Nudging Enforcers: How Norm Perceptions and Motives for Lying Shape Sanctions

ETH zürich

Center for Law & Economics Working Paper Series

Number 13/2020

Nudging Enforcers: How Norm Perceptions and Motives for Lying Shape Sanctions

Eugen Dimant
Tobias Gesche

October 2020

The most recent version of the working paper can always be downloaded following this link.

Nudging Enforcers: How Norm Perceptions and Motives for Lying Shape Sanctions

Eugen Dimant[1,2], Tobias Gesche[3]

[1] *University of Pennsylvania, Center for Social Norms and Behavioral Dynamics*
[2] *CESifo, Munich*
[3] *Center for Law & Economics, ETH Zurich*

Abstract

The enforcement of social norms is the fabric of a functioning society. Through the lens of two experiments, we examine how motives for lying and norm perceptions steer enforcement. Our contribution is to investigate the extent to which norm breaches are sanctioned, how norm-nudges affect the observed punishment behavior, and how the enforcement is linked to norm perceptions. Using a representative U.S. sample, Experiment 1 provides robust evidence that norm-enforcement is not only sensitive to the extent of the observed transgression (= size of the lie) but also to its consequences (= whether the lie remedies or creates payoff inequalities). We also find norm enforcers to be sensitive to different norm-nudges that convey social information about actual lying behavior or its social disapproval. To explain the punishment patterns, Experiment 2 examines how norms are perceived across different transgressions and how norm-nudges change these perceptions. We observe a malleability of social norm perceptions: norm nudges are most effective when pre-existing norms are vague. Importantly, we also find that punishment patterns in the first experiment closely follow these norm perceptions. Our findings suggest that norm enforcement can be successfully nudged and thus represent an expedient alternative to standard incentive-based interventions.

Keywords: Lying, Norm-Nudges, Nudging, Punishment, Social Norms
JEL: B41, D01, D9

*We would like to thank Giuseppe Danese, Mathias Ekström, Jana Freundt, Martin Kocher, Nils Köbis, Nikos Nikiforakis, Jan Schmitz, and Weiwei Tasch for helpful comments and Tanner Nichols for his support with the data collection for the second experiment. We acknowledge financial support from the German Research Foundation (DFG) under Germany's Excellence Strategy — EXC 2126/1– 390838866.

Email addresses: edimant@sas.upenn.edu (Eugen Dimant), tgesche@ethz.ch (Tobias Gesche)

This version: October 30, 2020

Introduction

Social norms are ubiquitous in human societies. They inform both our individual behavior and our interactions with others in a variety of socially and economically interesting domains such as collective action, altruistic sharing, and deviance.[1] The emerging consensus in the literature is that norm compliance can erode quickly and that enforcement is crucial to sustain compliance.[2]

One particularly promising approach to achieve behavior change has emerged in the form of nudging (Thaler and Sunstein, 2009). The existing nudge literature typically focuses on interventions that aim at directly changing behavior (e.g., Benartzi et al., 2017; DellaVigna and Linos, 2020). Recently, a growing body of literature has utilized so-called "norm-nudges": nudges that attempt to change behavior by eliciting and changing existing social norms through the manipulation of social expectations (for a theoretical conceptualization see Bicchieri and Dimant, 2019, for applications see, Hallsworth et al., 2017; Bhanot, 2018; Damgaard and Gravert, 2018; Allcott and Kessler, 2019; Bott et al., 2020; Bursztyn et al., 2020a,b; Dimant et al., 2020).

In this paper, we provide a complementary approach and utilize the power of norm-nudges to study its effects on those who *enforce* norm compliance of others. With that, we show how such enforcement patterns can be altered using norm-nudges – a perspective that is rather understudied in the existing literature. More specifically, across various decision scenarios, we examine interventions that use social information to affect norm enforcement through peers. In doing so, we connect three literature streams: the study of transgressions in the context of lying, the enforcement of norms through punishment, and the perception of social norms. To the best of our knowledge, this approach is the first to directly investigate ways to nudge norm enforcement as alternative to traditional policy interventions.

To accomplish this, we first investigate whether and how individuals differentiate punishment behavior based on different motives that lead to non-compliance with an existing norm (e.g., lying that achieves an unfair advantage over another person versus lying that restores equal chances). We then capitalize on the concept of norm-nudging. We do so by investigating the extent to which punishment decisions are sensitive to norm information with respect to what others *do* (empirical information) or what people *approve of doing* (normative information) - an approach we borrow directly from the affluent interdisciplinary social norms literature (e.g., Cialdini et al., 1990; Bicchieri, 2006; Tankard and Paluck, 2016; Young, 2016).

We run two well-powered and pre-registered controlled experiments with diverse subject pools that study norm enforcement and norm perceptions in conjunction. In Experiment 1, we

[1] See, e.g., Ostrom 2000; Kimbrough and Vostroknutov 2016; Albrecht et al. 2018; Fehr and Schurtenberger 2018; Bolton et al. 2019; Choi et al. 2019; Bott et al. 2020; Bicchieri et al. 2020a; Dimant 2020.

[2] See, e.g., Fehr and Gachter, 2000; Fehr and Fischbacher, 2004; Herrmann et al., 2008; Nikiforakis and Normann, 2008; Sutter et al., 2010; Balafoutas and Nikiforakis, 2012; Balafoutas et al., 2014; Fehr and Schurtenberger, 2018; Bolton et al., 2019; Bicchieri et al., 2020c; Brouwer et al., 2020. Also, punishment in (non-)monetary forms, such as shaming, are powerful tools to increase compliance, trust, and cooperation (e.g., Coleman, 1994; Coricelli et al., 2010; Dickinson et al., 2015; Heffner and FeldmanHall, 2019).

consider a situation where two participants solve a task in order to qualify for having the chance to win an indivisible price. One participant of each pair is then given the opportunity to lie. In our context, larger lies increase a participant's chance to win the fixed-size price. Using a representative U.S. sample of 1,240 participants acting as third-party punishers, we study how such behavior is sanctioned. We find that punishment increases in a linear fashion with the size of the lie. We also find that the equity nature of the lie matters: if the lie helps the lying participants to achieve equal chances in winning the price, then lying is less severely punished compared to when lying would improve one's winning chances relative to the victim. Finally, we observe that punishers react to norm-nudges and punish more when lying behavior is in conflict with the information provided in the norm nudge. This is independent of whether the information points to mere normative statements of others (i.e., disapproval of lying) or toward others' actual behavior (e.g., that others did not lie).[3] Thus, our results indicate that both types of norm-nudges affect punishment in similar ways.

Given the finding of our first experiment, we then ask: Are the observed patterns in norm enforcement rooted in norm perceptions that are consistent with the observed punishment patterns? If so, how would this help us better understand the variation in punishment that we have observed in Experiment 1? To answer this question, we capitalize on a separate set of 1,519 participants in Experiment 2. We use an incentive-compatible method to examine how the provision of norm information, the equity nature of the lie, and the interplay of these dimensions affect third parties' perception of social norms (i.e., what one thinks others consider appropriate behavior). We elicit social norm perceptions that are consistent with the findings from our first experiment: lying which generates an advantage for the liar, relative to the other participant, is perceived to be less appropriate than lying that yields more equal chances for both participants. Also reflecting on the first experiment's findings, we observe that providing norm information increases the perceived inappropriateness of lying, independent of the source of this information (normative statements that others perceive lying as inappropriate versus information that a majority of others did not lie, when given the chance). Examining this in more detail, we find that norm information works particularly well in situations where norms are otherwise perceived to be more lenient (i.e., when lying results in more equal chances).[4]

Taken together, our experiments not only allow us to investigate *when* and *how* norm breaches are sanctioned; but also show us that variations in norm enforcement are consistent with variations in the perception of social norms. Intuitively, this suggests that regulatory (top-down)

[3] We base this information on two auxiliary experiment. In one, a majority said that lying is not the right thing to do for participants in Experiment 1 (=normative information). In another auxiliary experiment, which resembles our Experiment 1, a majority of participants who could lie did not not do so (=empirical information).

[4] Our results also relate to the existing research examining how equity concerns, social perception, and the size of the lie affect dishonest behavior (see, e.g., Gino and Pierce, 2010; John et al., 2014; Dufwenberg and Dufwenberg, 2018; Gneezy et al., 2018; Abeler et al., 2019). Whereas this literature typically focuses on the factors determining the *intrinsic* costs of lying, we examine how these factors shape the *extrinsic* costs via punishment.

interventions implemented to change behavior can be complemented by social (bottom-up) enforcement through informal norm-nudging; at least where social norms are more clearly defined, transgressions are observable, and when they can be sanctioned.

Experiment 1: Norm Information and Norm Enforcement

Data collection and procedural details

Our experiment consists of two parts: a **pre-experiment** and a **main experiment**. The former was necessary to provide truthful data in the latter and elicit punishment behavior in an incentive-compatible manner. In the pre-experiment, some participants could improve their earnings by reporting dishonestly. In the main experiment, everyone was a "punisher", who could sanction dishonest behavior.

Behavior in a pre-experiment was collected from n=170 participants through Amazon Mechanical Turk (MTurk).[5] Participants were randomly assigned to one of two roles, *active* or *passive* players. *Active* players could then lie in return for a monetary incentive, thereby causing a cost on the *passive* player (see details below).[6] Punisher behavior in the main experiment was collected from n=1,240 participants who were recruited via a professional market research firm that obtained a diverse sample representative of the US working-age population. This allows us to draw generalizable inferences with respect to policy implications (see Levitt and List, 2007; Exadaktylos et al., 2013).[7] All data were collected in April/May (Experiment 1) and November 2019 (Experiment 2); tables A.2 and A.3 in the appendix provide descriptive characteristics for active players in the main experiment and punishers, respectively. Figure 1 illustrates our experimental design, which we discuss below.

Lying (Pre-Experiment): Subjects in the role of (potential) liars were recruited on MTurk. In addition to a base payment of $0.50, they could earn a bonus during two parts:

- Part 1 consisted of a real-effort counting task (counting the 1's in five matrices with numbers).[8] Participants solved this task in pairs, where both members in each pair had to independently

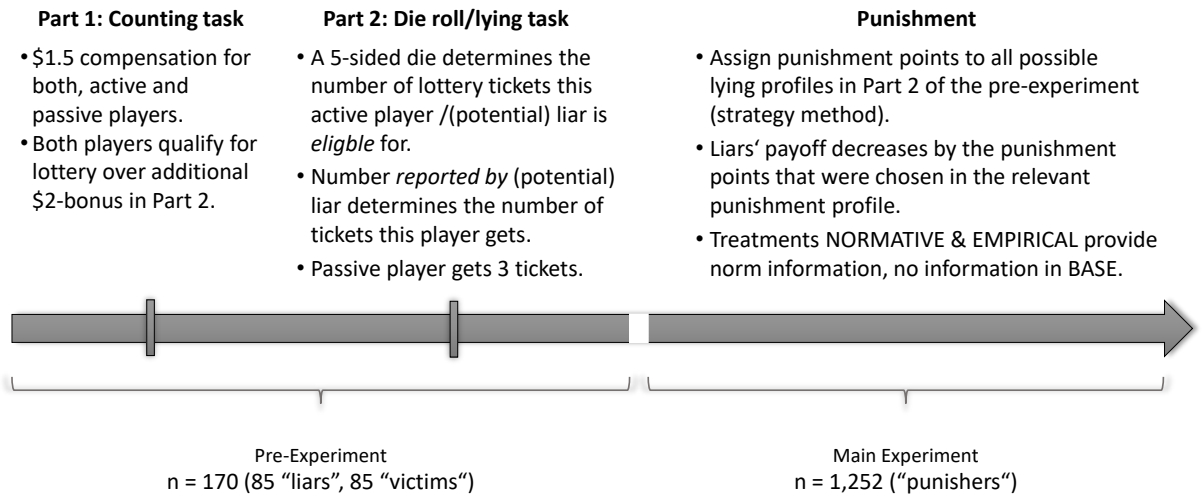
[5] To meet the criteria for robustness and generalizability of MTurk findings (e.g., Arechar et al., 2018; Snowberg and Yariv, 2020), we applied high quality restrictions on the sample: we utilize a combination of CAPTCHAs and comprehensive screening questions to avoid pool contamination. Participants had to be in the US and have an approval rate greater than 99%.

[6] Thus the matching ratio of punishers to active players was roughly 15:1. The instructions did not state or suggest any explicit value for this ratio. However, our exit questionnaire for punishers asked participants to provide an estimate for it. The average value (excluding one extreme outlier) implies a ratio of 1:0.87 – punishers therefore over-perceived the consequences of their decision. We also control for punishers' estimated implementation ratio in our regression analysis but do not find a significant effect in any our specifications.

[7] Data collection was conducted through market research firm Dynata (formerly "Research Now") by a quota-based sample, with the aim of having repetitiveness along age and gender for the US population.

[8] Participants were shown 10×10 matrices in the form of images. Wrong entries (more than +/-1 off the actual solution) required subjects to re-count and wait 45 seconds. The task took about seven minutes.

Figure 1: Design of Experiment 1



complete the same task. For successfully solving this tasked, they each earned a compensation of $1.50. In addition, solving this task also entitled them *both* to winning an additional bonus.

- In Part 2, players learned that one of them could earn an additional bonus of $2.50. To determine the winner, members of each pair were randomly chosen to either be in the *active* or *passive* role. The *passive* player always received 3 virtual lottery tickets by default and had no further decisions to make. The *active* player was entitled to the number of lottery tickets equal to the randomized outcome of a virtual 5-sided die. However, the *active* players were made aware that what they reported (rather than the actual outcome) would determine the number of lottery ticket they would get, thus providing an incentive to inflate the actual result. One ticket was drawn and the player who held it won the $2.50-bonus.

Note that active players had a monetary incentive to exaggerate the die roll's outcome to increase their chances to earn the bonus. Also note that this situation is deliberately designed to be non-strategic: active players cannot justify their behavior by (motivated) belief about the other player's behavior. Nevertheless, we observe lying as their reports are, on average, 29% higher than the actual outcome of the die (Wilcoxon signed rank test: $p<0.001$). In order to avoid deception but still obtain a truthful measure of lying behavior, active players were told after their submission of the report that another participant in the role of a punisher would later observe – and potentially punish – their behavior (amounting to punishment of up to $1.50). Participants were then given the opportunity to revise their initial report. In this case, the revised report was used to determine their punishment in the main experiment.[9]

[9]Since we elicited the behavior of punishers using the strategy method, the actual behavior was irrelevant for our analysis of punishment decisions. We observe that the *threat of punishment* works: The revised reports were, on average, lower by 0.44 than the the initial reports (signed rank test: $p<0.001$). They were, however, still higher

Table 1: Assignment of punishment points (Example)

I want to assign the following number of punishment points if Participant A...

	0	1	2	3	4	5
			*...had a die outcome of "**1**"*			
*...and reported "**2**"*	◉	○	○	○	○	○
*...and reported "**3**"*	○	◉	○	○	○	○
*...and reported "**4**"*	○	◉	○	○	○	○
*...and reported "**5**"*	○	◉	○	○	○	○
			*...had a die outcome of "**2**"*			
	0	1	2	3	4	5
*...and reported "**3**"*	○	○	◉	○	○	○
*...and reported "**4**"*	○	○	◉	○	○	○
*...and reported "**5**"*	○	○	◉	○	○	○
			*...had a die outcome of "**3**"*			
	0	1	2	3	4	5
*...and reported "**4**"*	○	○	○	◉	○	○
*...and reported "**5**"*	○	○	○	○	◉	○
			*...had a die outcome of "**4**"*			
	0	1	2	3	4	5
*...and reported "**5**"*	○	○	○	○	○	◉

Notes: The order determining whether punishment scenarios were presented by the die's actual outcome in an increasing manner (shown here) or decreasing manner was randomized. For punishers, an active player was referred to as "Participant A". For the original screen, see Appendix C. The lying/punishment scenarios presented here will, in this order, later be referred to as p12, p13, p14, p15, p23, p24, p25, p34, p35, and p45, respectively.

Here, participants played the role of "punishers" and were tasked with judging lying behavior by active players in the pre-experiment. Before the main experiment began, punishers read a description of the lying game and passed several comprehension checks. Punishers were presented with ten punishment scenarios. For each scenario, participants were asked to choose one of 6 punishments options (from 0 to 5 points). Table 1 shows an example of how punishers assigned punishment points. Punishers knew that if they were matched with an actual liar, that liar's earning would then be reduced by \$0.3 for each punishment point that they assigned in the scenario which corresponds to actual behavior by an active player.[10]

Our design of the lying task allowed for different "punishment scenarios" (i.e., different scenarios for lying which could then be punished). We denote them by a "p" and a number whose first digit indicates the actual outcome of the die and the second digit indicates the

(by about 0.46) than the actual outcomes (signed rank test: p=0.003). Figure A.1 in the appendix provides more details on the (revised) reporting pattern.

[10] For example, if such player rolled a "2" but reported a "4" and the matched punisher assigned $x \in \{0, ..., 5\}$ punishment points for this scenario, that player's earnings would be reduced by $x \times \$\,0.3$. Note that the relevant report for punishment was the second, potentially revised, report submitted by the active player. Also, while punishment affected that player's earnings if matched with the punisher, it did not affect the lottery tickets and chances to win the bonus.

reported outcome. For example, the scenario when the active player rolled a "1" but reported a "2" will be referred to as "p12". These ten scenarios differed in two main dimensions:[11]

1. The size of the lie (how much the reported outcome of the die exceeded the actual one).

   - **Lie size=1:** report larger by 1 than actual outcome (p12, p23, p34, and p45).
   - **Lie size=2:** report larger by 2 than actual outcome (p13, p24, p35).
   - **Lie size=3:** report larger by 3 than actual outcome (p14 and p25).
   - **Lie size=4:** report larger by 4 than actual outcome (p15).

2. The "equity nature" of a scenario (the chances of obtaining the bonus for the *active* player, relative to the *passive* player).

   - **Equity:** lying leads to more equity (reduces the gap) in the chances of winning the lottery (p12, p13, and p23).
   - **Inequity:** lying leads to (more) inequity in the chances of winning the lottery (p34, p35, and p45).
   - **Overclaiming:** starting from a situation with a disadvantaged active player, lying *reverts* inequality, leading to a now disadvantaged *passive* player (p14, p15, p24, and p25).

Note that goal of our above-described design is to carve out the effectiveness of norm-nudges in a methodologically clean manner. A key characteristic is the use of the strategy method as opposed to a direct response method. We acknowledge that there a pros and cons of either choice, as discussed in Brandts and Charness (2011). They look at punishment studies and find that overall, the use of the strategy method leads to lower effects than with the direct response. In this regard, our findings should indicate a lower bound effect of norm nudges. Note that due to the fact that we use the strategy method and the considered scenarios consistently for all subjects, any demand effect that may come with the use of this method does not correlate with our norm nudge treatments (for supporting discussion see Zizzo, 2010; De Quidt et al., 2018).

Another important ingredient of our design is to make punishment costless for the punisher. This is a necessary design choice in order to study the impact of the treatment variations mentioned above without having to consider the role of other factors (e.g., an additional monetary trade-off or risk assessment when there is the thread of counter-punishment; see Coffman, 2011 or Feess et al., 2018 for similar design choices).[12] With respect to the norm nudges as implemented in our study, this reflects our goal to study their effectiveness in affecting punishment patterns without the presence of incentives that counter the nudge intervention and could thus

[11] Table A.1 in Appendix A provides a summary of all punishment scenarios along these dimensions.

[12] For example, Nikiforakis (2008) shows that counter-punishment decreases enforcement of cooperation (see also Balafoutas et al., 2016, for recent findings of how concerns for couter-punishment impede norm enforcement).

mute effects that would have otherwise been observable. However, we also acknowledge that while this paper, as a first step, isolates and examines pure impact of norm-nudges in a controlled setting, punishment costs may be factor in future, extra-laboratory research.

Treatments: Our experiment has three conditions to which punishers were randomly allocated. The above describes our baseline (NO INFO). Alternatively, punishers were randomly allocated to one of two norm-nudge treatments. These treatments provide norm information based on actual behavior by subjects in two previous, auxiliary studies. This information is either descriptive (what other participants previously *did*; treatment EMPIRICAL) or normative (what other participants stated was the right thing to do; treatment NORMATIVE).[13] More precisely, if a punisher was allocated to a norm-nudge treatment, one of the two following messages was presented before punishment decisions could be made:

Norm-nudge in treatment EMPIRICAL

> *Please note the following:*
>
> *In another previous study, people were in the situation as participants of Study P for whom you will soon determine the punishment (however, they were not participants of Study P). Here is how they behaved:*
>
> ***The majority of Player As in the previous study reported the number truthfully (i.e., reported exactly what the die showed).***

Norm-nudge in treatment NORMATIVE

> *Please note the following:*
>
> *In a previous study, we asked people what they consider to be the right thing to do for participants in the role of Player A. Those people could not play the game themselves.*
>
> ***The majority of these people stated that the right thing to do for Player A is to report the number truthfully (i.e., report exactly what the die showed).***

This approach of providing norm information follows the long and established tradition of social norms literature (see, e.g., Cialdini and Goldstein, 2004; Bicchieri, 2006; Schultz et al., 2007; Goldstein et al., 2008), which has repeatedly validated the use of such "majority" messages – in which the majority of others' behavior or approval is utilized – to point to social norms (see, e.g., Allcott, 2011; Ferraro et al., 2011; Bhanot, 2018). A common feature of this

[13] The auxiliary experiment for NORMATIVE had n=51 subjects who were told of the active players' decision situation. They then had to indicate, for each possible die roll, what they consider "the right thing to do" in terms of reporting. The majority (68.6%) chose that reporting the the actual outcome (i.e., to tell the truth) was the right thing to do. The auxiliary experiment for EMPIRICAL had n=52 subjects who were in the same decision situation as active players in the pre-experiment, except that they could not revise their statement (since there was no punishment). The majority of them (57.7%) reported truthfully.

literature is is the idea that norms are behavioral patterns embedded in a shared understanding of acceptable actions within a reference group (Ostrom, 2000) and that social norms have two distinct components (see Cialdini et al., 1990; Bicchieri, 2006): an empirical component (often referred to as a *descriptive norm*) and a normative component (often referred to as an *injunctive norm*). From this literature follows that norms can be understood as coordination games among members of a reference group and that shared signal provided them can sustain norm adherence by facilitating coordination (Bicchieri, 2006; Young, 2016). Prior research that has utilized these concepts has shown that both ingredients of a social norm affect and guide behavior, but that their relative effectiveness can often differ (Schultz et al., 2007; Bicchieri and Xiao, 2009; Bhanot, 2018; Bicchieri et al., 2020b,c). For this reason, we make this distinction a central part of our experimental design, which enables us to capture heterogeneous effects of norm nudges depending on what "component" of a norm is breached.

Behavioral Predictions

Our experimental design allows us to extend insights from the existing literature in order to explore the relationship between norm breaching and punishment along different dimensions. First, we examine how the provision of the two types of norm-nudges in treatments NORMATIVE and EMPIRICAL affects the punishment of norm violations in the context of lying. Prior research has shown that providing norm-nudges, such as information on how relevant peers behave and what behavior they consider appropriate, can affect whether people adhere to norms (for an extensive discussion of theoretical and experimental literature see Bicchieri and Dimant, 2019). Here, we do not only analyze the effect of such norm-nudges (as opposed to no norm-nudge), but also how the provision of different forms of norm information (descriptive or normative) affects punishment. While some prior research points to instances in which norm violators – and even those who fail to punish norm violators – are punished (Winter and Zhang, 2018; Martin et al., 2019; Stamkou et al., 2019), there is currently no systematic research accounting for how breaching different types of norms affects their enforcement.

With this in mind, we can derive a set of pre-registered hypotheses.[14] First, existing research suggests that people are receptive to norm information and conform to both observed behavior and normative messaging (e.g., Goldstein et al., 2008; Bott et al., 2020; Dimant, 2019). We therefore hypothesize that norm nudges in EMPIRICAL and NORMATIVE lead to more punishment compared to NO INFO. In addition, theoretical and experimental insights from Bicchieri et al. (2020b) suggest that while people interpret honest behavior, which is a costly action, as a strong indicator of normative disapproval of lying, the reverse may not be true: merely saying what (not) to do does not necessarily have to be followed by the respective actions. Thus, we expect that empirical information may work as a stronger norm nudge since acting in breach of what people do is a stronger signal than acting breach of what people say:

[14] See Appendix B for details on the pre-registration.

