly upward the true realization. Moral self-licensing is the explanation that better fits our results. Behaving in a honest way most of the time helps to preserve one's self-image even when behaving in an opportunistic way occasionally. This second effect is small in magnitude but widespread so that it compensates the disappearance of cheating by large amounts observed in the other treatment. The implication in the labor market is that workers should be more likely to misreport their private information when observed over a long period, making slightly adjusted but plausible claims that do not signal opportunistic behavior.

Our results also reveal interesting gender differences. When reputation concerns are less relevant cheaters are typically males. In contrast, females' claims do not sizably differ from the theoretical distribution, signaling a higher intrinsic cost of cheating. At the same

time, males and females are equally prone to moral self-licensing. Given that (male-driven) severe cheating virtually disappears in the presence of reputation concerns, males and females turn out to be statistically indistinguishable when reporting multiple outcomes. Our findings suggest that observing multiple outcomes may be successful in reducing shirking among males, but it can backfire with females. The effectiveness of this measure to tackle information asymmetries can in the end depend on the gender composition of the workforce.

References

Abeler, J., Becker, A., Falk, A., 2014. Representative evidence on lying costs. Journal of Public Economics 113, 96–104.

Abeler, J., Nosenzo, D., Raymond, C., 2019. Preferences for truth-telling. Econometrica 87, 1115–1153.

Ariely, D., Garcia-Rada, X., Hornuf, L., Mann, H., 2015. The (true) legacy of two really existing economic systems .

Barfort, S., Harmon, N.A., Hjorth, F.G., Olsen, A.L., 2015. Dishonesty and selection into public service in denmark: Who runs the world's least corrupt public sector? .

Barron, K., 2019. Lying to appear honest. Discussion Paper 2019-307. WZB.

Blanken, I., van de Ven, N., Zeelenberg, M., 2015. A meta-analytic review of moral licensing. Personality and Social Psychology Bulletin 41, 540–558.

Blanken, I., van de Ven, N., Zeelenberg, M., Meijers, M.H., 2014. Three attempts to replicate the moral licensing effect. Social Psychology 45, 232.

Braun, S., Hornuf, L., 2015. Leadership and persistency in spontaneous dishonesty .

Clot, S., Grolleau, G., Ibanez, L., 2014. Smug alert! exploring self-licensing behavior in a cheating game. Economics Letters 123, 191–194.

Cohn, A., Fehr, E., Maréchal, M.A., 2014. Business culture and dishonesty in the banking industry. Nature 516, 86.

Cohn, A., Maréchal, M.A., 2015. Laboratory measure of cheating predicts misbehavior at school .

Conrads, J., 2014. The effect of communication channels on lying. Technical Report. Cologne Graduate School in Management, Economics and Social Sciences.

Crosetto, P., Filippin, A., 2013. The 'bomb' risk elicitation task. Journal of Risk and Uncertainty 47, 31–65.

Dufwenberg, M., Dufwenberg, M.A., 2018. Lies in disguise–a theoretical analysis of cheating. Journal of Economic Theory 175, 248–264.

Effron, D.A., Conway, P., 2015. When virtue leads to villainy: Advances in research on moral self-licensing. Current Opinion in Psychology 6, 32–35.

Ermisch, J., Gambetta, D., 2006. People's Trust: The Design of a Survey-Based Experiment. resreport. IZA Discussion Paper No. 2216.

Ermisch, J., Gambetta, D., Laurie, H., Siedler, T., Noah Uhrig, S., 2009. Measuring people's trust. Journal of the Royal Statistical Society: Series A (Statistics in Society) 172, 749–769.

Falk, A., Fehr, E., 2003. Why labour market experiments? Labour Economics 10, 399–406.

Falk, A., Kosfeld, M., 2006. The hidden costs of control. American Economic Review 96, 1611–1630.

Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171–178.

Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise - an experimental study on cheating. Journal of the European Economic Association 11, 525–547.

Fosgaard, T.R., 2013. Asymmetric default bias in dishonesty–how defaults work but only when in one's favor. University of Copenhagen Discussion Paper .

Garbarino, E., Slonim, R., Villeval, M.C., 2018. A method to estimate mean lying rates and their full distribution. Journal of the Economic Science Association 4, 136–150.