Hypothesis 1. *The amount of punishment assigned increases over our three treatments in the following order: NO INFO < NORMATIVE < EMPIRICAL*

Second, an extensive literature in social science has established the determinants of lies and lying costs (Abeler et al., 2014; Dufwenberg and Dufwenberg, 2018; Gneezy et al., 2018; Abeler et al., 2019). Little is known, however, regarding whether and how these findings are reflected in the punishment of lies, especially not within the context of social norm breaching. Based on the existing theories and experimental evidence with regards to lying, we hypothesize that not only the occurrence of a lie but also its size matters:

Hypothesis 2. *The amount of punishment assigned increases with the size of the lie, the reported minus actual outcome.*

Third, the existing literature emphasizes the importance of equity concerns, including in the context of deviant behavior (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Bolton et al., 2019). An unexplored question, however, is whether such motivations matter for the assessment of a norm breach and, consequently, affect the severity of punishment. Note that a mere aversion to unequal chances in getting the bonus is not able to explain the hypothesized effects of the norm nudges (because they do not change allocation). It does also not explain why larger lies should attract more punishment: this is because larger lies do, on average, increase inequity as much as they decrease it.[15] However, for a given lie size, some punishment scenario do increase this inequity while others decrease it. We hypothesize that breaching a norm in the form of over-reporting for the purposes of "getting ahead in unfair ways" is assessed differently from the purpose of leveling the playing field (see also Gino and Pierce, 2010; Bortolotti et al., 2017). While this logic also leads us to expect that lying in the Inequity-scenarios or Overclaim-scenarios will be punished harsher than in the Equity-scenarios, how punishments differ between the former two scenarios is an empirical question that we will investigate in our analysis.

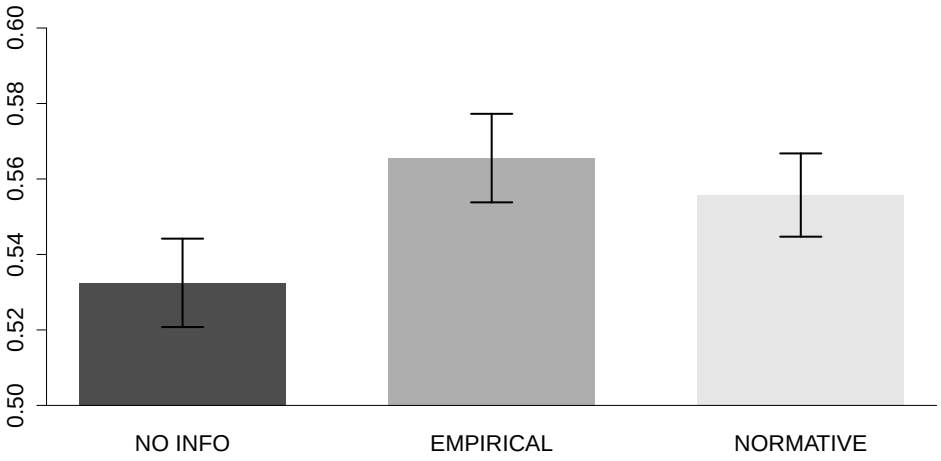
Hypothesis 3. *The equity nature of the lie matters. For a given size of the lie, the amount of punishment assigned*

a) in Equity-scenarios is lower than in Inequity-scenarios,

b) in Equity-scenarios is lower than in Overclaim-scenarios.

[15] Across the ten punishment scenarios, there is either always a scenario where, for a given lie size, an inequity increase in the chance to obtain the bonus is offset by a scenario where the same lie size decreases inequity (relative to an equal split of 3: p14/p25, p13/p35, p12/p45, and p23/p34) or where a lie does only reverses inequity (p15 and p24) but does not change it. Note that the punishment can also not be used to decrease (expected) inequity via the punishment amount subtracted from the lying player's earnings. The reason is that the size of the lie and inequity are uncorrelated. Higher punishment for larger lies would thus increase the (expected) inequity.

Figure 2: Punishment in the different norm information treatments



Notes: Punishment assigned as a share of total punishment points available in the 10 punishment scenarios, by norm-information treatments. Error bars denote SEM.

Results of the punishment experiment

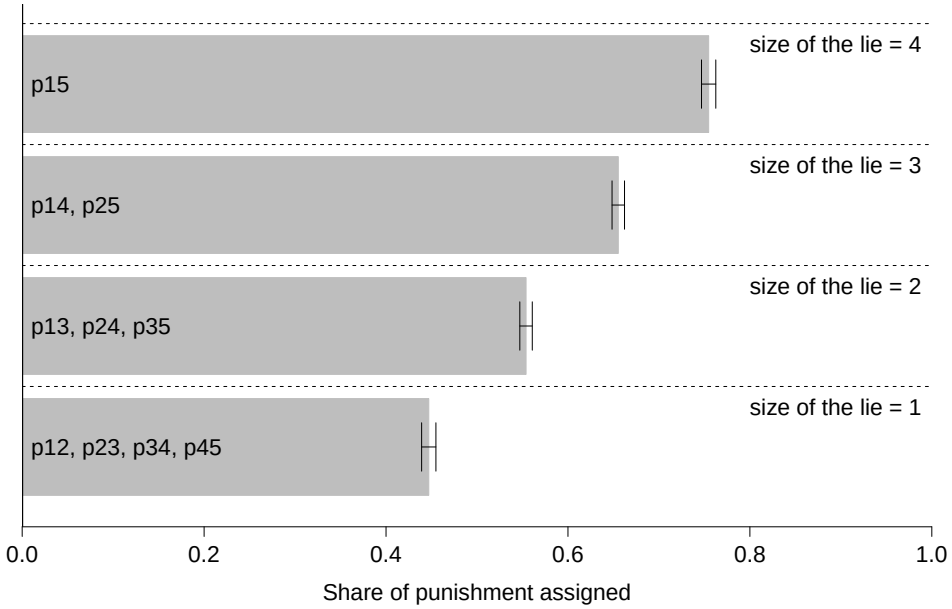
In the following, we report the punishers' actions. In doing so, we follow the order of the hypotheses described above using non-parametric methods. In a last step, we corroborate our results in a regression framework.

Punishment and norm-nudges: We start by looking at the aggregate impact of norm nudges on punishment. To test this, we compute the share of the possible punishment that subjects assigned across the 10 punishment scenarios shown to them.[16] We then compare the mean of the share of punishment assigned by punishers in our three treatments. Figure 2 displays the means and their associated standard errors.

Following Hypothesis 1, we expect behavior that is in conflict with others' behavior (EMPIRICAL treatment) should be punished at least as harshly as behavior that is in conflict with what is deemed appropriate by others (NORMATIVE treatment). We find conclusive evidence for our hypothesis in that both information treatments affect the extent of inflicted punishment. In particular, we observe the highest punishment in the EMPIRICAL treatment, which is significantly different from punishment in the NO INFO treatment (Wilcoxon rank-sum test: p=0.038). Punishment in EMPIRICAL is also directionally – but not significantly – larger than punishment in the NORMATIVE condition (Wilcoxon rank-sum test: p=0.734). Our result also show that punishment in the NORMATIVE condition is larger than punishment in the NO INFO condition (Wilcoxon rank-sum test: p=0.058). In sum, we find evidence that is consistent

[16] For example, consider a punisher who assigned 2 punishment points (=40% of available punishment) in five scenarios and 3 punishment points (=60% of available punishment) in the remaining five scenarios. That punisher's share of punishment assigned would then be 50%.

Figure 3: Punishment by size of the lie



Notes: Punishment assigned as a share of total punishment points available in the punishment scenarios for a given size of the lie (associated punishment scenarios are displayed in each bar). Error bars denote SEM.

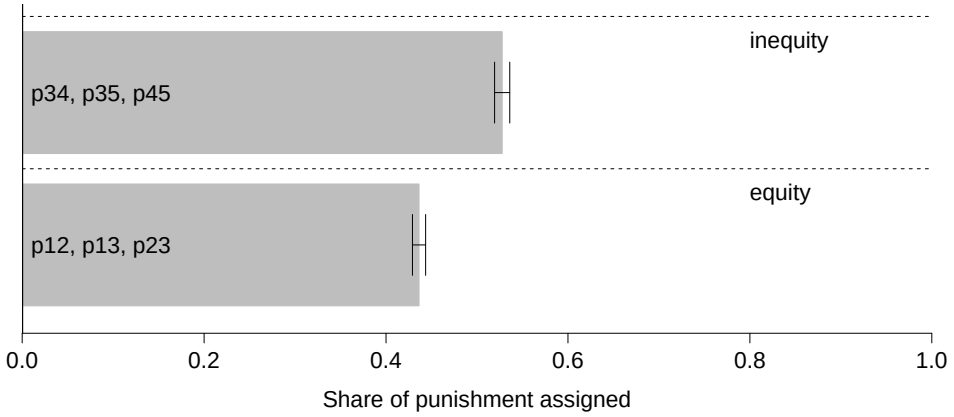
with Hypothesis 1 to the degree that observing lying behavior in combination with norm-related information leads to higher punishment.

Punishment and the size of the lie: In a next step, we examine the relationship between the size of the lie and punishment. To do so, we calculate the mean of punishment assigned share across different sizes of lies.[17] Figure 3 shows the results and provides a clear answer. Consistent with our intuition as formulated in Hypothesis 2, punishment increases significantly with the size of the lie. Specifically, we observe that the smallest possible lie (size = 1) is only assigned a share of punishment of 44.7% of the maximum possible punishment. With each larger lie, the share of punishment increases by about 10 percentage points in a linear manner. This increase is confirmed by Wilcoxon signed rank tests, which show that the shares of punishment under different sizes of the lies always differ significantly ($p<0.001$ for all six pairwise comparisons). In consequence, these results provide compelling evidence that individuals do not only punish norm breaches per se, but also take into consideration the extent to which norms are breached.

Equity nature of the lie: Hypothesis 3 posits that while individuals care about the extent of a norm breach, not all norm breaches are created equal and, thus, not punished equally. That is, lying to correct an initial unfair situation might be judged – and punished – differently than

[17] For example, if the size of the lie = 2, it groups the share of punishment assigned in the scenarios p13, p24, and p35 (Figure A.2 in the Appendix shows the ungrouped results for every single punishment scenario).

Figure 4: Punishment by equity norm



Notes: Punishment assigned as a share of total punishment points available in the punishment scenarios for a given equity norm (associated punishment scenarios are displayed in each bar). Error bars denote SEM.

lying to exacerbate an already unfair situation. Consequently, we examine Equity-, Inequity-, or Overclaim-Lying separately.

Figure 4 displays the means for punishment in the scenarios that feature Equity- or Inequity-Lying.[18] Note that in both cases, there are three associated scenarios with the same lie sizes (2×Lie size=1 and 1×Lie size=2). While punishment for lies that achieve equity amounts to 43.6%, the share of punishment for lies that achieve inequity is 9.2 percentage points higher, at a level of 52.8%. We also find that the punishment choices across these two equity norms are significantly different (Wilcoxon signed rank test: $p<0.001$). This confirms Hypothesis 3a. The figure does not display the average punishment for Overclaim-Lying. The reason is that there are four associated punishment scenarios, rather than the three scenarios for Equity- and Inequity-Lying and that the also have different lie sizes (1×Lie size=2, 2×Lie size=3, and 1×Lie size=4). This makes it difficult to compare it directly with punishment rates in the Equity- and Inequity-Lying scenarios. We provide this comparison as a part of the subsequent regression analysis, which allows us to account for these differences and answer Hypothesis 3b.

Regression analysis: We report results from fitting a regression of the following form:

$$\begin{aligned} p_{is} = \alpha \quad & +\beta_1 \,\text{EMPIRICAL}_i + \beta_2 \,\text{NORMATIVE}_i \\ & +\beta_3 \,\text{LieSize}_s + \beta_4 \,\text{Inequity}_s + \beta_5 \,\text{Overclaiming}_s + \gamma \,\mathbf{Controls_i} + \epsilon_{is} \end{aligned} \tag{1}$$

In the above, the dependent variable p_{is} is the share of punishment by subject i in scenario s, expressed as a number between 0 and 100. Consequently, the coefficient on the independent variables can be interpreted as percentage point shifts. The independent variables in (1) are the following: EMPIRICAL_i and NORMATIVE_i are dummies indicating whether subject i was in

[18] Figure A.3 in Appendix A displays the average punishment level for each single punishment scenario.

Table 2: Determinants of punishment – regression results

	Share of punishment assigned in %			
	(1)	(2)	(3)	(4)
EMPIRICAL	3.174*			3.174*
	(1.708)			(1.708)
NORMATIVE	2.978*			2.978*
	(1.687)			(1.687)
Size of the Lie		10.490***		10.261***
		(0.308)		(0.338)
Inequity			9.142***	9.142***
			(0.553)	(0.553)
Overclaiminig			22.246***	5.144***
			(0.544)	(0.420)
Constant	52.602***	33.933***	43.273***	27.280***
	(7.784)	(7.605)	(7.583)	(7.819)
N	11,340	11,340	11,340	11,340
R^2	0.008	0.128	0.103	0.144

F-test: coeff. for EMPIRICAL = NORMATIVE; p=0.629/0.629 (Col. 1/4)
F-test: coeff. for Size of the Lie; p<0.001 (all pairwise comparisons in Col. 1 or 2)
F-test: coeff. for Inequity = Overclaiming; p<0.001/0.001 (Col. 3/4)
Notes: OLS results of regressing the share of punishment assigned on indicators for the norm information treatment (EMPIRICAL or NORMATIVE), the size of the lie, and the equity nature of the lie (inequity, overclaiming); the baseline category is therefore a subject in treatment NO INFO and an equity-based punishment scenario with lie size=1. Additional control variables include age, gender, education, and controls for how punishment scenarios were presented (increasing/decreasing) and the estimated implementation probability). 10 punishment scenarios per punisher; standard errors are clustered at the punisher level. */**/***: p<0.10/0.05/0.01.

one of the norm nudge treatments (or in NO INFO, the baseline category), $LieSize_s$ measures the size of the lie in punishment scenario s, while Inequity_s and Overclaiming_s are dummies indicating the equity norm of scenario s (with an Equity-scenario being the baseline). Age, gender, and education are collected in the $\mathbf{Controls_i}$-vector for each punisher. This vector also contains a dummy to control for the order in which scenarios were presented (counterbalanced over treatments), and punisher i's estimate for the ratio of punishers to active player in Part 1 (i.e., the implementation probability). We then fit the above model using OLS using standard errors clustered on the subject level i.

Table 2 reports the regression results. The first three columns repeat our previous statistical analysis parametrically while controlling for punishers' personal characteristics. In Column (1) we only add the NormInfo_i-indicators and find that providing empirical and normative information respectively leads to a (marginally) significant increase in punishment of about 2.7 and 3.5 percentage points (these coefficients do not differ significantly from each other; F-test: p=0.629). In Column (2), we also replicate that larger lies lead to significantly more punishment, by about 10 percentage point for each unit increase in the size of the lie. Overall, this re-confirms hypotheses 1 and 2. Column (3) shows that relative to Equity-based punishment scenarios, Inequity- and Overclaim-scenarios lead to a significant increase in the share of punishment assigned by about 9 and 22 percentage points, respectively. We also find that the implied difference of 13 (=22−9) percentage points between Inequity- and Overclaim-scenarios are statistically significantly different (F-test: p<0.001). However, these estimates could also be due to the differing sizes of lies in the Inequity- and Overclaim-scenarios, respectively.

The full regression model presented in Column (4) allows us to measure the effect of different equity scenarios while controlling for differing lie sizes in the associated scenarios the underlying lies differ (i.e., in the Overclaiming-scenarios). We find that with these controls, the coefficient for the Overclaim-scenarios shrinks by about three quarters. This is consistent above-mentioned caveat that a large share of the original effect for Overclaiming is due to the larger lies in its underlying scenarios. In fact, the change by 17.1 punishment points of the Overclaim coefficient captures the change in average size of the lie on the point when controlling for the size of the lie (=22.2-5.1, comparing columns 3 and 4): it corresponds very closely to the 1.7-increase in the average size of the lie (3.0 in Overclaim vs. 1.3 in the Equity- and Inequity scenarios) multiplied with the "Size of the Lie"-coefficient of about 10. However, even when controlling for these larger lie sizes, we find that the Overclaim-scenarios yield a significant increase of 5.1 percentage relative to the baseline Equity-scenarios. Thus, while smaller than suggested by the initial estimates without controls for lie sizes, we find support for Hypothesis 3b: (in)equity concerns do not only matter when lying creates inequality but also when it reverts pre-existing inequality that benefits a liar. The regression analysis re-confirms Hypotheses 1, 2, and 3.

Experiment 2: Norm Information and Norm Perception

The preceding results suggest that people punish norm violations differently depending on the norm information that they receive and the equity nature of the lie. To better understand the mechanism of *why* norm enforcement varies, we examine whether the observed punishment patterns have an analogous variation in the perception of social norms. Investigating this is important: to achieve prolonged norm adherence, enforcement needs to reflect a shared (perceived) social norm and will be less effective if the norms are in conflict with formal rules or comprised of idiosyncratic judgments (Acemoglu and Jackson, 2017; Bicchieri et al., 2020c).

However, prior research has shown that the provision of norm information can – but does not necessarily have to – change the perception of social norms (for a recent example where this approach succeeded see Bursztyn et al., 2020b; and for where it didn't see Dimant et al., 2020). Thus, examining whether changes in the the punishment patterns are consistent with changes in social norms perceptions in our context is an empirical question. We therefore devised Experiment 2 in which we vary, as in Experiment 1, the type of norm information and equity nature of lies. However, instead of eliciting punishment decision, we elicit how the perception of social norms differ across these dimensions.[19]

Design

We recruited a new set of n=1,519 subjects on MTurk in order to elicit their norm perceptions regarding the lying behavior as observed in Experiment 1. Therefore, while this follow-up experiment reflects key features of Experiment 1, we did not elicit punishment behavior here. Instead, we elicit subjects' perceptions with regards to the normative appropriateness of lying in the different scenarios using the incentive-compatible procedure by Krupka and Weber (2013).

Prior to the elicitation of their norm perceptions, participants were informed about the original lying task in the same way that it was explained to punishers in Experiment 1 (i.e., about the structure of Part 1 and Part 2; see Figure 1). Subsequently, each subject was presented with one lying situation. The presented situations varied along two main dimensions. The first dimension was whether no norm information (NO INFO), normative information (NORMATIVE) or empirical information (EMPIRICAL) was provided. The second dimension was the equity nature of the lie. In this case, we picked three lying scenarios that reflected Equity-Lying (p13), Inequity-Lying (p35), and Overclaim-Lying (p24). Note that for the purpose of comparability, we kept the size of the lie constant at 2.

This yields a between-subjects design that varies norm information and equity nature of a lie in a fully factorial manner over 3 (norm information) $\times$ 3 (equity lying scenarios) treatments

[19] As customary in the norms literature, we elicit norm-related beliefs from participants who have not participated in our Experiment 1 (this avoids, among others, priming- or demand effects and the generation of post-hoc justifications by subjects; see d'Adda et al., 2016 for a discussion).

to which subjects were randomly assigned. We measured our dependent variable of interest by asking participants to rate the extent to which other subjects deemed the observed lying behavior socially (in)appropriate. They did so using a 4-point Likert-like scale ranging across "very socially inappropriate" (VSI), "somewhat socially inappropriate" (SSI), "somewhat socially appropriate" (SSA), and "very socially appropriate" (VSA). In each treatment, participants were given a monetary incentive to guess the modal answer, thus allowing for an incentive-compatible elicitation of norm perceptions.[20]

Results

Figure 5 illustrates the distributions of social (in)appropriateness ratings, split by whether norm information is provided (columns) and by the equity nature of the lie (rows). We first examine the role of different (in)equity-scenarios, then the role of norm information, and lastly their interaction.

Equity nature of the lie: Aggregated over the norm-nudge information treatments (i.e., row-wise comparisons in Figure 5), the average norm appropriateness rating in the Equity-scenario is higher compared to ratings for the Inequity- and Overclaiming-scenarios. This pattern is also marginally significant, according to pairwise ranksum tests (Equity vs Inequity $p=0.063$, Equity vs Overclaiming $p<0.066$, Inequity vs Overclaiming $p=0.983$; all ranksum test reported here and in the following are two-sided). Thus, our results show that the perception of social norms across different equity settings reflect the punishment patterns we observed in Experiment 1.

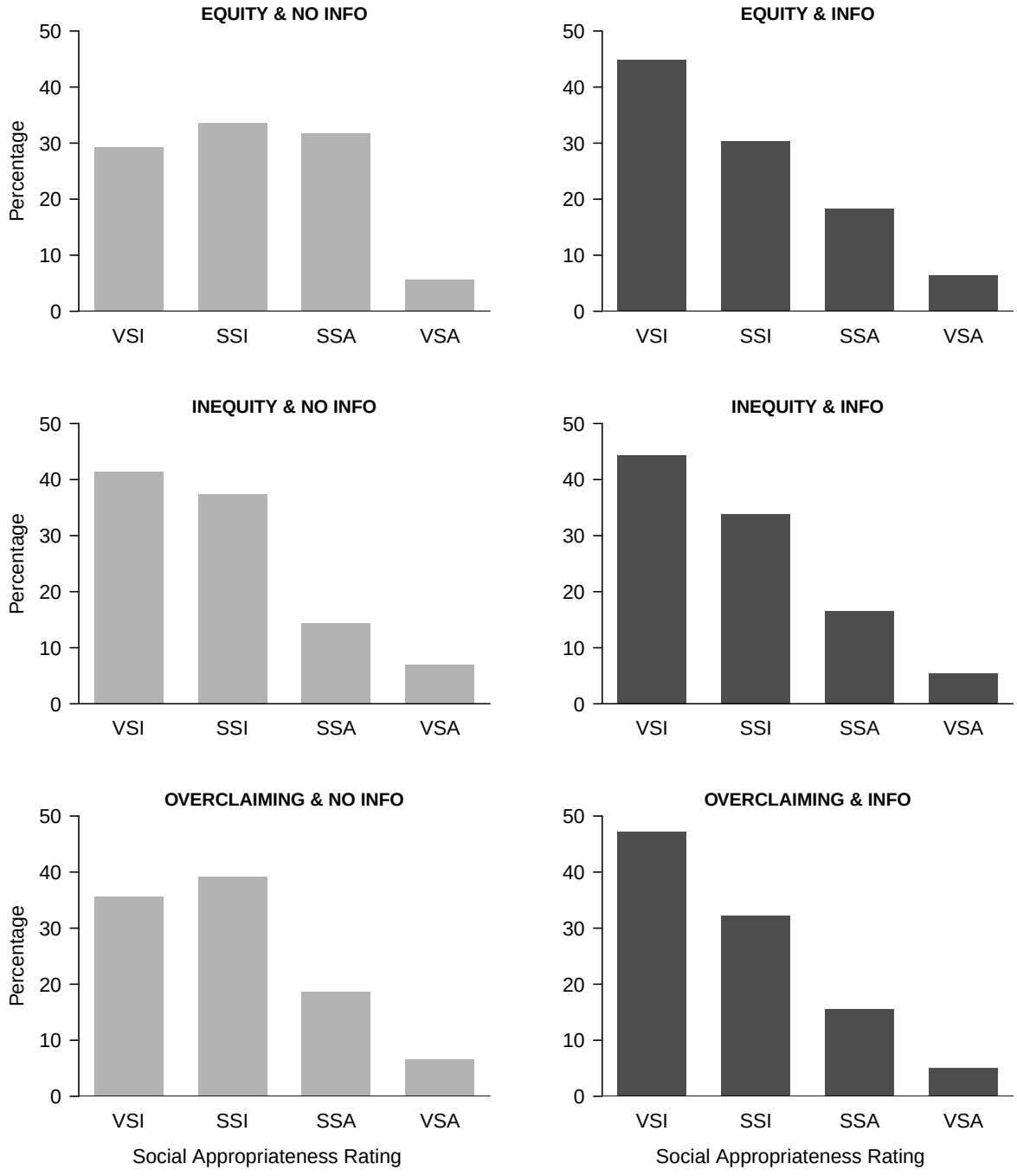
Norm-nudge: Likewise, we find that that providing information about a norm shifts the norm perception in a way that is similar to Experiment 1's findings. In particular, we do not see a difference in responses between the two norm information treatments NORMATIVE and EMPIRICAL (ranksum test: $p=0.218$). Consequently, we pool these two scenarios for the remainder of the analysis.[21] When we compare the different pooled equity-scenarios over the norm nudges (i.e., compare the two columns in Figure 5), we find that providing norm information (INFO) leads to a lower average social appropriateness rating of lying (ranksum-test: NO INFO vs. INFO, $p<0.001$).

Interaction of norm information and equity nature: From the distribution for NO INFO, one can observe that the perceived normative appropriateness is relatively heteroge-

[20] For a breakdown of observations per treatment see Table A.4 in the appendix where we also describe subject characteristics and provide randomization checks. The average duration of the experiment was about four minutes. Participants were paid a show-up fee of \$0.20 with an opportunity to receive a \$0.20 bonus based on their answers, and thus paid well above the average MTurk pay (Hara et al., 2018). We employed the same participant pool restrictions as in Experiment 1 (see Footnote 5).

[21] Figure A.4 in Appendix A shows when the distributions of responses when NORMATIVE and EMPIRICAL are separately displayed; the results are largely the same. Specifically, we find that appropriateness ratings in EMPIRICAL and NORMATIVE are significantly higher than ratings in NO INFO (ranksum-tests: EMPIRICAL vs. NO INFO, $p=0.012$; NORMATIVE vs. NO INFO, $p<0.001$).

Figure 5: Social appropriateness of lying over different norm information and equity situations



Notes: Norm information varies by column, and the equity nature of the lie varies by rows. The social norm is measured via a 4-items Likert scale ranging over "Very Socially Inappropriate" (*VSI*), "Somewhat Socially Inappropriate (*SSI*)", "Somewhat Socially Appropriate (*SSA*)", "Very Socially Appropriate (*VSA*)". INFO pools treatments the norm information treatments NORMATIVE and EMPIRICAL.

neous across the various equity-scenarios. For the Equity-scenario, displayed in the first row, the perceived social norm towards lying is relatively forgiving, with *VSI*, *SSI*, and *SSA* each obtaining about one-third of the total ratings. In contrast, most subjects seem to be aware that lying in the Inequity-scenario (second row) is "Very socially inappropriate" (*VSI*). For the Overclaim-scenario (third row), the pattern is similar, even though less pronounced. In line with these observations, responses in the Inequity- and Overclaiming-scenarios do not differ significantly (ranksum test: p=0.261). In contrast, responses in the Equity-scenario are (marginally) significantly different from responses in the Inequity- and Overclaim-scenarios (ranksum tests: p=0.003 and p=0.061, respectively). Importantly, we observe a very different pattern across all three equity-scenarios when norm-related information is provided (right column of Figure 5). One can clearly observe that providing norm information generates almost identical distributions, independent of the underlying equity-scenario. Across all scenarios, "Very Socially Inappropriate" is the modal answer, whereas almost no one considers lying in those scenarios to be appropriate. This homogeneous pattern induced by norm information across equity conditions is also reflected in ranksum tests that do not indicate significant differences ($p \geq 0.311$ for all pairwise comparisons).

It is noteworthy that the provision of norm information helps people to coordinate on the modal value: in the INFO treatments, there is a clear spike in the frequency with which the modal value (here, *VSI*) is selected. In the NO INFO treatments, such coordination is less pronounced. These findings are also supported by a regression analysis of subjects' norm ratings where we control for subject characteristics (see Table A.5 in Appendix A). Thus, providing a norm-nudge leads people to adapt and anticipate the transmitted information, independent of the original equity situation. In consequence, we observe significant differences in perceived social norms only when norm information is provided in an equity-scenario with a relatively unclear pre-existing norm perception. That is, lying is to be less appropriate with norm information than without norm information for the Equity-scenario (ranksum test: $p<0.001$) and the Overclaiming-scenario (ranksum test: p=0.020). In contrast, we do not find a significant difference in the case of Inequity-based lying where the pre-existing norm against lying was already relatively strong (ranksum-test: p=0.645).

Discussion

In Experiment 1, we investigate the drivers of norm enforcement in the context of lying. Across different norm information settings, participants observe the behavior of liars who differ in how much they lie (as measured by the difference between stated and actual outcomes) and the equity-consequences of the lie. We find that punishment is higher for larger lies and lies that increase inequity for the liar. In contrast, punishment is less severe when the lie serves the benefit of offsetting an ex-ante imbalance. We also find evidence that punishment is more pronounced when norm information is provided.

In Experiment 2, we investigate the reasons for *why* norm enforcement varies across the settings that subjects considered in Experiment 1. We do so by using a separate online sample to elicit social norm perceptions across these lying settings. We observe that the norm perception resemble the punishment patterns in the first experiment. That is, our results point out that inequity-based lies are perceived to be less acceptable than equity-based lies. We also show that providing norm information decreases the perceived acceptability of lying downwards; this works primarily through a downward-shift in the acceptability of equity-based lies. Therefore, these parallel patterns in the variation of punishment and norm perception suggest a close link between these components.[22]

Our main insights derived from the results of both experiments are that norm-nudges such as providing norm information do, in principle, foster norm enforcement. We also find that while they work, their nature does not matter – "hard facts" such as empirical information are not more effective than mere normative information in our context. A deeper analysis of our findings suggests that the effect of such information occurs through shifting norm perceptions and helping to coordinate. This applies in particular for equity-based lies, for which norm perceptions dispersed the most without norm information. Together, this suggests that norm nudges also work by reducing normative uncertainty for the punishers. Conversely, we show that where clear norms exist but are not necessarily honoured (as in the case of inequity-based lying), norm-nudges via norm information are less effective.

Conclusion

Enforcing social norms make up the fabric of a functioning society, and this fabric is undermined where transgression goes unpunished (Bicchieri, 2006). While individually costly, the society's collective gains can be substantial through facilitating coordination and norm adherence (see Xiao, 2018, for a recent review). While much of the existing literature has studied the *effect* of punishment in social interactions and its ability to uphold social norms, less is known about the *drivers* of such punishment. It is therefore imperative to understand the circumstances for which an observed norm-transgression is punished and the role that the motives behind the transgression play. Our contribution is showing how multi-layered norm enforcement really is, how it is affected by norm-nudges, and how it is linked to the perception of norms.

We analyze the extent to which norm enforcement is sensitive to the consequences of the observed transgression (e.g., to achieve equity versus achieving an unfair advantage) as well as the type of norm-nudges (empirical versus normative) with which the observed behavior is inconsistent. We examine questions through the lens of two experiments: the first experiment

[22] Recall that Experiment 1 is a within-subject design with respect to the different lying scenarios whereas Experiment 2 varies all treatment dimensions between subjects. The fact that results are nevertheless very consistent across these independent experiments suggests that our results are robust to both of these design choices (see also Clifford et al., 2020, for a recent discussion on the robustness of within-designs).

aims to measure norm-*enforcement* behavior, whereas the second experiment aims to measure the associated norm-*perceptions* of the observed lies. This approach allows us not only to understand how norm transgressions are punished, but also the extent to which variations in norm enforcement align with variations in norm perceptions. Indeed, we find that both map onto each other very well.

From a policy perspective, our results emphasize the necessity to capture heterogeneous social motives, especially when "soft" interventions such as nudges are used to achieve behavioral change. Existing work has shown that norm-based interventions such as norm-nudges need to be meticulously mapped onto the social environment in which they are implemented. Taking this into account, our results are also informative of ongoing scholarly debate on the reasons why prior studies that used such soft norm-nudge interventions had mixed success (Fellner et al., 2013; Castro and Scartascini, 2015; Hallsworth et al., 2017; Kettle et al., 2017; Dimant et al., 2020). For example, our findings suggest that norm-nudges exhibit the largest impact where pre-existing norm perceptions are rather inconclusive (that is, where our norm-elicitation measure did not identify a uniquely prevailing norm, as in the Equity & NoInfo setting).

We help to advance this scholarly debate by pointing towards what works, but also to highlight the reasons for unsuccessful nudging. On the one hand, our results suggest that norm enforcers unhesitatingly utilize punishment in a way that is sensitive to the motives underlying a norm breach. Thus, a "hands-off" approach can be warranted as long as norm following is self-enforcing through peer punishment. On the other hand, we find that using simple norm-nudges does not necessarily change the norm perception of the transgression – and the extent to which it is punished – if a norm already exists firmly. Consequently, to achieve behavior change and foster norm enforcement in such contexts, one first need to better understand the context and the existing norm in which the intended behavior should occur. Depending on the setting, one may be able to rely on gentle norm-nudges, as studied here, or rather have to turn to stronger, nudge-adjacent interventions such as *shoves* and *boosts*, or even explicit economic incentives (see Kahan, 2000; Grüne-Yanoff and Hertwig, 2016; Gino et al., 2019).

References

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

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*.

Acemoglu, D. and Jackson, M. O. (2017). Social norms and the enforcement of laws. *Journal of the European Economic Association*, 15(2):245–295.

Albrecht, F., Kube, S., and Traxler, C. (2018). Cooperation and norm enforcement-the individual-level perspective. *Journal of Public Economics*, 165:1–16.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, 95(9-10):1082–1095.

Allcott, H. and Kessler, J. B. (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics*, 11(1):236–76.

Arechar, A. A., Gächter, S., and Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21(1):99–131.

Balafoutas, L. and Nikiforakis, N. (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review*, 56(8):1773–1785.

Balafoutas, L., Nikiforakis, N., and Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, 111(45):15924–15927.

Balafoutas, L., Nikiforakis, N., and Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, 7.

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., and Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, 28(8):1041–1055.

Bhanot, S. P. (2018). Isolating the effect of injunctive norms on conservation behavior: New evidence from a field experiment in california. *Organizational Behavior and Human Decision Processes*.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. and Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice*.

Bicchieri, C., Dimant, E., Gaechter, S., and Nosenzo, D. (2020a). Observability, social proximity, and the erosion of norm compliance. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3355028.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2020b). It's not a lie if you believe the norm does not apply: Conditional norm-following with strategic beliefs. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3326146.

Bicchieri, C., Dimant, E., and Xiao, E. (2020c). Deviant or wrong? The effects of norm information on the efficacy of punishment. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3294371.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Bolton, G., Dimant, E., and Schmidt, U. (2019). The fragility of nudging: Combining (im)plausible deniability with social and economic incentives to promote behavioral change. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3294375.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Bortolotti, S., Soraperra, I., Sutter, M., and Zoller, C. (2017). Too lucky to be true-fairness views under the shadow of cheating. Working Paper Available at SSRN: https://ssrn.com/abstract=3014734.

Bott, K. M., Cappelen, A. W., Sorensen, E., and Tungodden, B. (2020). You've got mail: A randomised field experiment on tax evasion. *Management Science*, 66:2801–2819.

Brandts, J. and Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14(3):375–398.

Brouwer, T., Galeotti, F., and Villeval, M. C. (2020). Teaching norms in the streets. Mimeo.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020a). From extreme to mainstream: the erosion of social norms unravel. *American Economic Review, forthcoming.*

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2020b). Misperceived social norms: Women working outside the home in saudi arabia. *American Economic Review*, 110(10):2997–3029.

Castro, L. and Scartascini, C. (2015). Tax compliance and enforcement in the pampas evidence from a field experiment. *Journal of Economic Behavior & Organization*, 116:65–82.

Choi, D. D., Poertner, M., and Sambanis, N. (2019). Parochialism, social norms, and discrimination against immigrants. *Proceedings of the National Academy of Sciences*, 116(33):16274–16279.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015–1026.

Clifford, S., Sheagley, G., and Piston, S. (2020). Increasing precision in survey experiments without introducing bias. *APSA Preprints. doi: 10.33774/apsa-2020-tbd3c.*

Coffman, L. C. (2011). Intermediation reduces punishment (and reward). *American Economic Journal: Microeconomics*, 3(4):77–106.

Coleman, J. S. (1994). *Foundations of social theory.* Harvard University Press.

Coricelli, G., Montmarquette, C., and Villeval, M. C. (2010). Cheating, emotions, and rationality: an experiment on tax evasion. *Experimental Economics*, 11(3):226–247.

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

Damgaard, M. T. and Gravert, C. (2018). The hidden costs of nudging: Experimental evidence from reminders in fundraising. *Journal of Public Economics*, 157:15–26.

De Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.

DellaVigna, S. and Linos, E. (2020). Rcts to scale: Comprehensive evidence from two nudge units. Technical report, Working Paper, UC Berkeley.

Dickinson, D. L., Masclet, D., and Villeval, M. C. (2015). Norm enforcement in social dilemmas: An experiment with police commissioners. *Journal of Public Economics*, 126:74 – 85.

Dimant, E. (2019). Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73:66–88.

Dimant, E. (2020). Hate trumps love: The impact of political polarization on social preferences. Working Paper Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3680871.

Dimant, E., Gerben, A. v. K., and Shalvi, S. (2020). Requiem for a nudge: Framing effects in nudging honest. *Journal of Economic Behavior & Organization*, 172:247–266.

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

Exadaktylos, F., Espín, A. M., and Branas-Garza, P. (2013). Experimental subjects are not different. *Scientific Reports*, 3(1):1–6.

Feess, E., Schildberg-Hörisch, H., Schramm, M., and Wohlschlegel, A. (2018). The impact of fine size and uncertainty on punishment and deterrence: Theory and evidence from the laboratory. *Journal of Economic Behavior & Organization*, 149:58–73.

Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2):63–87.

Fehr, E. and Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fehr, E. and Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7):458.

Fellner, G., Sausgruber, R., and Traxler, C. (2013). Testing enforcement strategies in the field: Threat, moral appeal and social information. *Journal of the European Economic Association*, 11(3):634–660.

Ferraro, P. J., Miranda, J. J., and Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *American Economic Review*, 101(3):318–22.

Gino, F., Hauser, O. P., and Norton, M. I. (2019). Budging beliefs, nudging behaviour. *Mind & Society*, pages 1–12.

Gino, F. and Pierce, L. (2010). Lying to level the playing field: Why people may dishonestly help or hurt others to create equity. *Journal of Business Ethics*, 95(1):89–103.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–53.

Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, 35(3):472–482.

Grüne-Yanoff, T. and Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, 26(1-2):149–183.

Hallsworth, M., List, J. A., Metcalfe, R. D., and Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148:14–31.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 449. ACM.

Heffner, J. and FeldmanHall, O. (2019). Why we don't always punish: Preferences for non-punitive responses to moral violations. *Scientific Reports*, 9(1):1–13.

Herrmann, B., Thöni, C., and Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319(5868):1362–1367.

John, L. K., Loewenstein, G., and Rick, S. I. (2014). Cheating more for less: Upward social comparisons motivate the poorly compensated to cheat. *Organizational Behavior and Human Decision Processes*, 123(2):101–109.

Kahan, D. M. (2000). Gentle nudges vs. hard shoves: Solving the sticky norms problem. *The University of Chicago Law Review*, pages 607–645.

Kettle, S., Hernandez, M., Sanders, M., Hauser, O., and Ruda, S. (2017). Failure to captcha attention: Null results from an honesty priming experiment in guatemala. *Behavioral Sciences*, 7(2):28.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3):608–638.

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

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2):153–174.

Martin, J. W., Jordan, J. J., Rand, D. G., and Cushman, F. (2019). When do we punish people who don't? *Cognition*, 193:104040.

Maxwell, S. E., Kelley, K., and Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, 92(1-2):91–112.

Nikiforakis, N. and Normann, H.-T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics*, 11(4):358–369.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3):137–158.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Snowberg, E. and Yariv, L. (2020). Testing the waters: Behavior across participant pools. Forthcoming American Economic Review.

Stamkou, E., van Kleef, G. A., Homan, A. C., Gelfand, M. J., van de Vijver, F. J., van Egmond, M. C., Boer, D., Phiri, N., Ayub, N., Kinias, Z., et al. (2019). Cultural collectivism and tightness moderate responses to norm violators: Effects on power perception, moral emotions, and leader support. *Personality and Social Psychology Bulletin*, 45(6):947–964.

Sutter, M., Haigner, S., and Kocher, M. G. (2010). Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *The Review of Economic Studies*, 77(4):1540–1566.

Tankard, M. E. and Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, 10(1):181–211.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Winter, F. and Zhang, N. (2018). Social norm enforcement in ethnically diverse communities. *Proceedings of the National Academy of Sciences*, 115(11):2722–2727.

Xiao, E. (2018). Punishment, social norms, and cooperation. In *Research Handbook on Behavioral Law and Economics*. Edward Elgar Publishing.

Young, H. P. (2016). *Social Norms*, pages 1–7. Palgrave Macmillan UK, London.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98.

Norm Makers and Norm Breakers:
How Motives for Deviance Shape Punishment of Norm Violations

Online Appendix

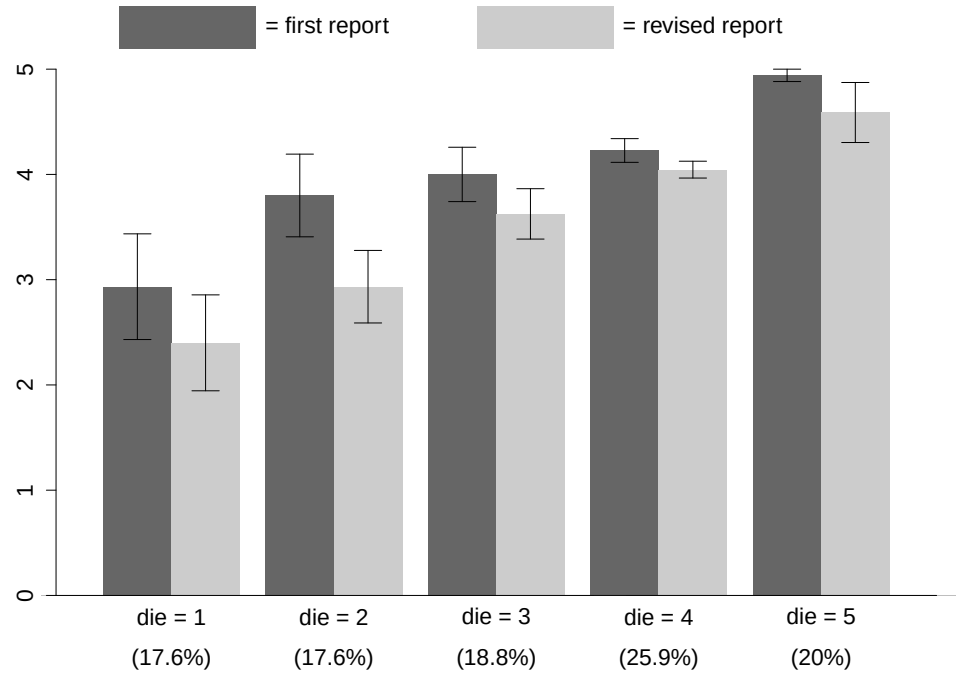
Contents:

Appendix A: Additional figures, tables, and data analysis

Table A.1: Dimensions of punishment scenarios

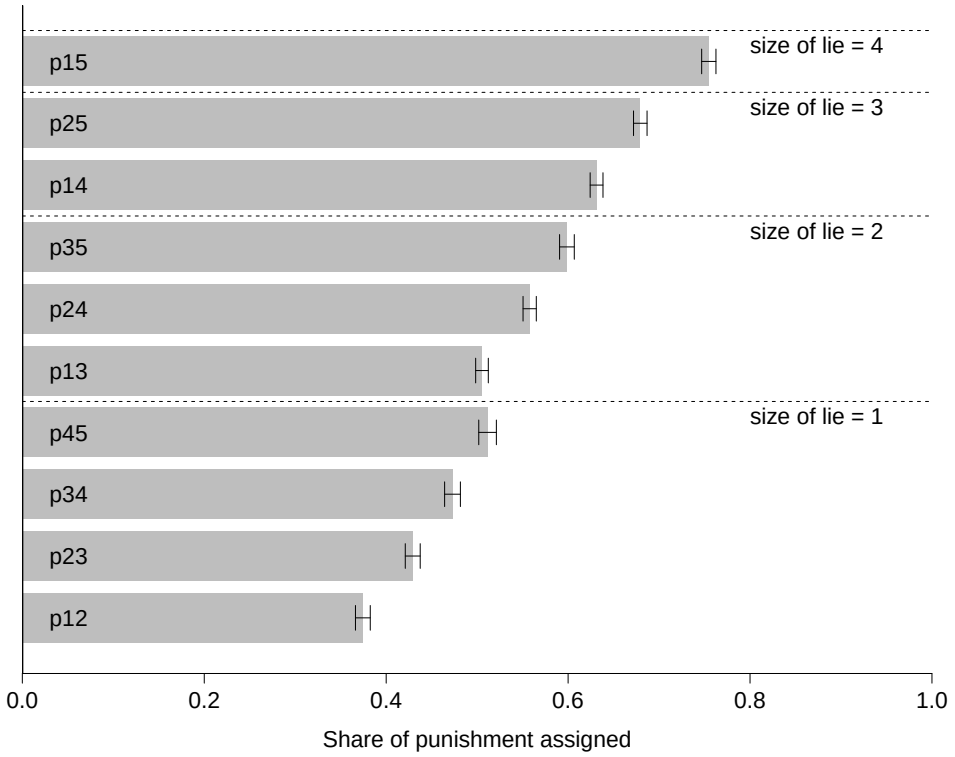
Punishment scenario	Actual outcome	Reported outcome	Size of the lie	equity nature
p12	1	2	1	equity
p13	1	3	2	equity
p14	1	4	3	overclaiming
p15	1	5	4	overclaiming
p23	2	3	1	equity
p24	2	4	2	overclaiming
p25	2	5	3	overclaiming
p34	3	4	1	inequity
p35	3	5	2	inequity
p45	4	5	1	inequity

Figure A.1: Liars' first and revised reports by actual result of the die



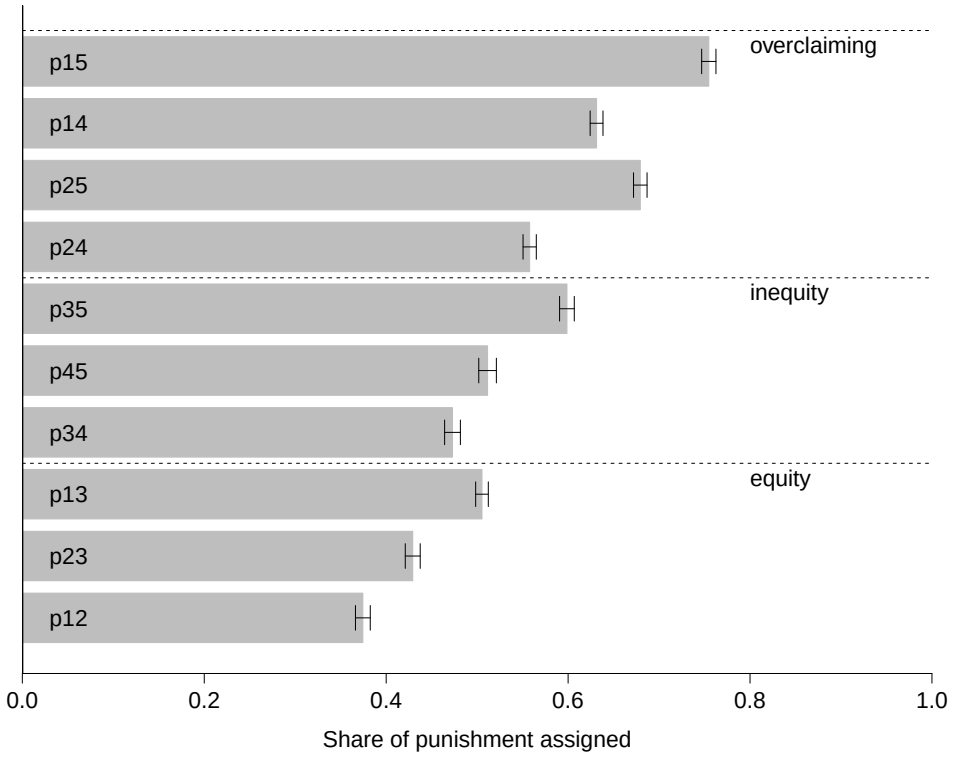
Notes: Means of first and revised reports for the outcome of the 5-sided die toss, grouped by the actual outcome of the die (frequency for each outcome in paratheses). Error bars denote SEM.