Geraldes, D., Heinicke, F., Rosenkranz, S., 2019. Lying on two dimensions and moral spillovers. Available at SSRN 3381277 .

Gino, F., Ariely, D., 2012. The dark side of creativity: original thinkers can be more dishonest. Journal of personality and social psychology 102, 445.

Gioia, F., 2016. Peer effects on risk behaviour: the importance of group identity. Experimental Economics , 1–30.

Gneezy, U., Kajackaite, A., Sobel, J., 2018. Lying aversion and the size of the lie. American Economic Review 108, 419–53.

Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with orsee. Journal of the Economic Science Association 1, 114–125.

Halevy, R., Shalvi, S., Verschuere, B., 2014. Being honest about dishonesty: Correlating self-reports and actual lying. Human Communication Research 40, 54–72.

Hitchcock, D.B., 2009. Yates and contingency tables: 75 years later. Electron. J. Hist. Probab. Stat 5, 1–14.

Hofmann, W., Wisneski, D.C., Brandt, M.J., Skitka, L.J., 2014. Morality in everyday life. Science 345, 1340–1343.

Ichino, A., Muehlheusser, G., 2008. How often should you open the door?: Optimal monitoring to screen heterogeneous agents. Journal of Economic Behavior & Organization 67, 820–831.

Khalmetski, K., Sliwka, D., 2019. Disguising lies — image concerns and partial lying in cheating games. American Economic Journal: Microeconomics 11, 79–110.

Kroher, M., Wolbring, T., 2015. Social control, social learning, and cheating: Evidence from lab and online experiments on dishonesty. Social Science Research 53, 311–324.

Mann, H., Garcia-Rada, X., Hornuf, L., Tafurt, J., Ariely, D., 2016. Cut from the same cloth: Similarly dishonest individuals across countries. Journal of Cross-Cultural Psychology 47, 858–874.

Monin, B., Miller, D.T., 2001. Moral credentials and the expression of prejudice. Journal of personality and social psychology 81, 33.

Pittarello, A., Leib, M., Gordon-Hecker, T., Shalvi, S., 2015. Justifications shape ethical blind spots. Psychological Science 26, 794–804.

Ploner, M., Regner, T., 2013. Self-image and moral balancing: An experimental analysis. Journal of Economic Behavior & Organization 93, 374–383.

Pollmann, M.M., Potters, J., Trautmann, S.T., 2014. Risk taking by agents: The role of ex-ante and ex-post accountability. Economics Letters 123, 387–390.

de Quidt, J., Haushofer, J., Roth, C., 2018. Measuring and bounding experimenter demand. American Economic Review 108, 3266–3302.

Sappington, D.E., 1991. Incentives in principal-agent relationships. Journal of economic Perspectives 5, 45–66.

Shalvi, S., Dana, J., Handgraaf, M.J., De Dreu, C.K., 2011. Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. Organizational Behavior and Human Decision Processes 115, 181–190.

Suri, S., Goldstein, D.G., Mason, W.A., 2011. Honesty in an online labor market. Human Computation 11.

Walker, B., 2000. Monitoring and motivation in principal-agent relationships: some issues in the case of local authority services. Scottish Journal of Political Economy 47, 525–549.

Whynes, D.K., 1993. Can performance monitoring solve the public services' principal-agent problem? Scottish Journal of Political Economy 40, 434–446.

Zizzo, D.J., 2010. Experimenter demand effects in economic experiments. Experimental Economics 13, 75–98.

AppendixA. Knowledge of joint probabilities

In a pre-experimental questionnaire we ask participants what is the probability to obtain a sum equal to 2 and to 7 rolling two dice.
The alternatives in the first case are:

a) *1/6*
b) *1/12*
c) *1/36*
d) *I don't know*

The alternatives provided for the likelihood of obtaining a sum equal to 7 are:

a) *1/6*
b) *1/12*
c) *1/36*
d) *I don't know*
e) *I don't know exactly, but smaller than obtaining a 2*
f) *I don't know exactly, but the same as obtaining a 2*
g) *I don't know exactly, but larger than obtaining a 2*

We define a participant as being unfamiliar with joint probabilities if the answer to the second question is either a) or g). Regardless of the correctness of their answers, participants receive the solution of the two questions and a brief explanation of why it is more likely to obtain a seven than a two.