Figure A.2: Punishment for each punishment scenario, ordered by size of the lie



Notes: Punishment assigned as a share of total punishment points available in each punishment scenarios, ordered by the size of the lie. Error bars denote SEM.

Figure A.3: Punishment for each punishment scenario, ordered by equity norm



Notes: Punishment assigned as a share of total punishment points available in each punishment scenarios, ordered by equity nature of the lie. Error bars denote SEM.

Table A.2: Descriptive statistics for active players

	Mean	S.D.
Age	36.835	10.194
Male	0.682	0.468
Edu.: some high school	0.012	0.108
Edu.: finished high school	0.118	0.324
Edu.: some college	0.341	0.477
Edu.: finished college	0.447	0.500
Edu.: higher degree	0.082	0.277
observations	N=	85

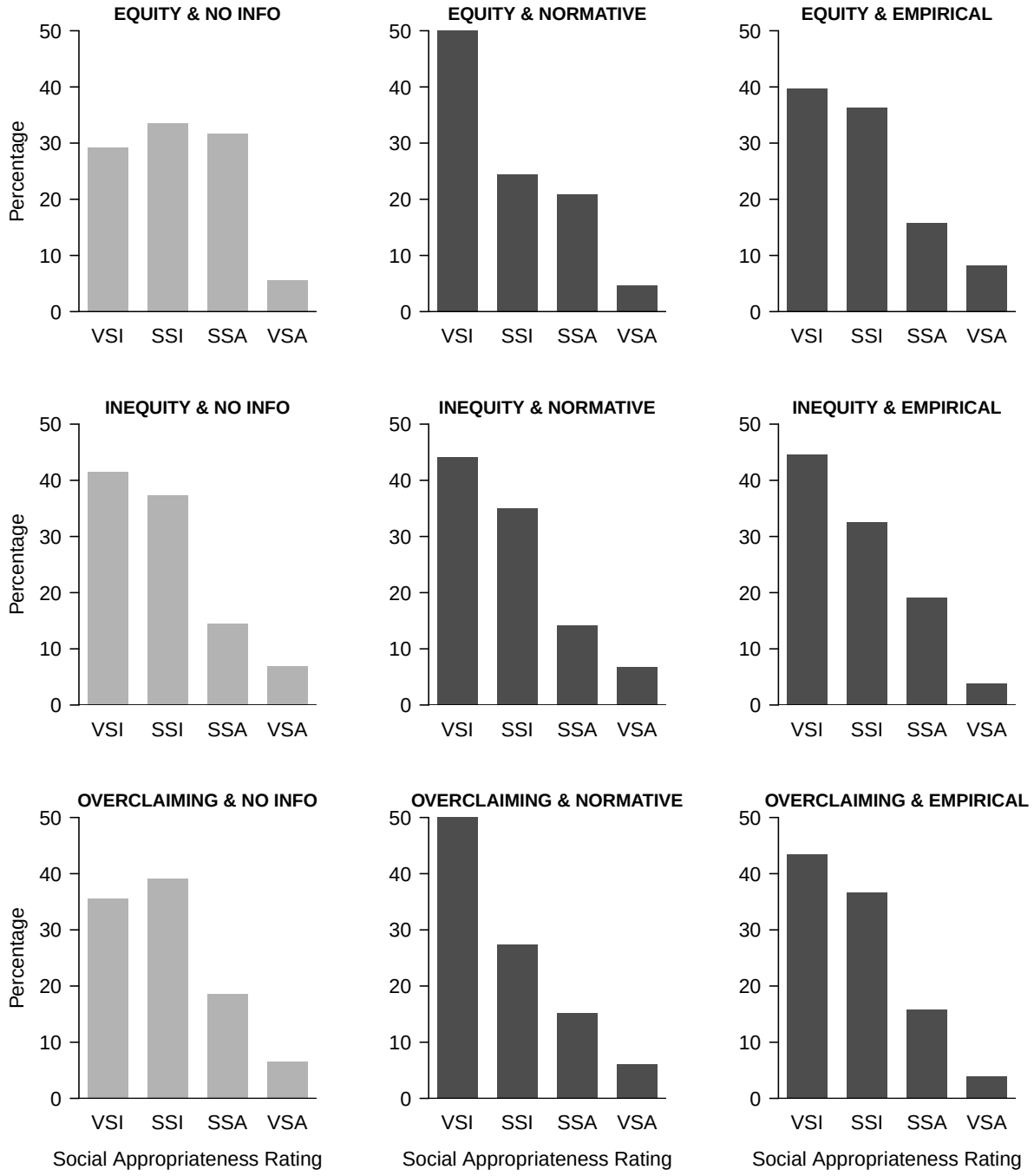
Notes: Personal characteristics of active players in the pre-experiment (mean and standard deviation).

Table A.3: Descriptive Statistics for punishers by norm information treatments

	NO INFO		NORMATIVE		EMPIRICAL	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Age	43.545	13.478	42.552	14.122	43.303	14.089
Male	0.468	0.500	0.477	0.500	0.482	0.500
Edu.: some high school	0.037	0.189	0.039	0.193	0.024	0.152
Edu.: finished high school	0.163	0.370	0.245	0.430	0.194	0.396
Edu.: some college	0.280	0.449	0.257	0.437	0.277	0.448
Edu.: finished college	0.300	0.459	0.293	0.456	0.314	0.465
Edu.: higher degree	0.208	0.406	0.160	0.367	0.175	0.380
observations	N=	404	N=	413	N=	423

Notes: Personal characteristics of punishers in the main experiment, by norm-information treatment (mean and standard deviation). We do not observe statistically significant differences between the treatments for age (Kruskal-Wallis test: p=0.548), gender (χ^2-test: p=0.916), or education (χ^2-test: p=0.190).

Figure A.4: Social appropriateness of lying over different norm information and equity treatment



Notes: Norm information varies by column, the equity nature of the lie by rows. The social norm is measured via a 4-items Likert scale ranging over "Very Socially Inappropriate" (*VSI*), "Somewhat Socially Inappropriate (*SSI*)", "Somewhat Socially Appropriate (*SSA*)", "Very Socially Appropriate (*VSA*)".

Table A.4: Descriptive Statistics for subjects in Experiment 2

	Panel a) NO INFO					
	Equity		Inequity		Overclaiming	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Age	37.862	13.791	39.529	13.686	39.843	12.835
Male	0.469	0.501	0.483	0.501	0.506	0.501
Educ.: less than high school degree	0.000	0.000	0	0.000	0.000	0.000
Educ.: high school graduate	0.094	0.292	0.080	0.273	0.133	0.340
Educ.: some college but no degree	0.150	0.358	0.161	0.369	0.193	0.396
Educ.: associate degree in college	0.094	0.292	0.092	0.290	0.090	0.288
Educ.: bachelor degree in college	0.481	0.501	0.477	0.501	0.398	0.491
Educ.: master degree	0.131	0.339	0.172	0.379	0.157	0.365
Educ.: doctoral degree	0.025	0.157	0.006	0.076	0.006	0.078
Educ.: professional degree (JD, MD)	0.025	0.157	0.011	0.107	0.024	0.154
observations	N=	160	N=	174	N=	166
	Panel b) INFO					
	Equity		Inequity		Overclaiming	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Age	38.437	12.483	38.508	12.715	38.303	12.696
Male	0.469	0.500	0.456	0.499	0.435	0.497
Educ.: less than high school degree	0.012	0.108	0.000	0.000	0.003	0.054
Educ.: high school graduate	0.076	0.265	0.048	0.214	0.094	0.292
Educ.: some college but no degree	0.178	0.383	0.213	0.410	0.171	0.377
Educ.: associate degree in college	0.105	0.307	0.111	0.315	0.118	0.323
Educ.: bachelor degree in college	0.466	0.500	0.411	0.493	0.432	0.496
Educ.: master degree	0.128	0.335	0.180	0.385	0.144	0.352
Educ.: doctoral degree	0.012	0.108	0.012	0.109	0.026	0.161
Educ.: professional degree (JD, MD)	0.023	0.151	0.024	0.153	0.012	0.108
observations	N=	343	N=	333	N=	340

Notes: Personal characteristics of subjects in Experiment 2 (mean and standard deviation), split by the different treatments (Equity norm: Equity, Inequity, and Overclaiming times NO INFO and INFO). We do not observe statistically significant differences between the six treatments for age (Kruskal-Wallis test: p=0.502), gender (χ^2-test: p=0.749), or education (χ^2-test: p=0.197). NO INFO pools the the observations from treatments NORMATIVE and EMPIRICAL. If taken individually, the number of observations in NORMATIVE are n=172, n=176, and n=163 for the Equity-, Inequity-, and Overclaim-treatments, respectively; the corresponding number of observations in EMPIRICAL are n=171, n=157, and n=177.

Table A.5: Social appropriateness ratings – regression results

	Social Appropriateness score (1 – 4)		
	(1)	(2)	(3)
INFO	-0.154^{***}		-0.261^{***}
	(0.050)		(0.088)
Inequity		-0.109^{*}	-0.268^{***}
		(0.058)	(0.100)
Overclaim		-0.109^{*}	-0.163
		(0.058)	(0.101)
INFO × Inequity			0.232^{*}
			(0.122)
INFO × Overclaiming			0.078
			(0.123)
Constant	1.348^{***}	1.215^{***}	1.472^{***}
	(0.186)	(0.176)	(0.195)
N	1,516	1,516	1,516

F-test: coeff. for INFO + INFO × Inequity = Inequity; p=0.140
F-test: coeff. for INFO + INFO × Overclaiming = Overclaiming; p=0.908

Notes: OLS results of regressing the social appropriateness score (coded as "Very Socially Inappropriate"=1, "Somewhat Socially Inappropriate"=2, "Somewhat Socially Appropriate"=3, "Very Socially Appropriate"=4) on indicators for the norm information treatments (INFO – pools EMPIRICAL and NORMATIVE) and the equity nature of the lie (inequity, overclaim); the baseline category is therefore a subject in treatment NO INFO and an equity-based punishment scenario with lie size=1. Additional control variables include age, gender, education. Robust standart errors in paratheses. */**/***: p<0.10/0.05/0.01.

Appendix B: Information regarding pre-registration

Both experiments have been pre-registered on `http://aspredicted.org`. Below are the links to the anonymized pre-registration files. The files can be de-anonymized and made publicly accessible upon publication.

Links:

- Experiment 1 – `http://aspredicted.org/blind.php?x=u94s84`
- Experiment 2 – `http://aspredicted.org/blind.php?x=kp3kd3`

Notes regarding data collection for Experiment 1:

- Data for Experiment 1 was collected in April, May, and November 2019.
- A first batch of 106 observations was collected in April 2016 to test functionality of the experimental interface with subjects from the representative pool provided by the online survey firm employed by us. Based on these observations, we pre-registered the study and started to collect additional observations in May.
- Due to miscommunication with the data collection company, we obtained almost twice as many observations (987 instead of the targeted 500) for the punishers by May.
- Upon inspecting the data, we found that the quote-based sample was representative of the US population across our defined age and gender-bins but not for the cross product of those bins (e.g., 50-59 years old females). We then approached the survey firm to rectify this issue. By an courtesy agreement, it sampled additional observations in November in order to have a representative sample also along these cross bins. This then yielded or final sample with a total of 1,240 observations.
- Given this windfall in statistical power and the resulting improvement of the precision of our estimates (see, e.g., Maxwell et al., 2008) and to avoid a waste of resources, we decided to utilize the full sample. As a robustness check, we also repeated our analysis while excluding the 106 observations collected in April before we pre-registered Experiment 1. All results are similar in term of significance and magnitude to those stated in the main text.
- For all punisher observations, we collected observations for the potentially punished active players according to our stated 15:1-ratio-rule (see Footnote 6).

Appendix C: Instructions

In the following, we display screenshots for all of our experiments. Specifically, we present them in the following order:

1. Instructions for the pre-experiment of Experiment 1
2. Instructions for the main experiment of Experiment 1
3. Instructions for the Experiment 2

Notes:

- In order to facilitate a comprehensive display, we omit repeating elements such as "continue"-buttons.
- If a screen does not fit a single page in this appendix, it is split over two pages and the screen number gets ab an "a" or "b"-suffix for the first and second part of the screen (e.g., "Screen 7a).

C.1: Instructions for the pre-experiment of Experiment 1

The following pages display the screens for the pre-experiment of Experient 1 (see Figure 1). Note that subjects had different roles in the experiment: "liars" and "victims". We present the screens in the order in which they were presented to subjects. Some screens were the same, independent of a subject's role. If not otherwise noted in the respective captions, screens are shown to subjects in both roles.

Figure C.1.1: Pre-experiment (for Experiment 1), Screen #1a

WELCOME!

In the following, you will be asked to answer questions and make some decisions.

Note: Sometimes, it will take a moment until you can click the "continue"-button at the bottom.

Confidentiality

The data collected in this study do not include any personally identifiable information about you. Your data will be kept separate from your MTurk ID. By participating, you understand and agree that the research data gathered during this study will be used by ETH Zurich and aggregated results will be published.

Contact Information

This scientific study is conducted by the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland. If you have any questions concerning this study, please write to descil@ethz.ch.

We recommend participating in this study only from a computer, not a touchscreen device.

Please click each of the boxes below to indicate that you have understood the rules and accept them.

You will not be able to return to previous pages.

You have to finish the entire study to get your additional payment.

The last page of this study will provide you with a verification code. Please provide the verification code on the MTurk-page where you accepted the assignment. Without this code, we cannot approve your submission and cannot pay you for it.

Please read the instructions carefully. There will be simple control questions which will check your understanding. If you do not answer these questions correctly you will not be able to finish the entire study.

I am a US resident and 18 years or older.

I would like to participate in this study and agree to the above rules.

In this study you can earn earnings in addition to the announced wage (payed as a bonus). However, the transfer of this additional payment will take up to a week.

Figure C.1.2: Pre-experiment (for Experiment 1), Screen #1b

RULES IN THIS STUDY

You have received $0.50 for participating in this study and you can earn an additional bonus.
In this study, you will perform a task and make decisions, which can influence your bonus payment. Any payment will only occur upon completion of the study.

Please read the rules carefully to maximize your earnings:

Your decisions

- The task is to count numbers. You are requested to do so for about seven minutes. You will receive a bonus of $1.50 for its completion.
- Upon completion, you will be paired with another participant who also finished the task.
- One of you will be chosen at random to make an additional decision. The details about this decision will be given to you in time.
- Both of you will enter a lottery with the chance to win an additional bonus of $2.50 (in addition to the $1.50 you can earn for completing the task).
- Everyone who decides to participate has equal chances to win this bonus.
- The experiment ends with some demographic questions.

Your payment

- $0.50 for participation.
- A bonus payment of up to $4.00 (up to $1.50 for completion and $2.50 as an additional bonus), depending on the outcome of the study.
- The final bonus is subject to decisions made by you and others in throughout the experiment as well as the outcome of the lottery.

If you do not agree with the above, please close this window.

By clicking the button below I agree that I have understood the above rules.

Figure C.1.3: Pre-experiment (for Experiment 1), Screen #1b

RULES IN THIS STUDY

You have received $0.50 for participating in this study and you can earn an additional bonus.
In this study, you will perform a task and make decisions, which can influence your bonus payment. Any payment will only occur upon completion of the study.

Please read the rules carefully to maximize your earnings:

Your decisions

- The task is to count numbers. You are requested to do so for about seven minutes. You will receive a bonus of $1.50 for its completion.
- Upon completion, you will be paired with another participant who also finished the task.
- One of you will be chosen at random to make an additional decision. The details about this decision will be given to you in time.
- Both of you will enter a lottery with the chance to win an additional bonus of $2.50 (in addition to the $1.50 you can earn for completing the task).
- Everyone who decides to participate has equal chances to win this bonus.
- The experiment ends with some demographic questions.

Your payment

- $0.50 for participation.
- A bonus payment of up to $4.00 (up to $1.50 for completion and $2.50 as an additional bonus), depending on the outcome of the study.
- The final bonus is subject to decisions made by you and others in throughout the experiment as well as the outcome of the lottery.

If you do not agree with the above, please close this window.

By clicking the button below I agree that I have understood the above rules.

Figure C.1.4: Pre-experiment (for Experiment 1), Screen #3

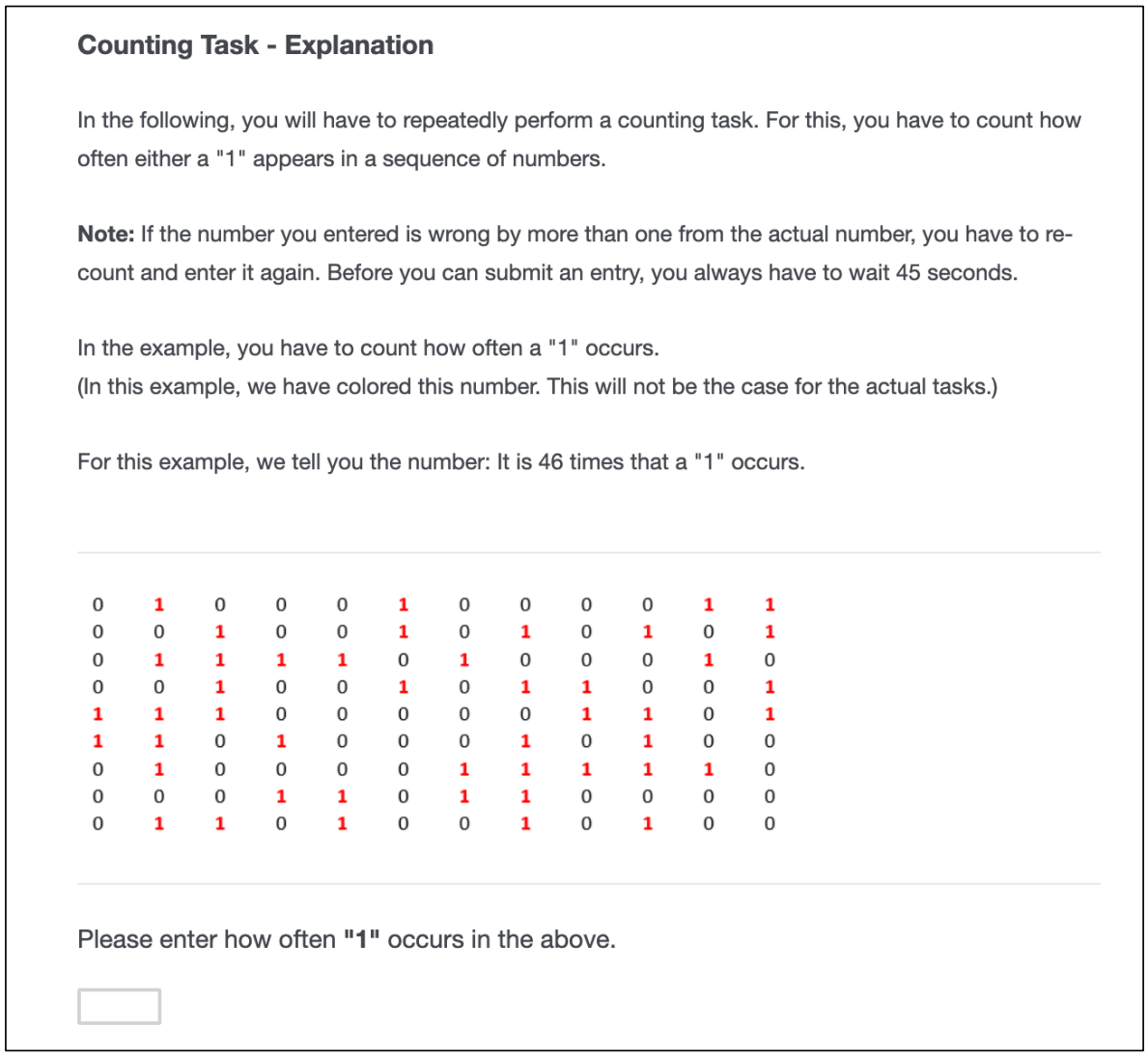


Figure C.1.5: Pre-experiment (for Experiment 1), Screen #4

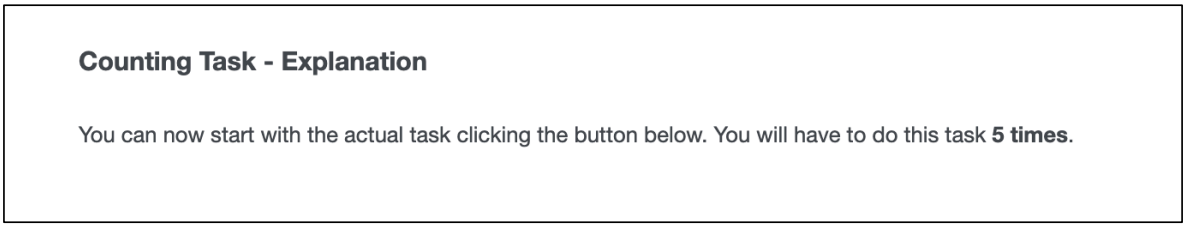


Figure C.1.6: Pre-experiment (for Experiment 1), Screen #5 (five such screens, "Counting Task 1" – "Counting Task 5")

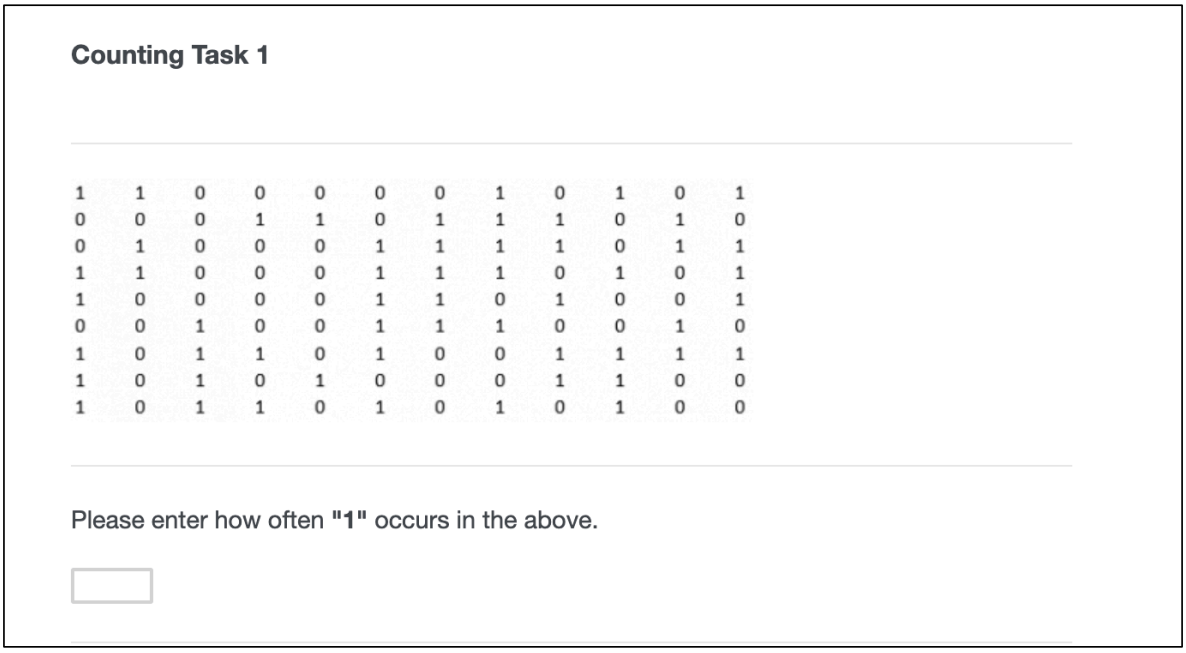


Figure C.1.7: Pre-experiment (for Experiment 1), Screen #6a ("liars" only)

Earning the additional bonus - Explanation

You will now be paired with another participant who also finished the task. Both of you are given the chance to win an additional bonus. However, only one of you will actually win it. Who will win it is determined by chance.

How it works:

- You have been randomly chosen to toss a fair, five-sided die (with numbers "1", "2", "3", "4", "5"). You will do so on the next page.
- You are eligible for as many lottery tickets as the die shows.
- We will ask you to report what the die shows.
- You will receive the number of lottery tickets you report.
- The other participant with whom you are paired has been randomly chosen to not toss a die. Instead, this participant gets a fixed amount of 3 lottery tickets.
- Your and the other participant's lottery tickets will be put into a (virtual) raffle. One ticket will be randomly drawn. The person to whom this ticket belongs earns the bonus of $2.50.

Figure C.1.8: Pre-experiment (for Experiment 1), Screen #6b ("liars" only)

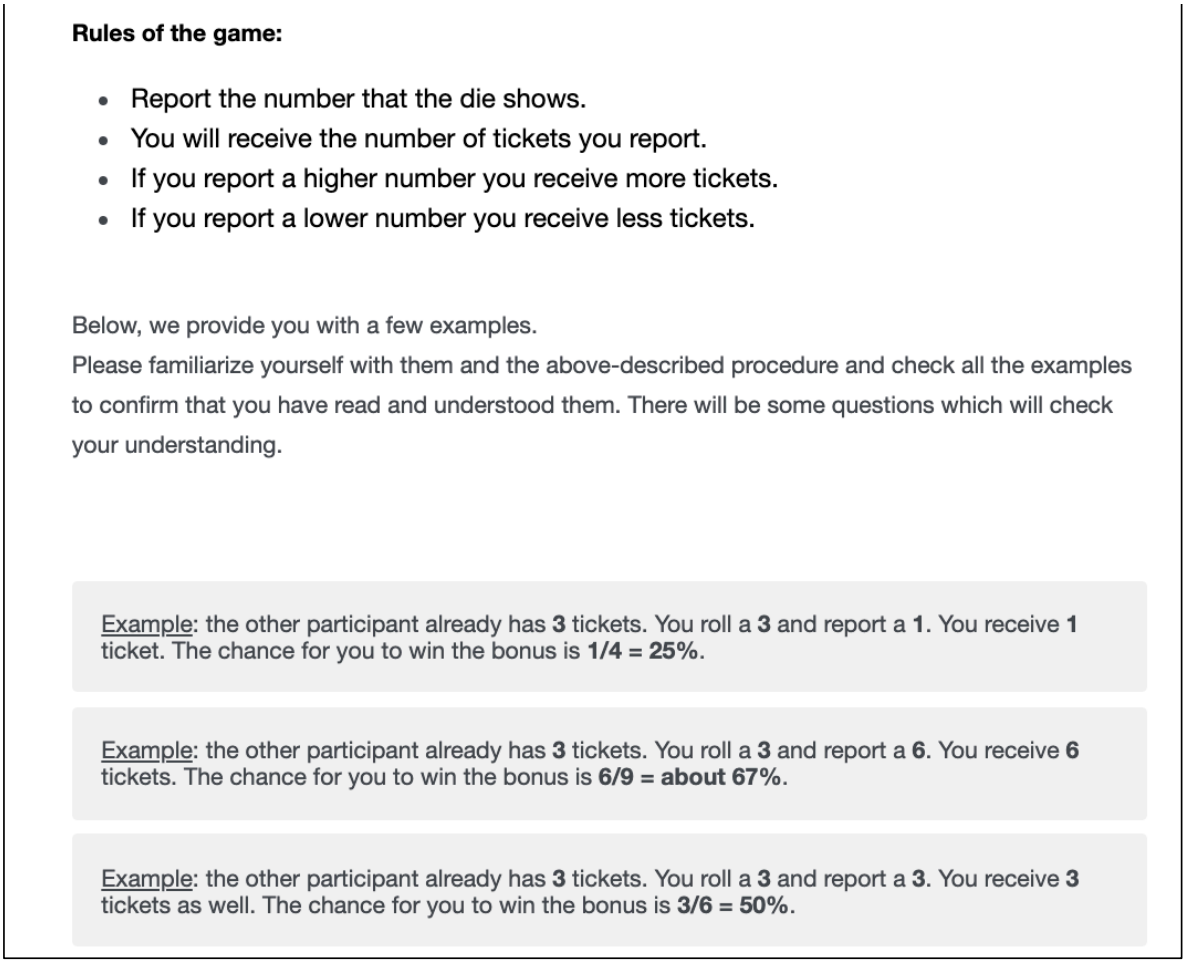


Figure C.1.9: Pre-experiment (for Experiment 1), Screen #7 (”liars” only)

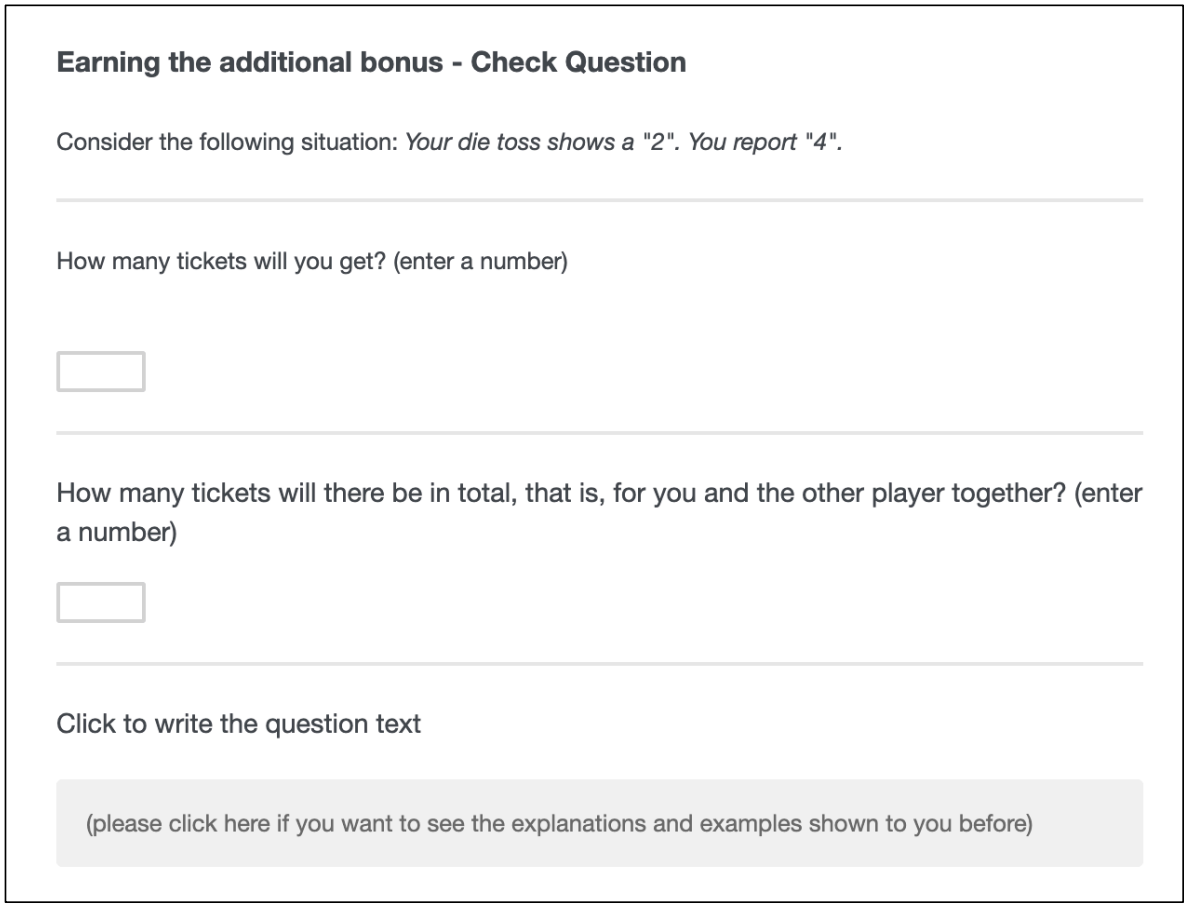


Figure C.1.10: Pre-experiment (for Experiment 1), Screen #8 (”liars” only)

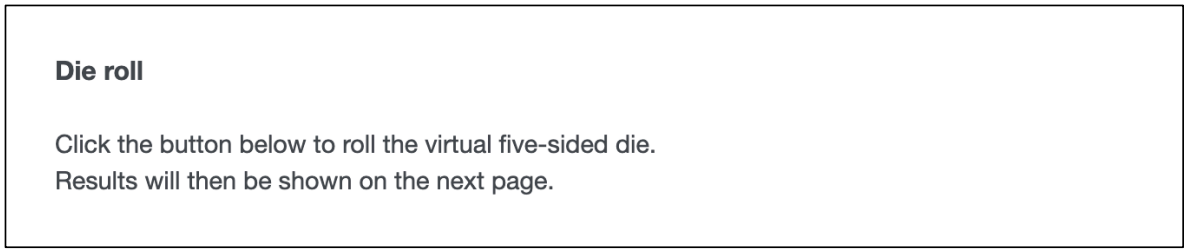


Figure C.1.11: Pre-experiment (for Experiment 1), Screen #9 (”liars” only)

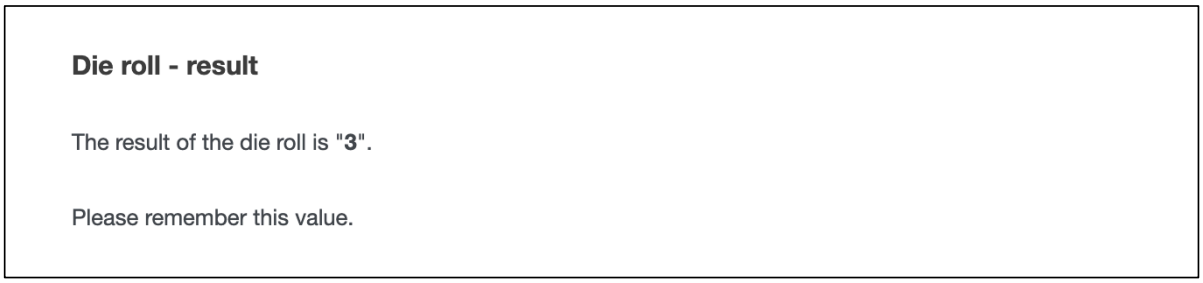
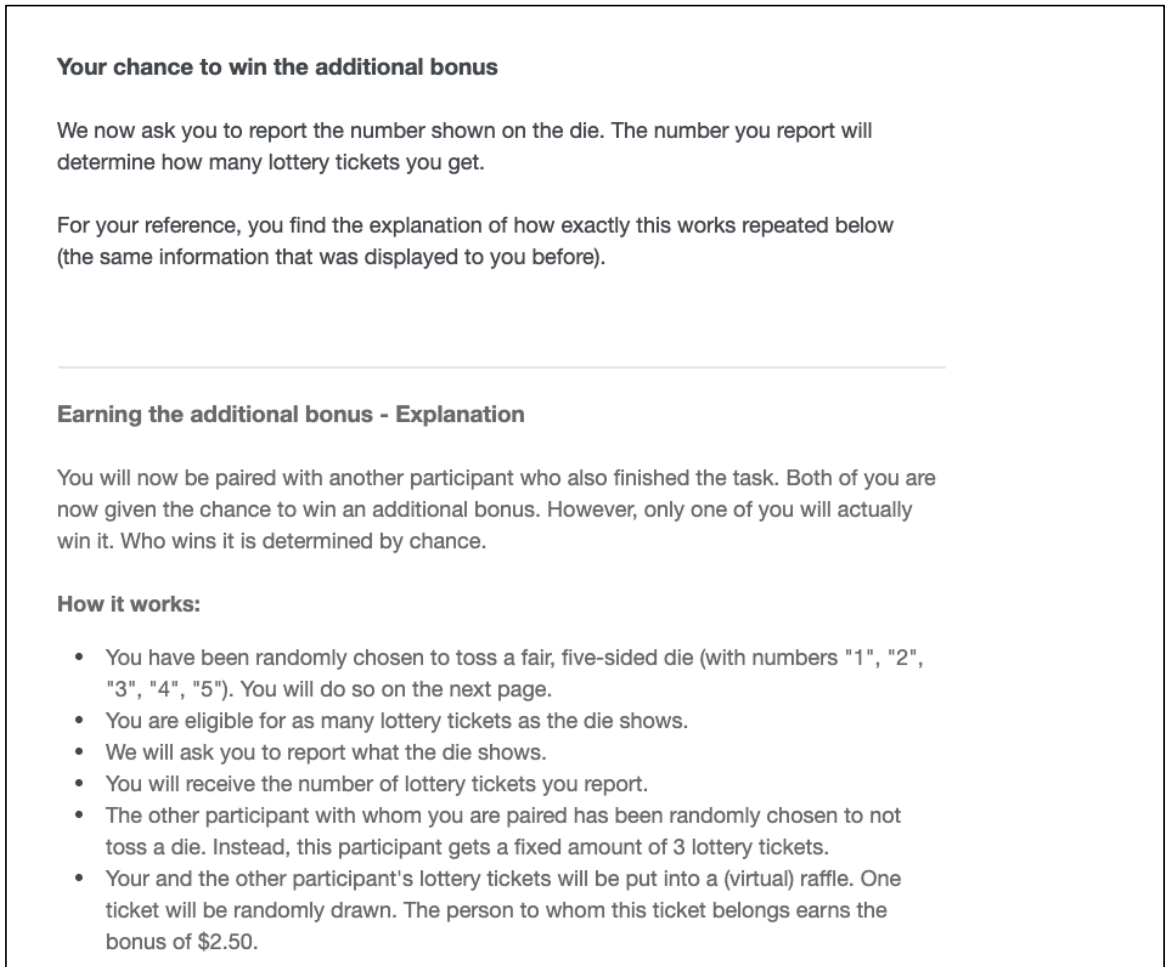


Figure C.1.12: Pre-experiment (for Experiment 1), Screen #10a ("liars" only)



Your chance to win the additional bonus

We now ask you to report the number shown on the die. The number you report will determine how many lottery tickets you get.

For your reference, you find the explanation of how exactly this works repeated below (the same information that was displayed to you before).

Earning the additional bonus - Explanation

You will now be paired with another participant who also finished the task. Both of you are now given the chance to win an additional bonus. However, only one of you will actually win it. Who wins it is determined by chance.

How it works:

- You have been randomly chosen to toss a fair, five-sided die (with numbers "1", "2", "3", "4", "5"). You will do so on the next page.
- You are eligible for as many lottery tickets as the die shows.
- We will ask you to report what the die shows.
- You will receive the number of lottery tickets you report.
- The other participant with whom you are paired has been randomly chosen to not toss a die. Instead, this participant gets a fixed amount of 3 lottery tickets.
- Your and the other participant's lottery tickets will be put into a (virtual) raffle. One ticket will be randomly drawn. The person to whom this ticket belongs earns the bonus of $2.50.

Figure C.1.13: Pre-experiment (for Experiment 1), Screen #10b ("liars" only)

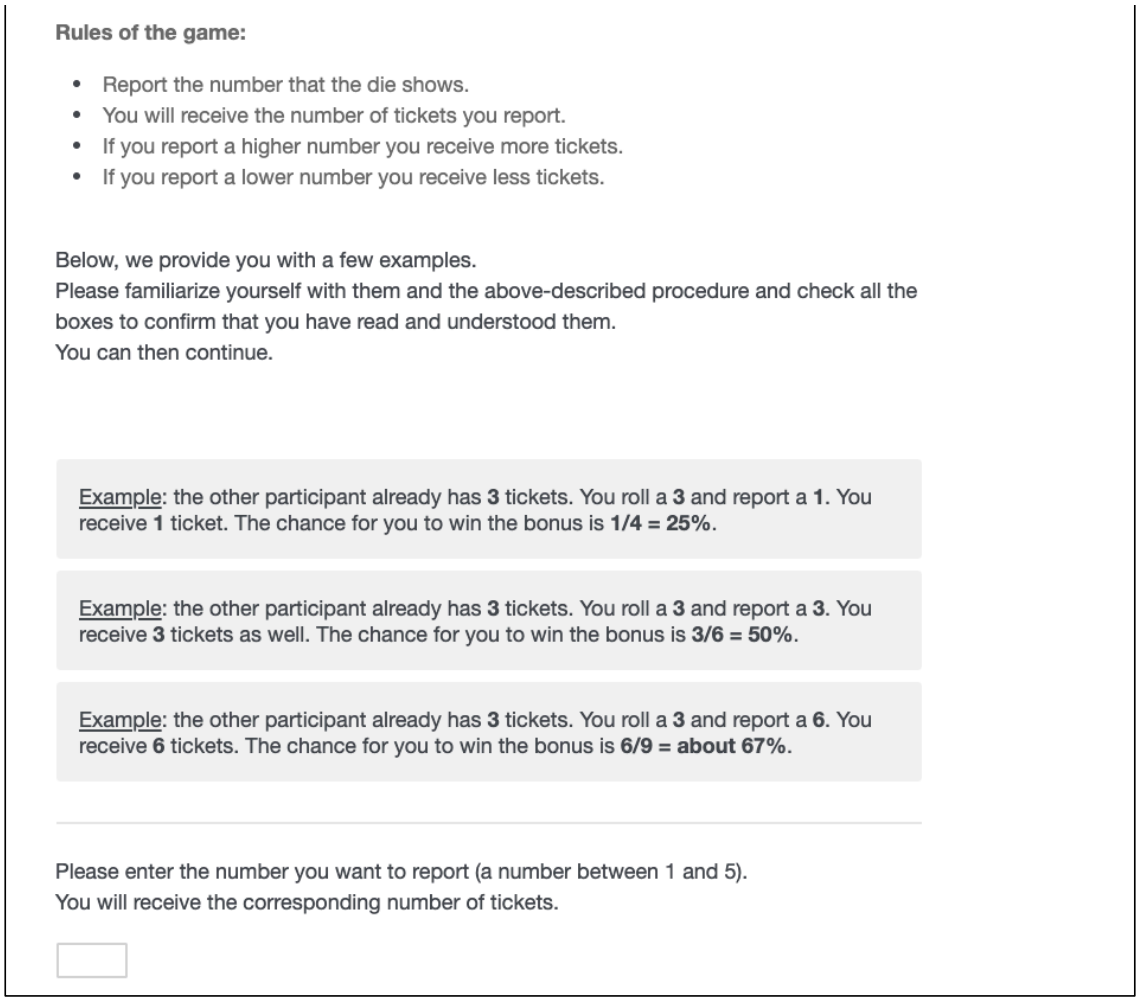


Figure C.1.14: Pre-experiment (for Experiment 1), Screen #11 ("liars" only)

Possibility to revise your report

One the previous page, you reported the following number:
This is what your die actually showed:

Punishment:

- The information above will be used to let other participants in another study decide whether they want to punish you.
- These other participants do not play the game you have just played but they know its rules.
- They also know that the rule was to report the number which the die showed.
- They can punish you only if you reported a higher number than you actually rolled.
- Punishment will be in the form of "punishment points". Each punishment point is worth minus $0.30 and you can receive up to five punishment points.
- The sum of your punishment points will be subtracted from the $1.50 you receive as an individual bonus for having successfully completed the counting task (not from the $2.50-bonus which either you or the other participant you have been paired with can receive).

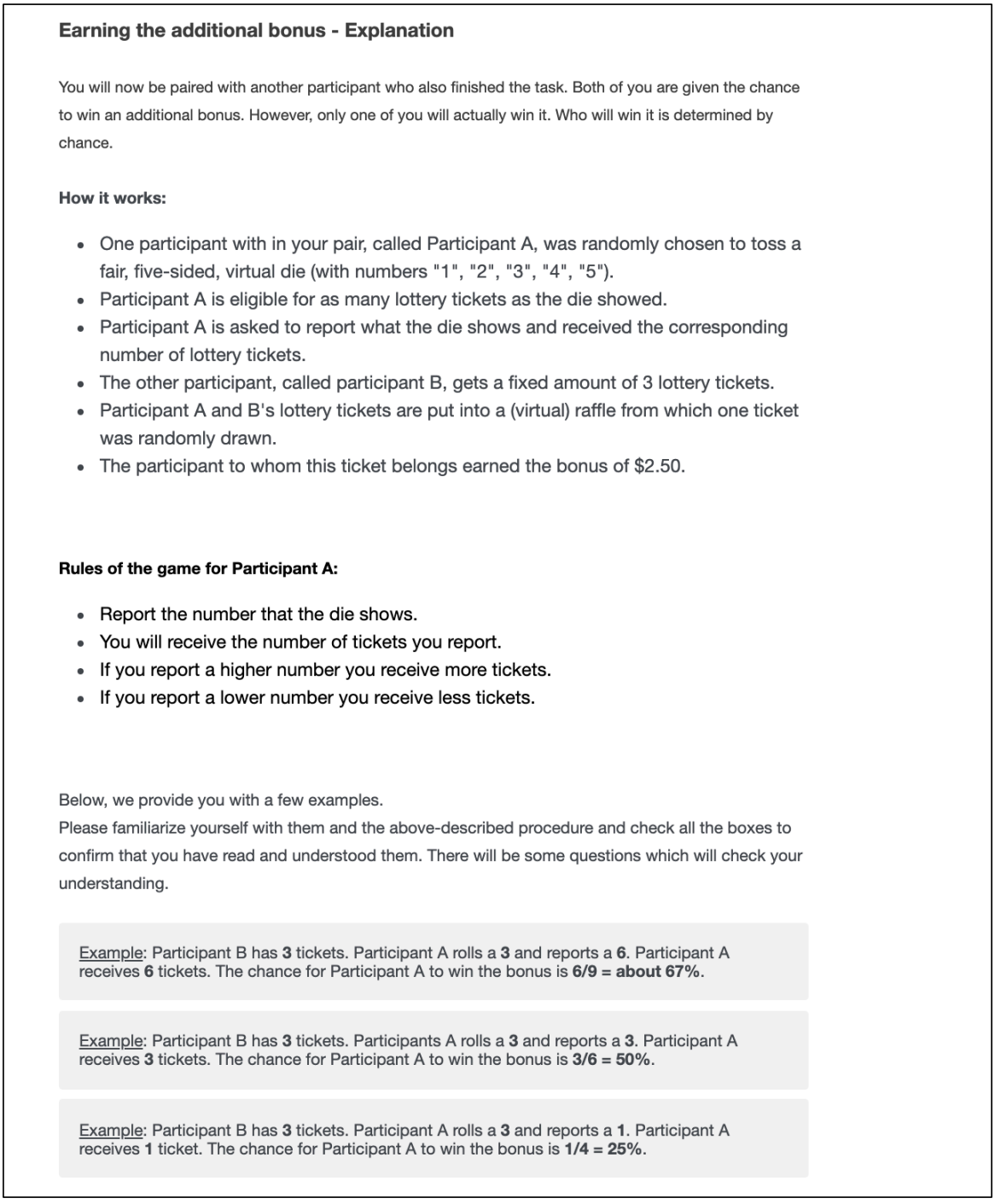
Example: You receive two punishment points. You will earn an individual bonus of 1.50-2x0.30=$0.90.

Example: You receive five punishment points. You will earn an individual bonus of 1.50-5x0.30=$0.00

Revision:
You can now revise the number you reported. If you do so, this revised number - not your initially reported number - will be used to determine the number of lottery tickets for winning the additional $2.50-bonus (the other player with whom you are paired still receives three tickets).
Also, your punishment will be based only on the revised number (the other participants who can punish you will never learn your originally reported number).

If you want to revise your report, please do so below by entering a new number.
If you do not want to revise your report, you re-enter your initial number.

Figure C.1.15: Pre-experiment (for Experiment 1), Screen #12 ("victims" only)



Earning the additional bonus - Explanation

You will now be paired with another participant who also finished the task. Both of you are given the chance to win an additional bonus. However, only one of you will actually win it. Who will win it is determined by chance.

How it works:

- One participant with in your pair, called Participant A, was randomly chosen to toss a fair, five-sided, virtual die (with numbers "1", "2", "3", "4", "5").
- Participant A is eligible for as many lottery tickets as the die showed.
- Participant A is asked to report what the die shows and received the corresponding number of lottery tickets.
- The other participant, called participant B, gets a fixed amount of 3 lottery tickets.
- Participant A and B's lottery tickets are put into a (virtual) raffle from which one ticket was randomly drawn.
- The participant to whom this ticket belongs earned the bonus of $2.50.

Rules of the game for Participant A:

- Report the number that the die shows.
- You will receive the number of tickets you report.
- If you report a higher number you receive more tickets.
- If you report a lower number you receive less tickets.