AppendixB. Instructions (translation from Italian)

Welcome! Thank you for participating in this experiment.

The instructions are displayed on your screen and they will also be read aloud. Please, follow carefully the instructions since your final earning will depend on the decisions taken during the experiment. Thus, it is in your personal interest to have a thorough comprehension of the experiment.

If you have any question at any time please raise your hand and wait for an experimenter to come to your desk and answer it privately. During the whole experiment it is prohibited to talk with other participants.

The experiment starts with a short questionnaire and will then continue with three Phases. We start reading the instructions of Phase 1. You will then receive the instructions for Phase 2 and 3 in due course. In each Phase, you will earn Experimental Currency Units (ECU). The earnings of each Phase are independent and you will be paid the sum of your earnings in every Phase. The total amount in ECU will be converted using the exchange rate

20 ECU = 1 EURO

and paid privately in cash at the end of the experiment. For your participation you will also receive an additional amount of 2.5 euro.

All the decisions made during the experiment are anonymous. At the end of the experiment we will call you individually using the number randomly drawn when entering the lab and we will give you the total amount earned in cash.

PHASE 1

A grid with 100 boxes will appear on your computer.

Your task is to choose how many boxes to collect. So, you will be asked to choose a number between 1 and 100. Boxes will be collected in numerical order, starting from the box in the top left corner of the grid.

Every box collected is worth 1 ECU. However, this earning is only potential since one of this boxes hides a time bomb. You do not know where this bomb lies. You only know that the bomb can be in any box with equal probability. After choices have been made, the computer will randomly determine which box contains the bomb. This random draw is made at the individual level, thus the box containing the bomb can differ for every participant.

If the bomb is located in a box that you did not collect — i.e. the number of boxes you choose is smaller than the number of the box containing the bomb — you will earn 1 ECU for each box collected.

In contrast, if you happen to collect the box where the bomb is located — i.e. if the number of boxes you choose is greater than or equal to the number of the box containing the bomb — the bomb by exploding will destroy the earnings: thus, you will earn zero in Phase 1.

PHASE 2

In this phase two participants interacts: *Player A* and *Player B*. The Figure on your screen (see Figure 2) summarizes all possible outcomes of this game.

At the beginning, *Player A* is endowed with 10 ECU and must decide among two alernatives, called OPTION 1 and OPTION 2.

OPTION 1

Player A keeps the 10 ECU and nothing is passed to *Player B*. The game ends without *Player B* taking any decision and earnings for Phase 2 will be 10 ECU for *Player A* and 0 ECU for *Player B*, respectively.

OPTION 2

Player A passes 10 ECU to *Player B*. The 10 ECU will be multiplied by a factor of 4, thus *Player B* will receive 40 ECU.

In this case *Player B* must decide among OPTION 3 and OPTION 4:

OPTION 3

Player B keeps the 40 ECU and nothing is passed back to *Player A*. Earnings for Phase 2 will be 0 ECU for *Player A* and 40 ECU for *Player B*, respectively.

OPTION 4

Player B keeps the 18 ECU and passes back 22 ECU to *Player A*. Earnings for Phase 2 will be 22 ECU for *Player A* and 18 ECU for *Player B*, respectively.

After the choices have been made, the computer will randomly form groups of two players, randomly assigning a role (*Player A* and *Player B*) to each player. For this reason, you will be asked to make a decision both as *Player A* and as *Player B*. Once pairs have been formed, earnings will be automatically computed according to the decision made by each player in the assigned role.

PHASE 3

In this phase you are required to recognize the value and the issuing country of several euro coins. As shown in the example below on the left part of your screen, a table with several coins will appear. The corresponding value and the issuing country is reported below each coin. On the right part of your screen you will see one coin randomly selected from the table.

Your task is to identify both the value and the issuing country of the selected coin. You have to select the value and the country from the corresponding list and then confirm your choice. You will then be notified about the correctness of your response. Your task will then continue with the identification of another coin randomly chosen from a different table.