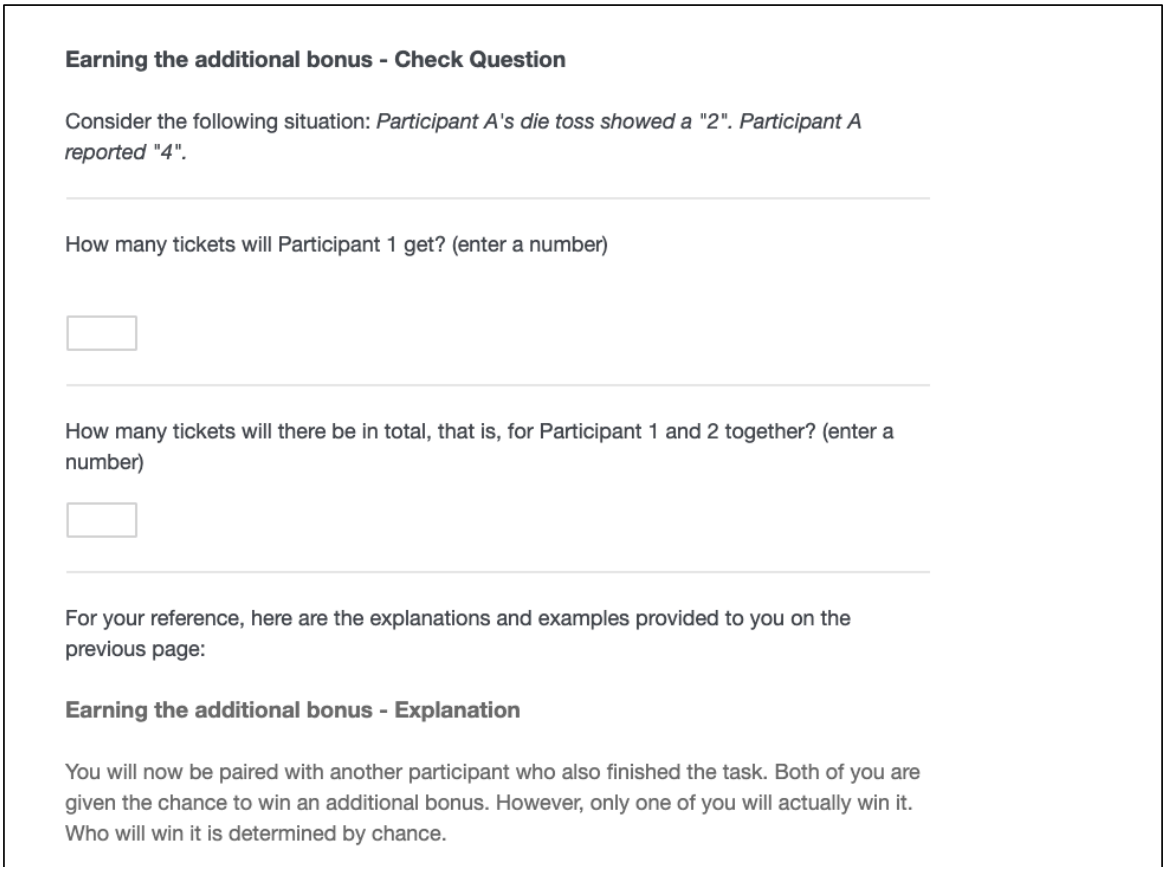
Below, we provide you with a few examples.
Please familiarize yourself with them and the above-described procedure and check all the boxes to confirm that you have read and understood them. There will be some questions which will check your understanding.

Example: Participant B has **3** tickets. Participant A rolls a **3** and reports a **6**. Participant A receives **6** tickets. The chance for Participant A to win the bonus is **6/9 = about 67%**.

Example: Participant B has **3** tickets. Participants A rolls a **3** and reports a **3**. Participant A receives **3** tickets. The chance for Participant A to win the bonus is **3/6 = 50%**.

Example: Participant B has **3** tickets. Participant A rolls a **3** and reports a **1**. Participant A receives **1** ticket. The chance for Participant A to win the bonus is **1/4 = 25%**.

Figure C.1.16: Pre-experiment (for Experiment 1), Screen #13a ("victims" only)



Earning the additional bonus - Check Question

Consider the following situation: *Participant A's die toss showed a "2". Participant A reported "4".*

How many tickets will Participant 1 get? (enter a number)

How many tickets will there be in total, that is, for Participant 1 and 2 together? (enter a number)

For your reference, here are the explanations and examples provided to you on the previous page:

Earning the additional bonus - Explanation

You will now be paired with another participant who also finished the task. Both of you are given the chance to win an additional bonus. However, only one of you will actually win it. Who will win it is determined by chance.

Figure C.1.17: Pre-experiment (for Experiment 1), Screen #13b ("victims" only)

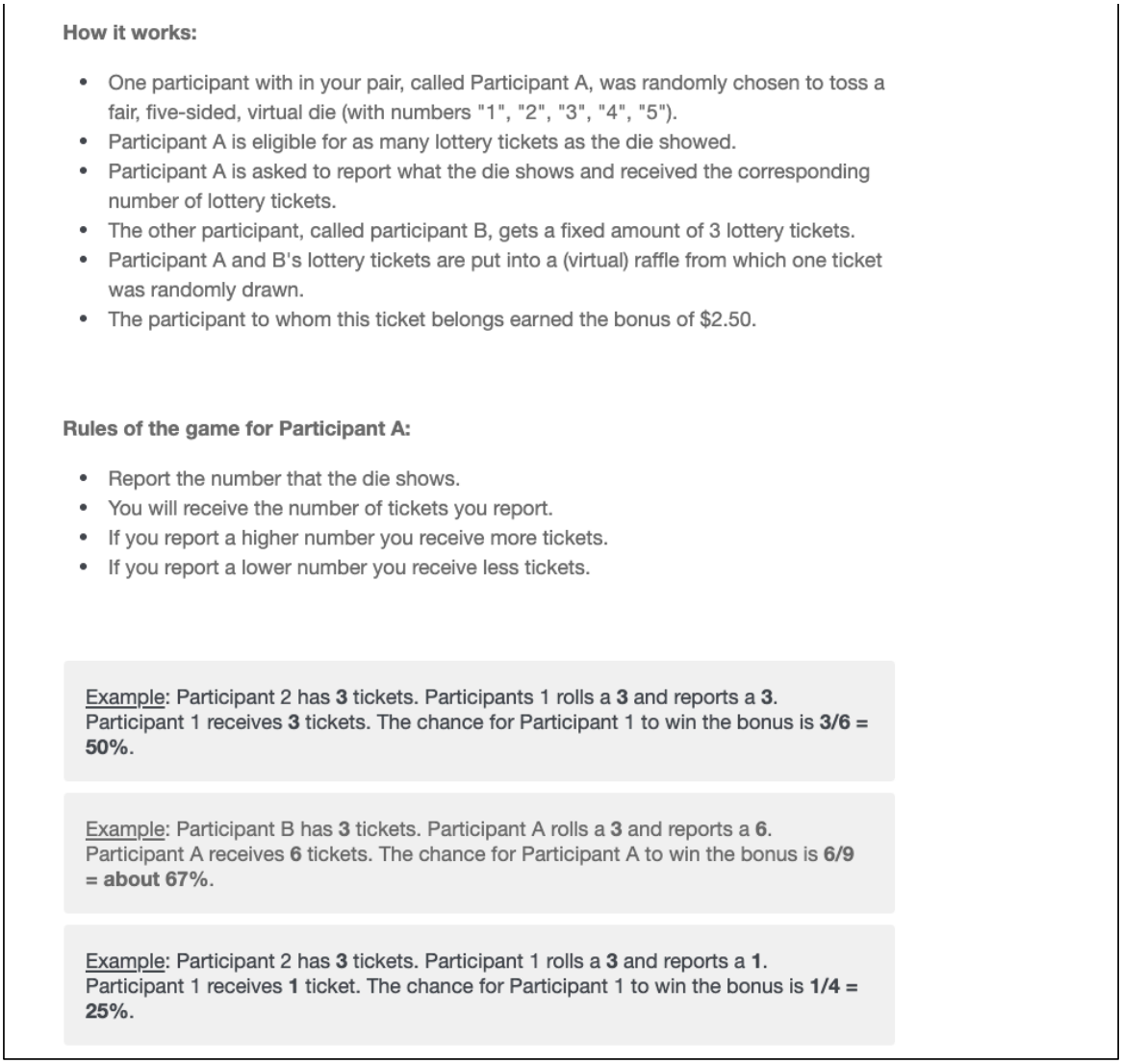
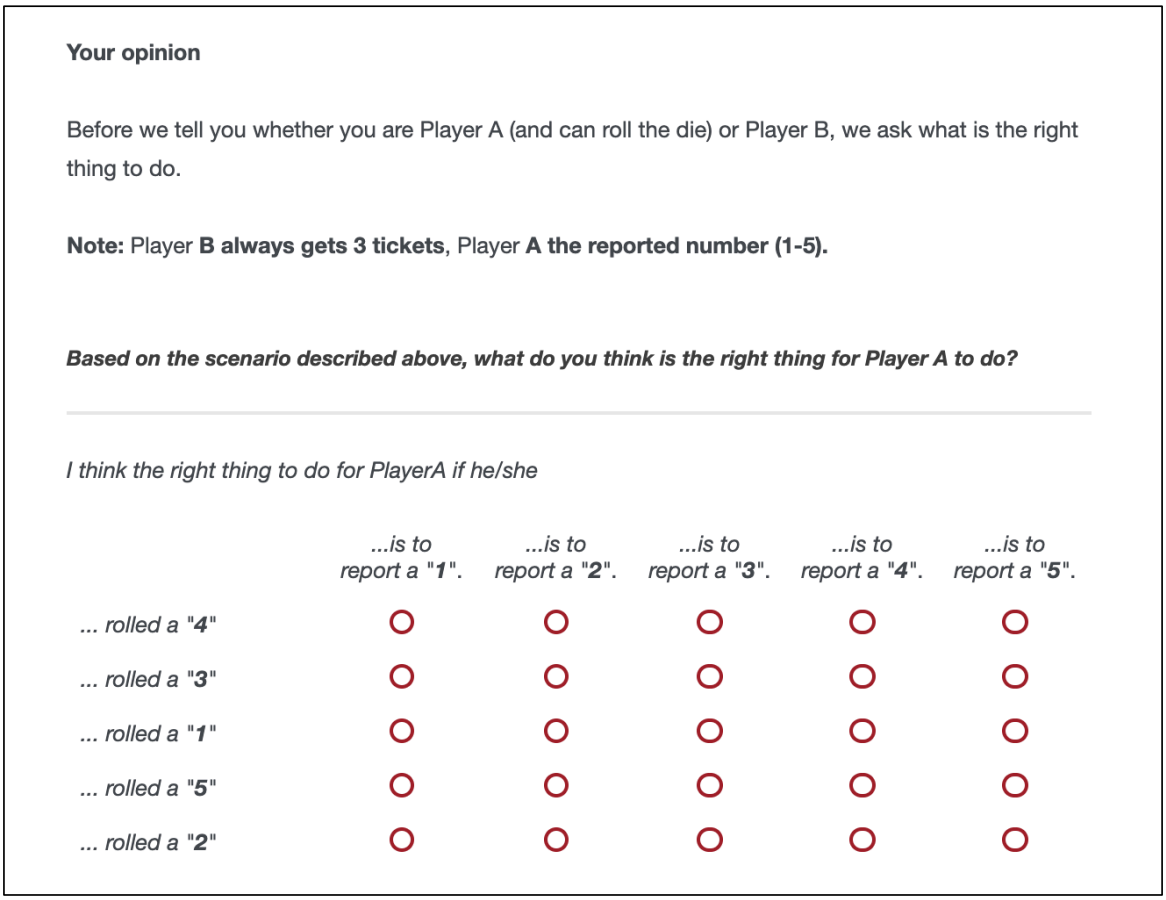


Figure C.1.18: Pre-experiment (for Experiment 1), Screen #14 ("victims" only)



Your opinion

Before we tell you whether you are Player A (and can roll the die) or Player B, we ask what is the right thing to do.

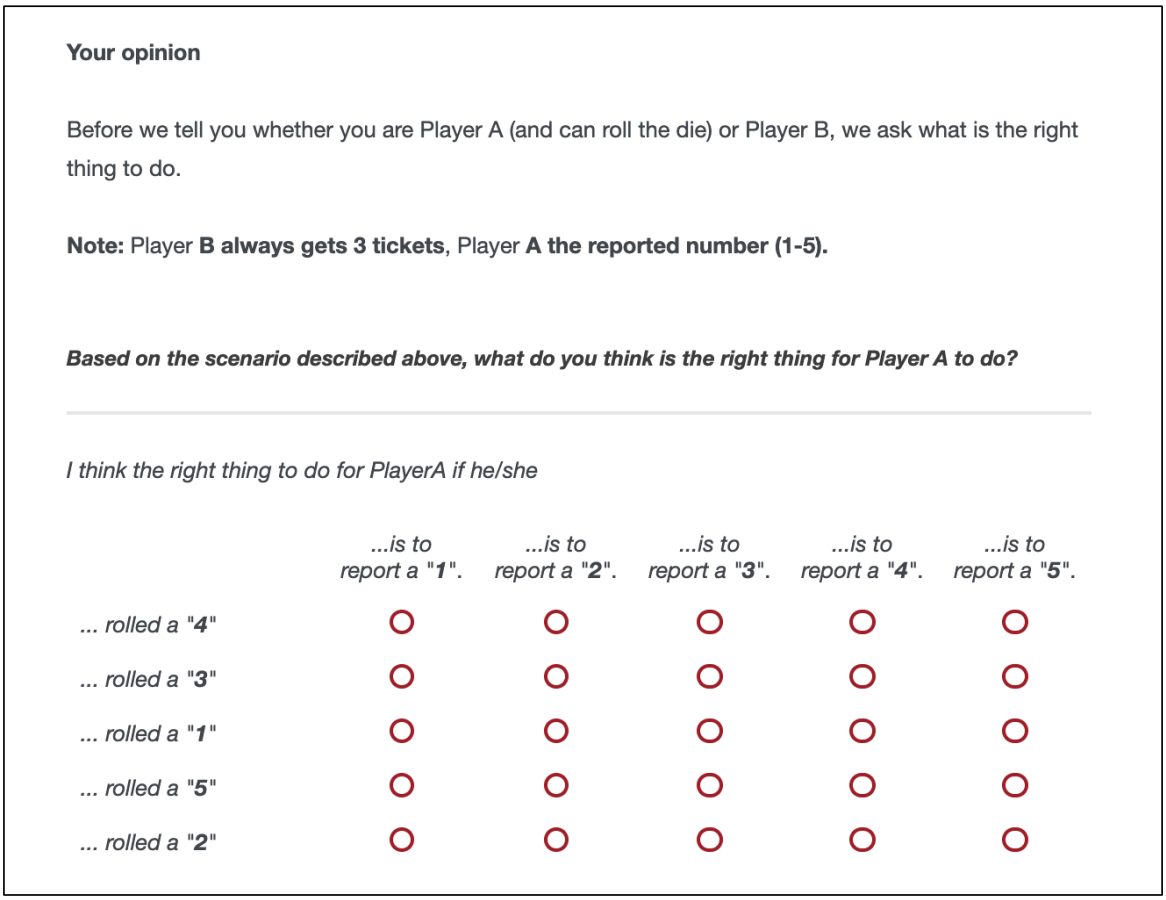
Note: Player **B always gets 3 tickets**, Player **A the reported number (1-5).**

Based on the scenario described above, what do you think is the right thing for Player A to do?

I think the right thing to do for PlayerA if he/she

	*...is to report a "**1**".*	*...is to report a "**2**".*	*...is to report a "**3**".*	*...is to report a "**4**".*	*...is to report a "**5**".*
*... rolled a "**4**"*	○	○	○	○	○
*... rolled a "**3**"*	○	○	○	○	○
*... rolled a "**1**"*	○	○	○	○	○
*... rolled a "**5**"*	○	○	○	○	○
*... rolled a "**2**"*	○	○	○	○	○

Figure C.1.19: Pre-experiment (for Experiment 1), Screen #14 ("victims" only)



Your opinion

Before we tell you whether you are Player A (and can roll the die) or Player B, we ask what is the right thing to do.

Note: Player **B always gets 3 tickets**, Player **A the reported number (1-5).**

Based on the scenario described above, what do you think is the right thing for Player A to do?

I think the right thing to do for PlayerA if he/she

	*...is to report a "**1**".*	*...is to report a "**2**".*	*...is to report a "**3**".*	*...is to report a "**4**".*	*...is to report a "**5**".*
*... rolled a "**4**"*	○	○	○	○	○
*... rolled a "**3**"*	○	○	○	○	○
*... rolled a "**1**"*	○	○	○	○	○
*... rolled a "**5**"*	○	○	○	○	○
*... rolled a "**2**"*	○	○	○	○	○

Figure C.1.20: Pre-experiment (for Experiment 1), Screen #15 ("victims" only)

Your role

You have been randomly chosen to be
Player B.

Figure C.1.21: Pre-experiment (for Experiment 1), Screen #16 ("victims" only)

Earning the additional bonus - Explanation

You will now be paired with another participant who also finished the task. Both of you are given the chance to win an additional bonus. However, only one of you will actually win it. Who will win it is determined by chance.

How it works:

- The other participant has been randomly chosen to toss a fair, five-sided die (with numbers "1", "2", "3", "4", "5").
- The other participant is eligible for as many lottery tickets as the die shows.
- The other participant will be ask to report what the die shows.
- The other participant will receive the number of lottery tickets he/she reports.
- You have been randomly chosen to not toss a die. Instead, you get a fixed amount of 3 lottery tickets.
- Your and the other participant's lottery tickets will be put into a (virtual) raffle. One ticket will be randomly drawn. The person to whom this ticket belongs earns the team bonus of $2.50.

Rules of the game:

- The other participant is supposed to the number that the die shows.
- The other participant receives the number of tickets he/she reports.
- If the other participant reports a higher number he/she receives more tickets.
- If the ioher participant reports a lower number he/she receives less tickets.

The draw will take place when the study is over. You will then receive your bonus payment after a few days.

Figure C.1.22: Pre-experiment (for Experiment 1), Screen #17

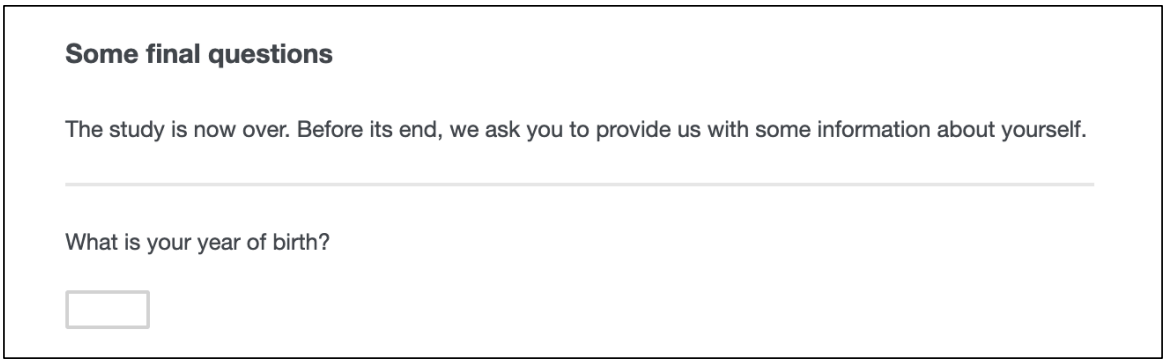
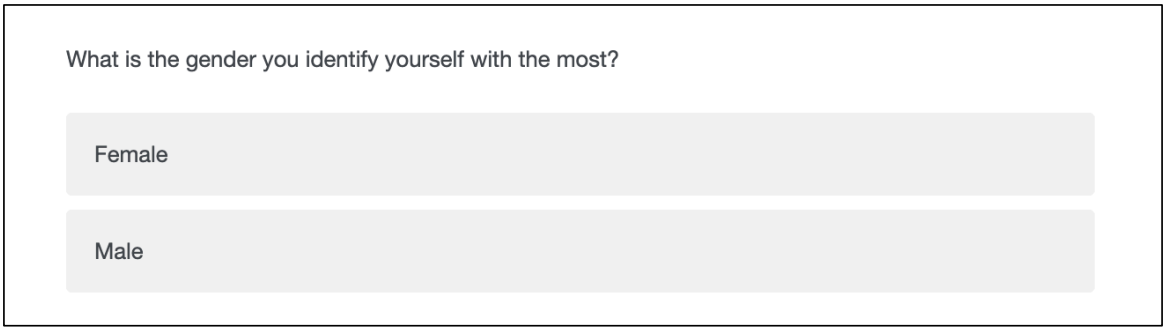


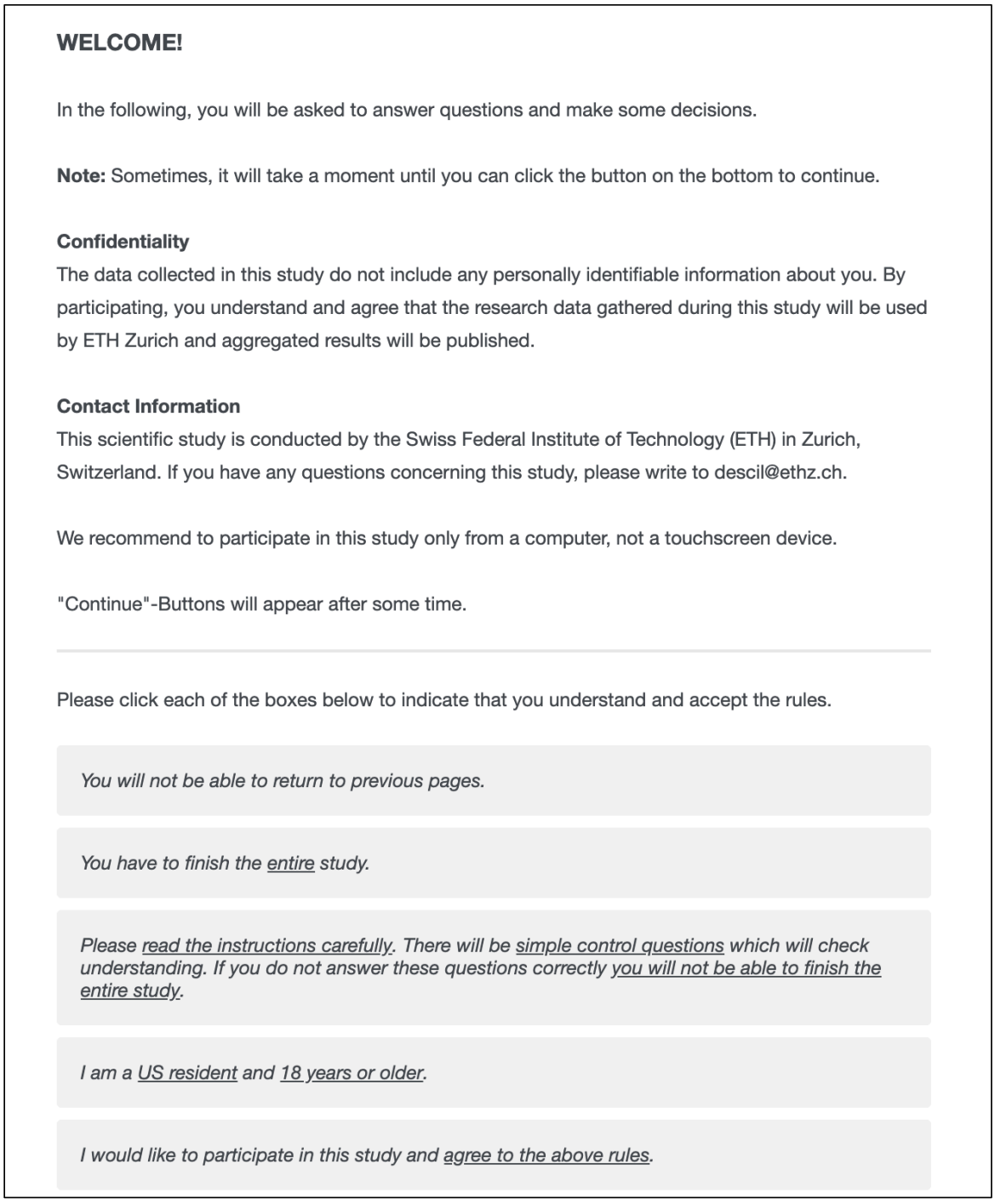
Figure C.1.23: Pre-experiment (for Experiment 1), Screen #18



C.2: Instructions for the main experiment of Experiment 1

The following pages display the screens for the main experiment of Experiment 1 (see Figure 1). All subjects were in the role of punishers but were assigned to different treatment, displaying different norm information (NO INFO, NORMATIVE, and EMPIRICAL). This norm information is displayed on Screen #8, here displaced in Figure C.2.9. This screenshot shows the norm information displayed in treatment NORMATIVE. For EMPIRICAL, the text was adjusted to the one reproduced in the main text; for NO INFO, no information was shown.

Figure C.2.1: Main experiment, Screen #1



WELCOME!

In the following, you will be asked to answer questions and make some decisions.

Note: Sometimes, it will take a moment until you can click the button on the bottom to continue.

Confidentiality
The data collected in this study do not include any personally identifiable information about you. By participating, you understand and agree that the research data gathered during this study will be used by ETH Zurich and aggregated results will be published.

Contact Information
This scientific study is conducted by the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland. If you have any questions concerning this study, please write to descil@ethz.ch.

We recommend to participate in this study only from a computer, not a touchscreen device.

"Continue"-Buttons will appear after some time.

Please click each of the boxes below to indicate that you understand and accept the rules.

You will not be able to return to previous pages.

You have to finish the entire study.

Please read the instructions carefully. There will be simple control questions which will check understanding. If you do not answer these questions correctly you will not be able to finish the entire study.

I am a US resident and 18 years or older.

I would like to participate in this study and agree to the above rules.

Figure C.2.2: Main experiment, Screen #2

Purpose of this study

In the current study (henceforth called "**Study C**"), you will be asked to judge how participants in another, the previous study have behaved. In this previous study (henceforth called "**Study P**"), participants could earn real money, potentially through cheating. Your judgment of the observed behavior will consist of deciding whether and how much to punish such behavior. Your inflicted punishment will have real monetary consequences for participants in Study P.

We will explain to you the situation in two parts:

Explanation Part a) - Situation of other participants in Study P: an explanation of what participants in the other study had to do, how they could cheat and earn money.

Explanation Part b) - **Punishment by you:** an explanation of how you can punish the other participants.

Note: please read the following information carefully. Participant in Study P have already participated but will not receive their payment until the current Study C is over. Your punishment will have real consequences on how much money participants in Study P can earn.

Figure C.2.3: Main experiment, Screen #3

Explanation A - Situation for other participants

In **Study P**, every participant was able to earn $2 plus an additional bonus of $2.5 for completing a counting task. The task was to count either how many zeros or ones appeared in a picture similar to the one shown below.

1	1	0	0	0	0	0	1	0	1	0	1
0	0	0	1	1	0	1	1	1	0	1	0
0	1	0	0	0	1	1	1	1	0	1	1
1	1	0	0	0	1	1	1	0	1	0	1
1	0	0	0	0	1	1	0	1	0	0	1
0	0	1	0	0	1	1	1	0	0	1	0
1	0	1	1	0	1	0	0	1	1	1	1
1	0	1	0	1	0	0	0	1	1	0	0
1	0	1	1	0	1	0	1	0	1	0	0

In total, participants in **Study P** had to do complete five such counting tasks.

(After a while, the button to proceed will appear below. If you have read the above, you can click it.)

Figure C.2.4: Main experiment, Screen #4

Explanation Part a) - Situation for other participants

After receiving $2 for completing the counting task, participants in **Study P** were paired with one other participant (who also finished the task). They were then given the opportunity to earn an **additional bonus of $2.5**. The rules were as follows:

How it works:

- One participant within a pair, called Participant A, was randomly chosen to toss a fair, five-sided, virtual die (with numbers "1", "2", "3", "4", "5").
- Participant A was then eligible for as many lottery tickets as the die showed.
- Participant A was asked to report what the die shows and received the corresponding number of lottery tickets.
- The other participant of the pair, called Participant B, always received a fixed amount of 3 lottery tickets without rolling the die.
- Participant A and B's lottery tickets were put into a (virtual) raffle from which one ticket was randomly drawn.
- The participant to whom this ticket belongs earned the bonus of $2.50.

The rules of the game given to Participants A was that they should report the number shown on the die. However, the participants could cheat and report any number between 1 and 5. Thus, reporting a higher or lower number yielded more or fewer tickets, respectively.

Below are three examples:
(please read and then click each of them in order to proceed)

Example: A Participant B has **3** tickets. Participants A rolls a **3** and reports a **3**. Participant A receives **3** tickets. The chance for Participant A to win the bonus is **3/6 = 50%**.

Example: A Participant B has **3** tickets. Participant A rolls a **3** and reports a **1**. Participant A receives **1** ticket. The chance for Participant A to win the bonus is **1/4 = 25%**.

Example: A Participant B has **3** tickets. Participant A rolls a **3** and reports a **5**. Participant A receives **5** tickets. The chance for Participant A to win the bonus is **5/8 = about 63%**.

Figure C.2.5: Main experiment, Screen #5

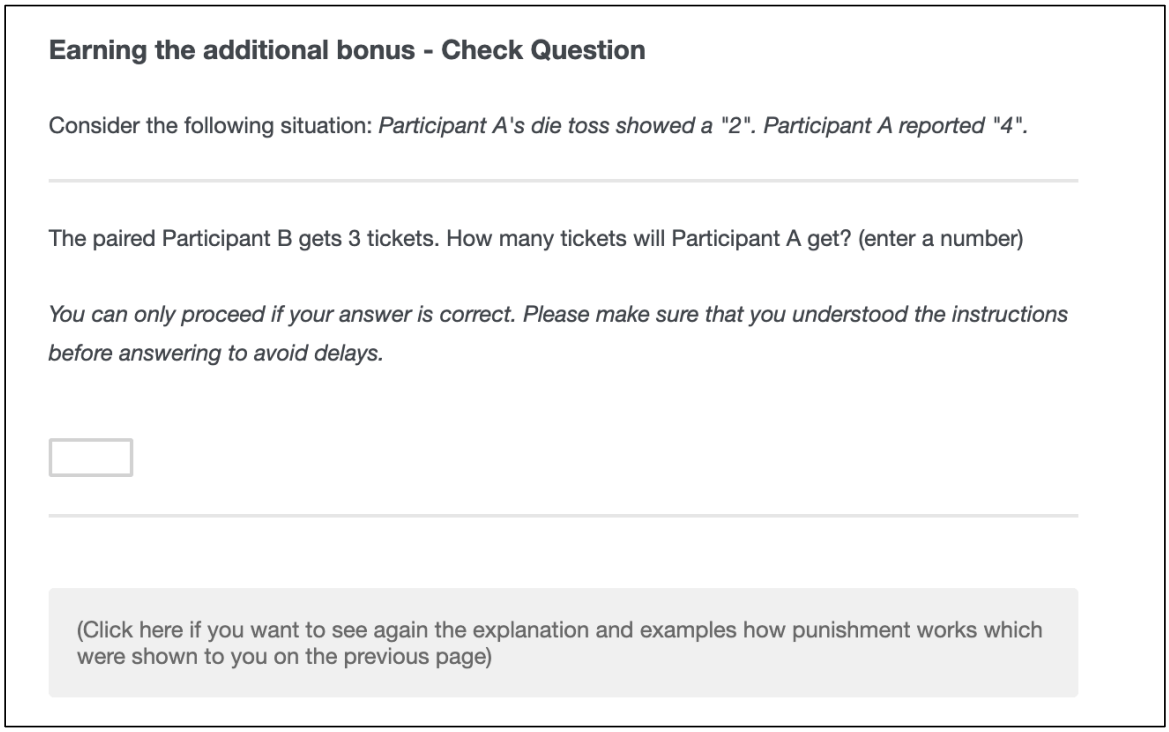
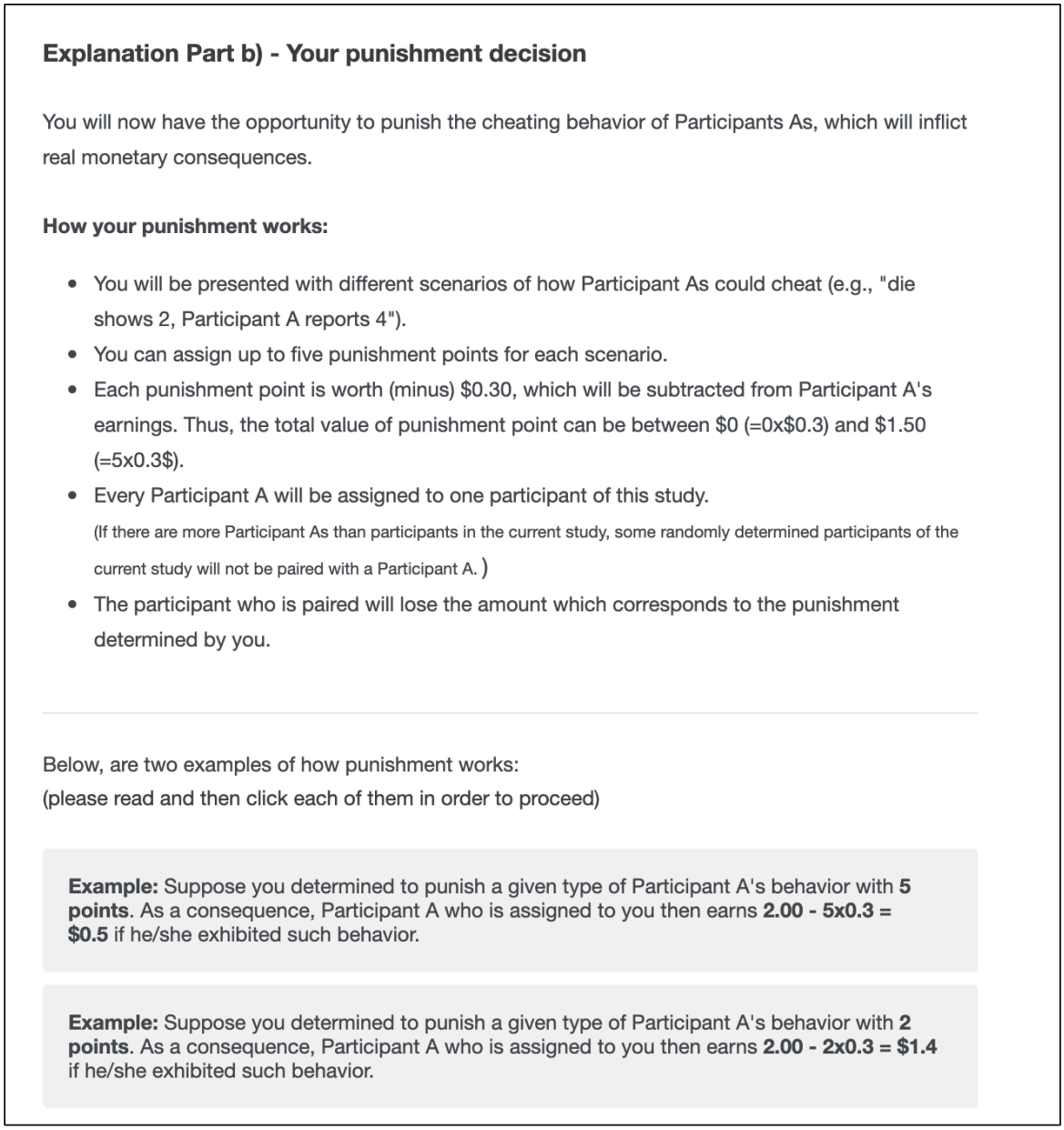


Figure C.2.6: Main experiment, Screen #6



Explanation Part b) - Your punishment decision

You will now have the opportunity to punish the cheating behavior of Participants As, which will inflict real monetary consequences.

How your punishment works:

- You will be presented with different scenarios of how Participant As could cheat (e.g., "die shows 2, Participant A reports 4").
- You can assign up to five punishment points for each scenario.
- Each punishment point is worth (minus) $0.30, which will be subtracted from Participant A's earnings. Thus, the total value of punishment point can be between $0 (=0x$0.3) and $1.50 (=5x0.3$).
- Every Participant A will be assigned to one participant of this study.
  (If there are more Participant As than participants in the current study, some randomly determined participants of the current study will not be paired with a Participant A.)
- The participant who is paired will lose the amount which corresponds to the punishment determined by you.

Below, are two examples of how punishment works:
(please read and then click each of them in order to proceed)

Example: Suppose you determined to punish a given type of Participant A's behavior with **5 points**. As a consequence, Participant A who is assigned to you then earns **2.00 - 5x0.3 = $0.5** if he/she exhibited such behavior.

Example: Suppose you determined to punish a given type of Participant A's behavior with **2 points**. As a consequence, Participant A who is assigned to you then earns **2.00 - 2x0.3 = $1.4** if he/she exhibited such behavior.

Figure C.2.7: Main experiment, Screen #7

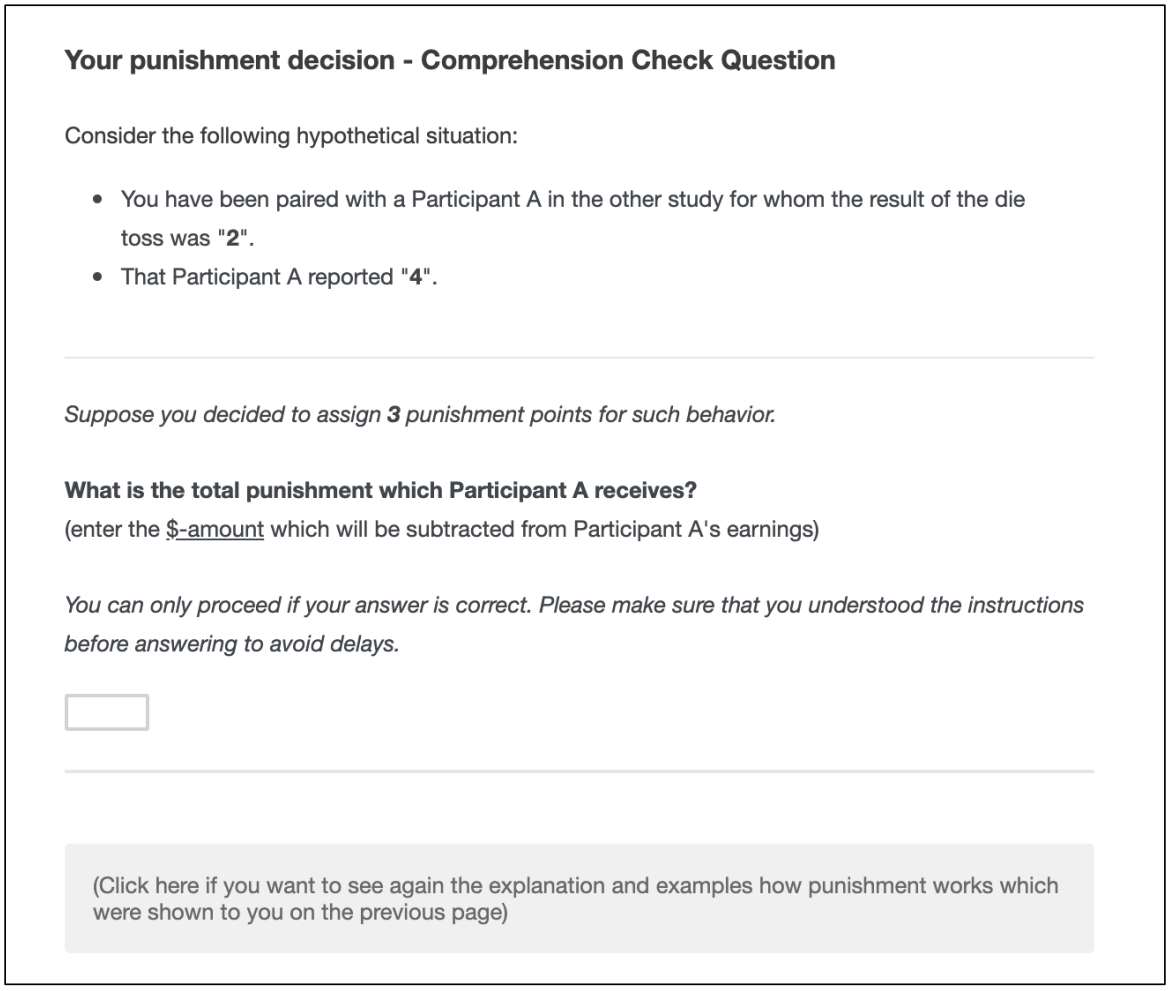


Figure C.2.8: Main experiment, Screen #8a

Your punishment decision

We now ask you to make punishment decisions for 10 different scenarios in which cheating could occur.

To repeat:

- If matched with you, you decide the punishment for Participant A's cheating behavior.
- Each punishment point will lead to a loss of $0.3 for the punished participant in the role of Participant A.

please click if you have read the above

Please note the following:
In a previous study, we asked people what they consider to be the right thing to do for participants in the role of Player A. Those people could not play the game themselves.

The majority of these people stated that the right thing to do for Player A is to report the number truthfully (i.e., report exactly what the die showed).

please click if you have read the above

Please click below how many punishment points (0-5) you want to assign for cheating behavior in each of 10 possible scenarios (you have to enter a value for each scenario).

I want to assign the following number of punishment points if Participant A...

Figure C.2.9: Main experiment, Screen #8b

Figure C.2.10: Main experiment, Screen #9

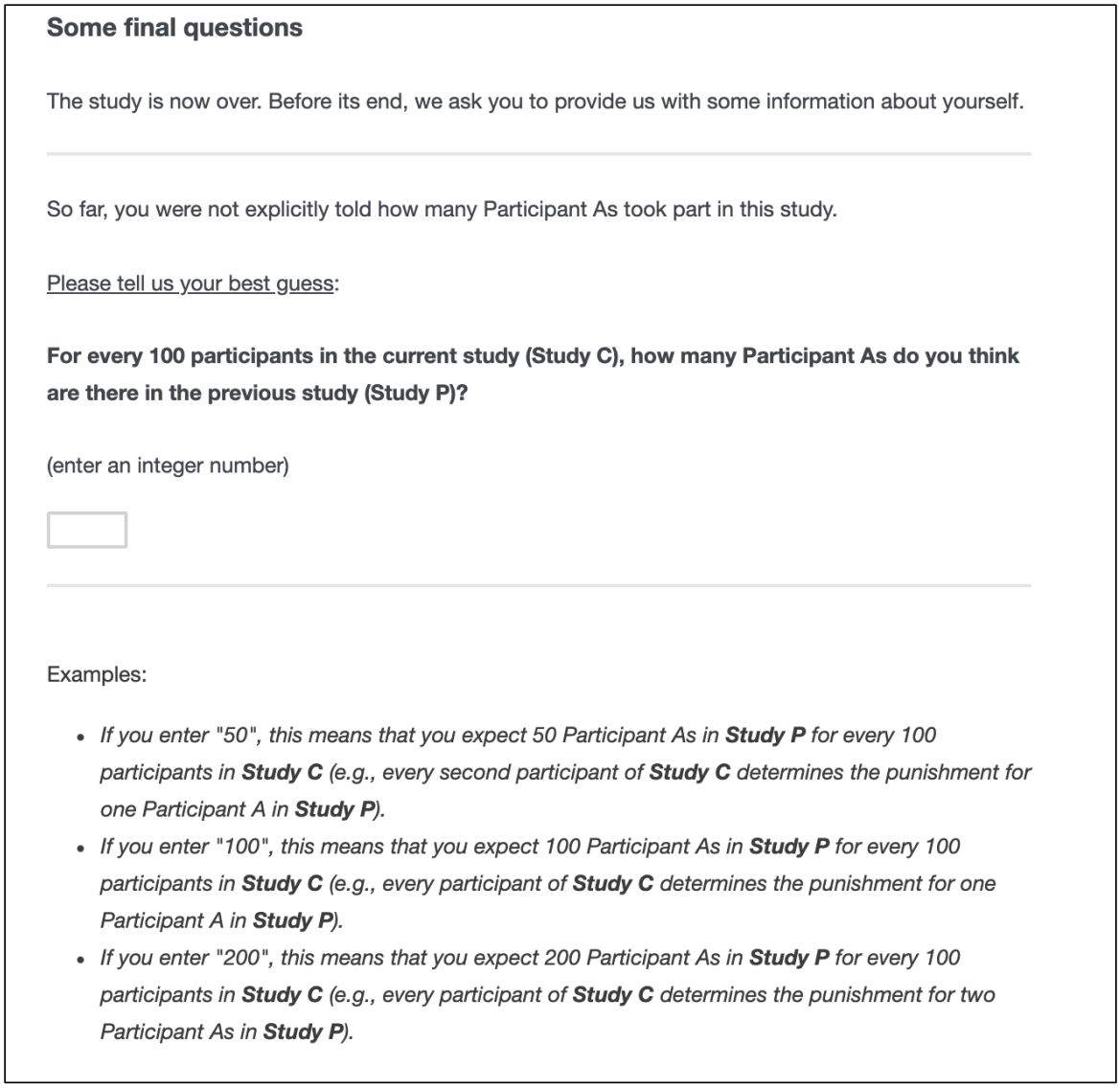


Figure C.2.11: Main experiment, Screen #10

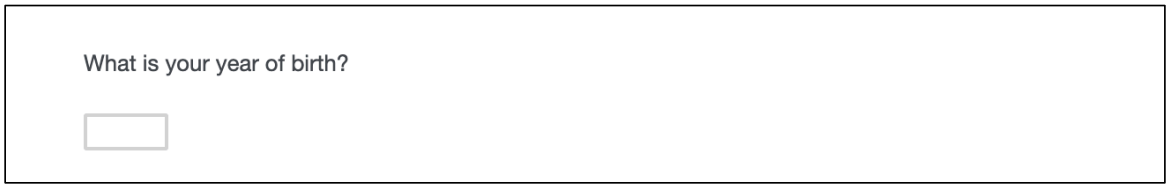


Figure C.2.12: Main experiment, Screen #11

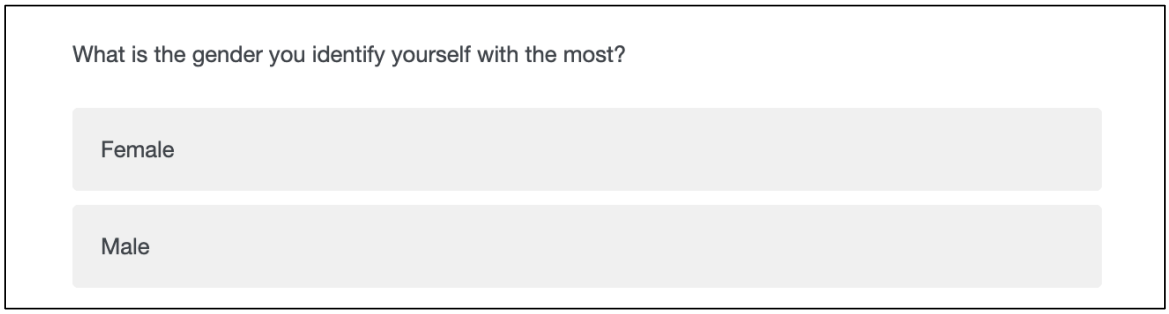
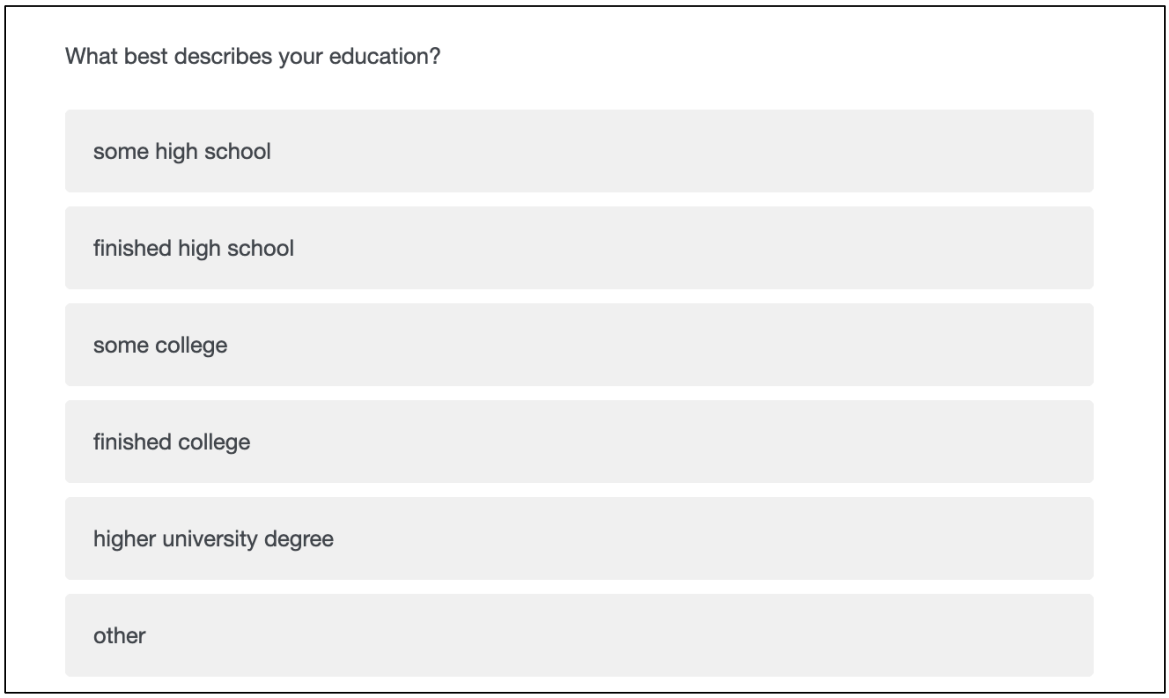


Figure C.2.13: Main experiment, Screen #12



C.3: Instructions for the main experiment of Experiment 2

The following pages display the screens for Experiment 2. All subjects were in the same role (determining the social appropriateness of lying) but the provided norm information and lying scenario differed across subjects. In each treatment, they were presented with either NO INFO, NORMATIVE, or EMPIRICAL information and the equity scenario was either Equity- (p13), Inequity- (p35) or Overclaiming-based (p24) lying. Here, the shown screen present the text for the EMPIRICAL norm information (see, e.g., Figure C.3.5); the content for the other treatments was adjusted accordingly.[23] The same applies for the Equity-scenario. Here, we present Inequity-based lying (see Figure C.3.7; p35: "*Player A rolls a 3 and reports a 5*"). The scenario description was adjusted accordingly for the other conditions.