Figure B.8: Screenshot of the coin task

You have 15 minutes to identify as many coins as possible, with a maximum of 20. There are 60 tables available, and therefore you can successfully complete the task even if you make some mistakes. At the end of each round you will also be informed about your total score up to that point. Before starting with the real task, you have the opportunity to practice for 60 seconds. The coins identified during this trial period do not count for your score.

Your earnings in this Phase of the experiment are composed by two parts:

1. A deterministic component linked to your score: you will receive 2 ECU for each coin correctly identified.

2. A variable component unrelated to the number of coins correctly identified.

1-Uniform: This part of your compensation is worth between 10 and 60 ECU, and will be privately determined randomly drawing one number. On your desk, there is bag containing 51 numbered tokens, from 10 to 60. Each number represents a value in ECU (for instance, number 10 means 10 ECU, number 11 means 11 ECU and so on).

Draw only one number from this bag and report it on your computer. The reported number, which must be by construction between 10 and 60, represents the variable part of your earnings in this phase.

10-Uniform: This part of your compensation is worth between 10 and 60 ECU, and will be privately determined randomly drawing one number ten times. On your desk, there is bag containing 6 numbered tokens, from 1 to 6. Each number represents a value in ECU (for instance, number 1 means 1 ECU, number 2 means 2 ECU and so on).
Draw one number from this bag for ten times and report the sum of the ten draws on your computer. The reported number, which must be by construction between 10 and 60, represents the variable part of your earnings in this phase. If helpful, feel free to use a sheet of paper or your smartphone to write down the numbers drawn and/or to compute the sum.
ATTENTION: each draw must be done with all the 6 numbers in the bag. Hence, remember to put the number back into the bag before proceeding with the subsequent draw. Repeat the procedure for exactly ten times.

1-Normal: This part of your compensation is worth between 10 and 60 ECU, and will be privately determined randomly drawing one number. You will access an online random generator of numbers with your smartphone using the QR code that will be displayed with the projector.
The random generator will draw ten times from a virtual bag containing 6 numbered tokens, from 1 to 6. Note that each draw is made with all the 6 numbers in the virtual bag and that draws are made separately for each of you. The random generator will show you only the sum of the outcomes obtained. You have to report this number on your computer. The reported number, which must be by construction between 10 and 60, represents the variable part of your earnings in this phase.

You will start with the identification of the coins, then you will proceed with the determination of the variable part of your earnings.

AppendixC. Additional results

Table C.5: Reputation concerns by gender

PANEL A: FEMALES

	Threshold													
	45	46	47	48	49	50	51	52	53	54	55	56	57	58
% of cheaters														
1-Normal	0.77	0	0	0	0	0	0	0	0	0	0	0	0	0
1-Uniform	7.82	8.45	8.08	8.76	9.48	6.11	6.70	5.47	4.36	4.84	3.73	4.21	4.82	6.62
p-value	0.169	0.081	0.089	0.069	0.053	0.169	0.136	0.209	0.312	0.259	0.390	0.323	0.255	0.132

PANEL B: MALES

	Threshold													
	45	46	47	48	49	50	51	52	53	54	55	56	57	58
% of cheaters														
1-Normal	1.43	1.76	2.07	2.32	2.51	2.63	2.70	2.74	2.76	2.77	2.77	2.77	2.77	2.77
1-Uniform	29.64	31.58	31.82	32.05	27.76	22.21	21.22	21.68	22.12	21.20	19.02	18.21	16.18	10.56
p-value	0.002	0.001	0.001	0.001	0.004	0.019	0.025	0.022	0.02	0.025	0.042	0.050	0.081	0.291

Note: Every column represents a binary partition of the claims between those lower than or equal to the indicated threshold from those strictly higher. The fraction of subjects lying for every threshold is estimated following Garbarino et al. (2018). The p-value refers to a Chi-square test for the difference in the proportions across treatments.