[23] For NORMATIVE, the text was: *"In a previous study, we asked people what they consider to be the right thing to do for participants in the role of Player A. Those people could not play the game themselves. The majority of these people stated that the right thing to do for Player A is to report the number truthfully (i.e., report exactly what the die showed)."* In NO INFO, such information was not displayed.

Figure C.3.1: Experiment 2, Screen #1a

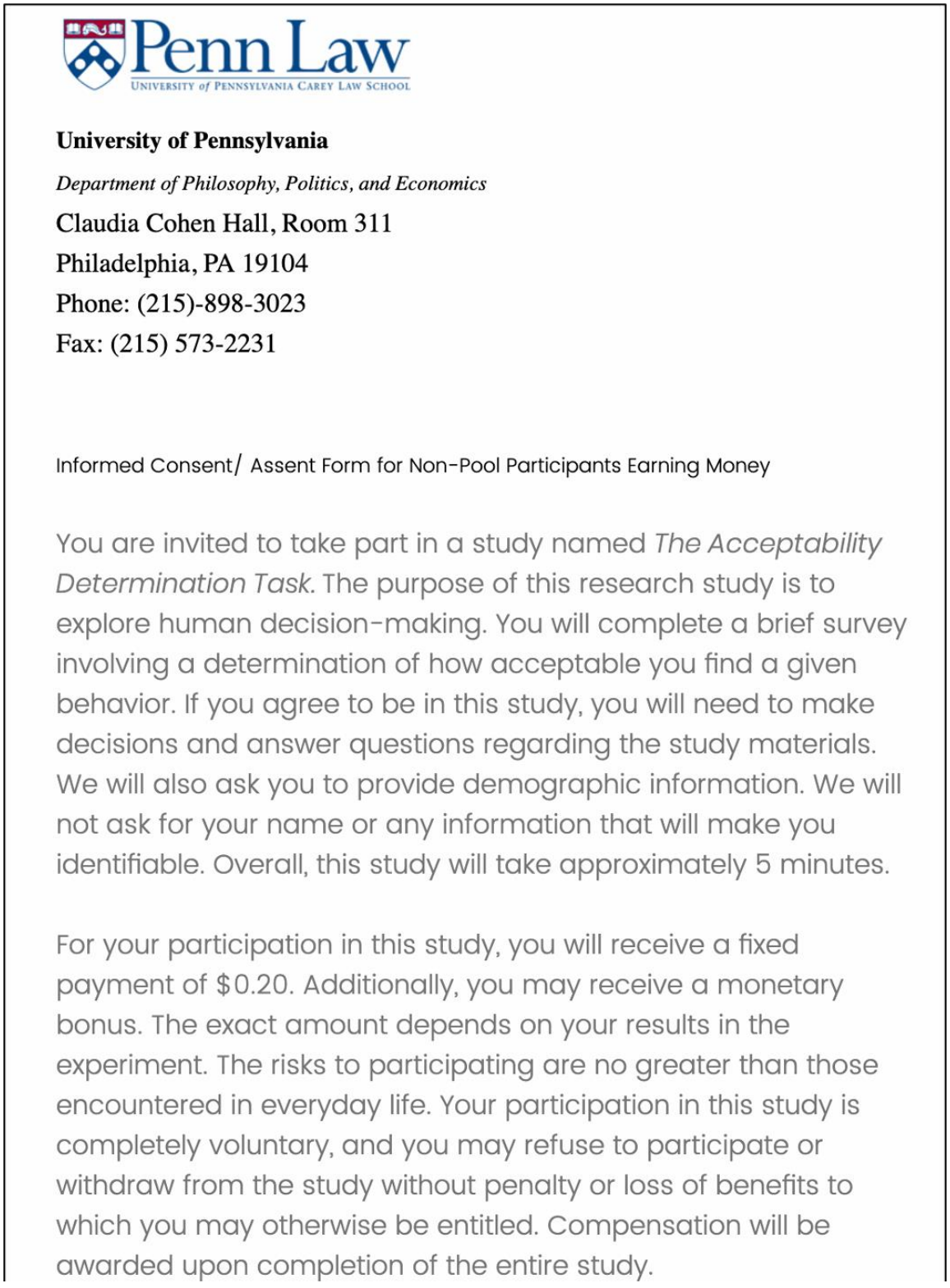


Figure C.3.2: Experiment 2, Screen #1b

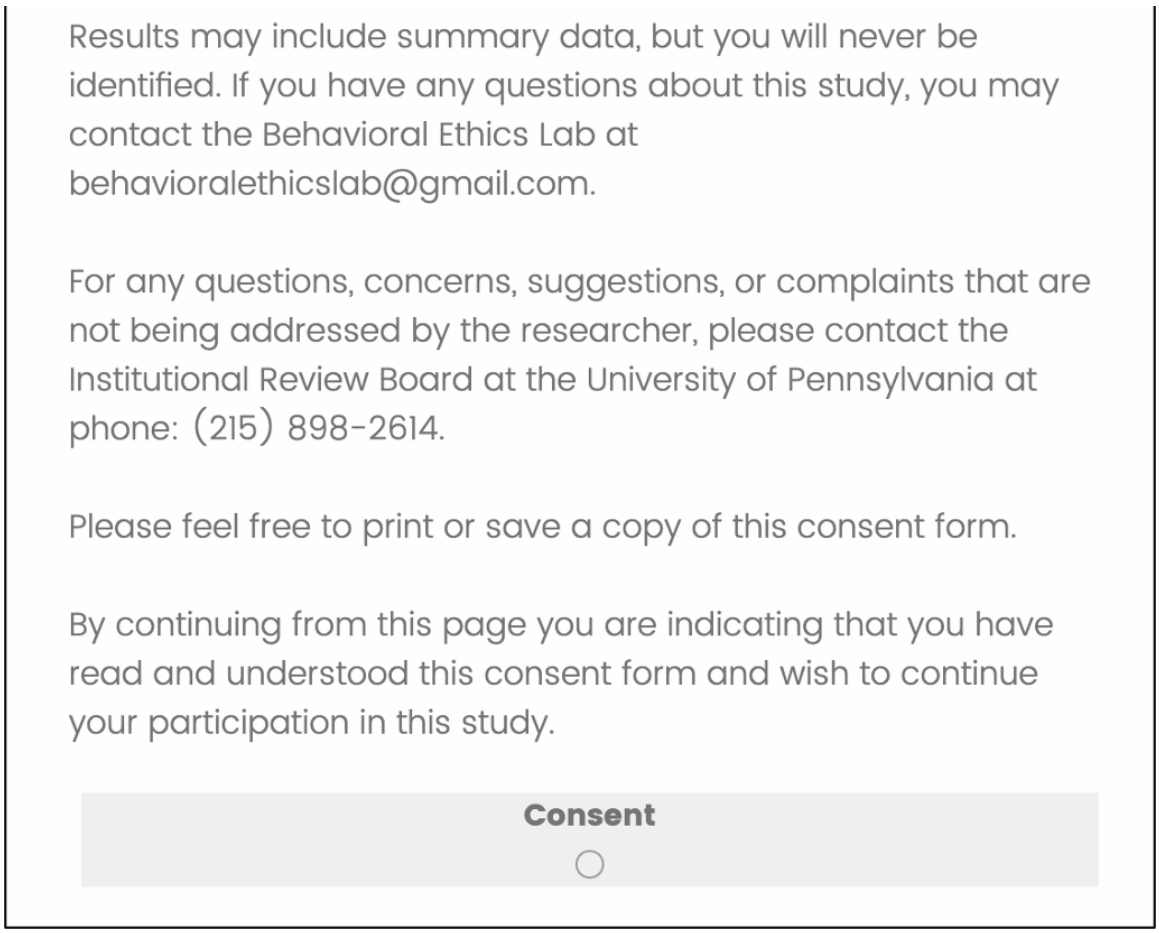


Figure C.3.3: Experiment 2, Screen #2

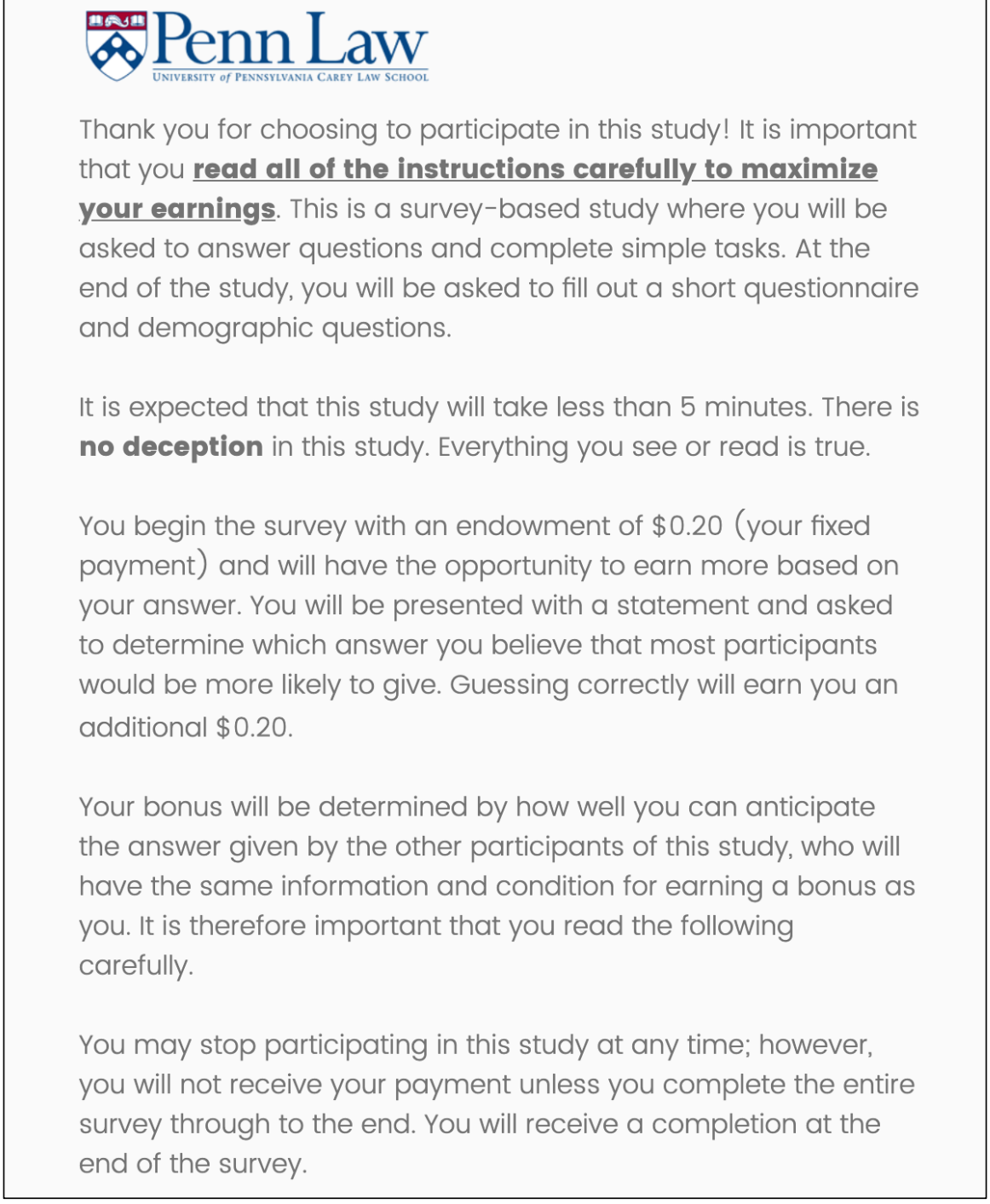


Figure C.3.4: Experiment 2, Screen #3a

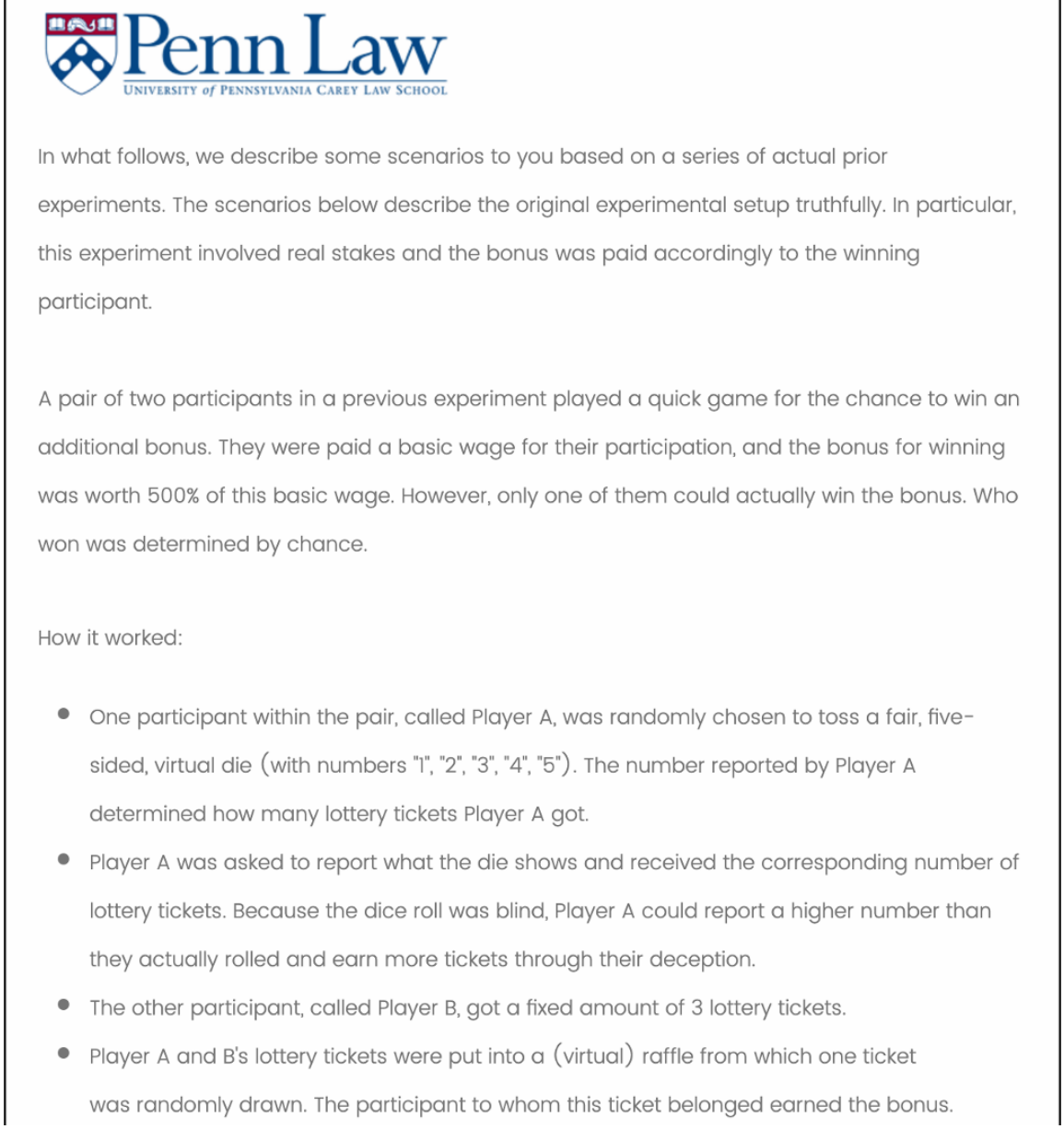


Figure C.3.5: Experiment 2, Screen #3b

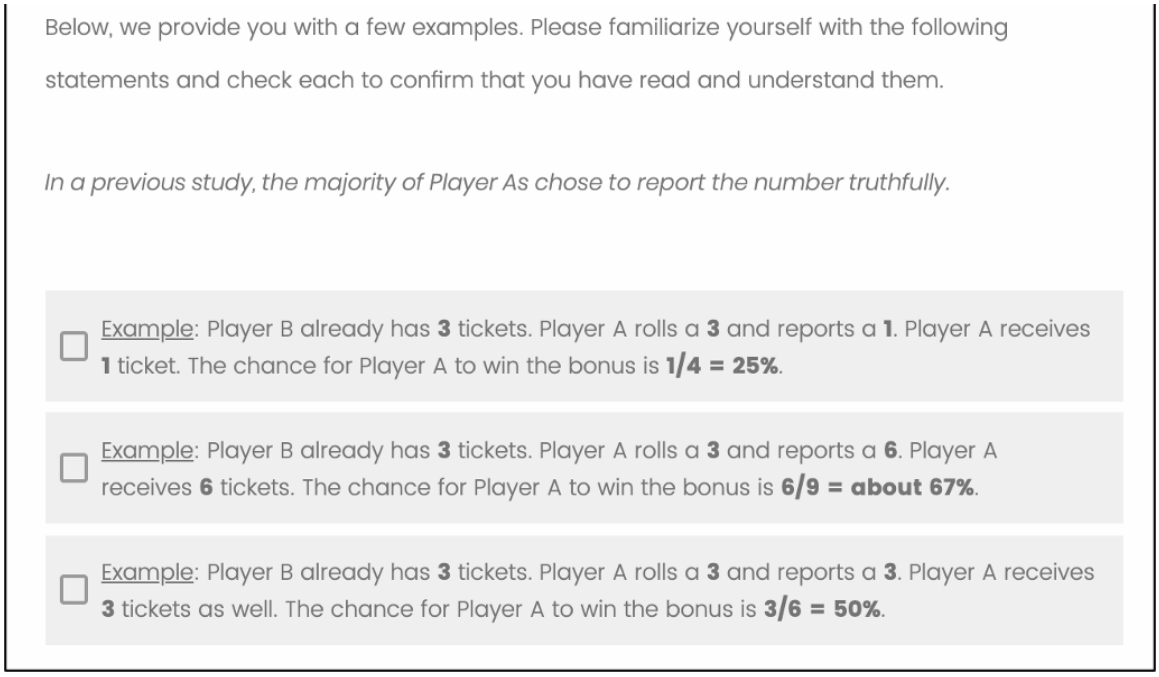


Figure C.3.6: Experiment 2, Screen #4

Figure C.3.7: Experiment 2, Screen #5

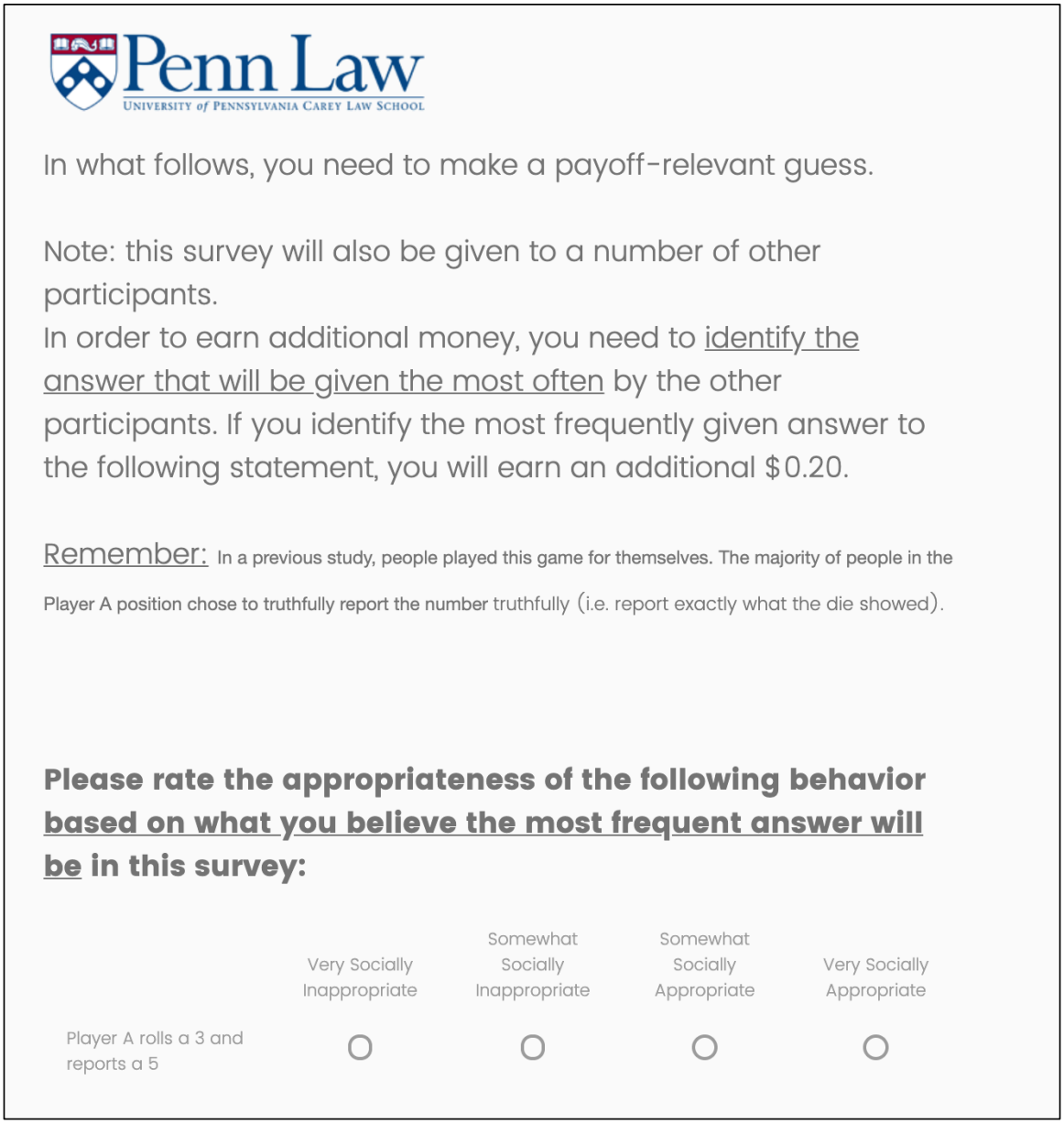


Figure C.3.8: Experiment 2, Screen #6