Table C.6: Observing multiple outcomes by gender

PANEL A: FEMALES

	Threshold													
	30	31	32	33	34	35	36	37	38	39	40	41	42	43
% of cheaters														
1-Normal	11.78	15.92	21.52	20.64	14.84	4.72	6.65	7.83	6.12	6.63	5.39	5.01	5.60	7.62
10-Uniform	67.32	58.41	53.87	56.93	50.36	38.02	42.14	39.01	30.44	25.16	17.47	21.28	15.24	13.23
p-value	0.030	0.059	0.115	0.040	0.020	0.006	0.003	0.005	0.013	0.044	0.148	0.044	0.228	0.566

	Threshold						
	44	45	46	47	48	49	50-59
% of cheaters							
1-Normal	2.54	0.77	0	0	0	0	0
10-Uniform	10.74	9.92	8.74	7.27	7.66	3.8	0
p-value	0.211	0.104	0.088	0.143	0.124	0.483	.

PANEL B: MALES

	Threshold													
	30	31	32	33	34	35	36	37	38	39	40	41	42	43
% of cheaters														
1-Normal	32.12	20.89	30.58	28.48	22.57	31.43	34.91	37.51	24.20	16.03	8.84	3.27	2.85	3.85
10-Uniform	35.77	38.66	49.92	45.60	43.03	41.32	44.20	46.21	40.44	38.30	32.79	27.23	24.28	23.65
p-value	1	0.716	0.578	0.564	0.358	0.747	0.736	0.732	0.317	0.103	0.044	0.019	0.027	0.041

	Threshold											
	44	45	46	47	48	49	50	51	52	53	54	55-59
% of cheaters												
1-Normal	1.12	1.43	1.76	2.07	2.32	2.51	2.63	2.70	2.74	2.76	2.77	2.77
10-Uniform	22.54	23.66	21.81	19.74	17.49	12.56	10.11	10.19	10.22	10.24	7.68	5.12
p-value	0.016	0.013	0.023	0.041	0.075	0.233	0.400	0.401	0.403	0.404	0.666	1

Note: Every column represents a binary partition of the claims between those lower than or equal to the indicated threshold from those strictly higher. The fraction of subjects lying for every threshold is estimated following Garbarino et al. (2018). The p-value refers to a Chi-square test for the difference in the proportions across treatments.

Table C.7: Survey of gender differences in repeated cheating tasks

Paper	Task	Number of repetitions	Number of Female	Number of Male	Method of Analysis
		Panel A: Females report significantly less			
Abeler et al. (2014)	coin	4	262	182	Logit, est. = –0.345, p = 0.027 Tobit, est. = –0.371, p = 0.046
Barfort et al. (2015)	die	40	400	462	Linear regression, est. = –0.061, $p<0.05$[1]
Cohn and Maréchal (2015)	coin	10	70	92	OLS, est. = –1.028, p = 0.000[2]
Halevy et al. (2014)	die	180	37	14	In text, not with data
		Panel B: No significant difference			
Ariely et al. (2015)	die	40	*259*		Probit, est. = 0.035, p = 0.671[2]
Braun and Hornuf (2015)	die	10	218	117	t-test, p = 0.1809[3]
Cohn et al. (2014)	coin	10	*128*		Probit, est. = 0.027, $p>0.1$[2]
Conrads (2014)	coin	4	121	125	Logit, est. = –0.074, $p>0.1$[2]
Kroher and Wolbring (2015)	die	2	*222*		OLS, est. = –0.135, p = 0.619[2]
Mann et al. (2016)	die	20	1157	1022	Linear reg. (student), est. = –0.002, p = 0.857 Linear reg. (not student), est. = 0.006, p = 0.526
Suri et al. (2011)	die	30	84	149	GLM, est. = 0.020, $p>0.1$
		Panel C: Females report significantly more			
Fosgaard (2013)	die	10	*420*		Tobit, est. = 1.403, $p<0.01$[2]

Notes: The table reports the device used in the cheating task, the number of repetitions, the numerosity of the sample (divided in males and females when available), the method of analysis implemented to test gender differences, the estimate of the gender difference (negative values mean that females cheat less than males), and the p-value (if reported).
[1] The difference is significant only when gender is analysed without further controls.
[2] We report only the specification with the largest set of controls, or with more observations.
[3] Mann–Whitney U-test detects a weakly significant effect of females cheating more (p = 0.073).