RATIONALITY & COMPETITION
CRC TRR 190

Do Women Comply More Than Men? Experimental Evidence from a General Population Sample

Müge Süer (The Halle Institute for Economic Research (IWH))
Nicola Cerutti (Oasis Loss Modelling Framework Ltd.)
Jana Friedrichsen (Kiel University)
Gyula Seres (National University of Singapore, N.1 Institute for Health and Institute for Digital Medicine)

Discussion Paper No. 519

December 20, 2024

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Georg Weizsäcker, Ph.D., Humboldt University Berlin, 10117 Berlin, Germany
info@rationality-and-competition.de

Do Women Comply More Than Men? Experimental Evidence from a General Population Sample *

Müge Süer[†] Nicola Cerutti[‡] Jana Friedrichsen[§] Gyula Seres[¶]

December 20, 2024

Abstract

Women are often perceived as more compliant than men; however, the literature provides inconclusive evidence. Using a novel experimental design comprising two complementary experiments, we test this claim in online samples representative of the German adult population. The first experiment (N=1600) features a probabilistic social dilemma game (PDG) in which participants can increase their individual payoff at the expense of exposing themselves and their group to probabilistic losses. In two treatment conditions, they receive either a recommendation on socially optimal behavior or a recommendation and information on weakly non-compliant peer behavior. We find that the recommendation strongly affects behavior but more so for women than for men. However, information on the non-compliant behavior of others does not induce significantly different responses in men and women. In the second experiment (N=522), we elicit empirical and normative expectations about behavior in the PDG with a recommendation to study the role of norms in following it. While men and women are expected to hold similar normative beliefs, men are expected to follow the recommendation less often, suggesting that compliance is a female social norm.

JEL-codes: J16, I12, D81, H41
Keywords: compliance, public good, social dilemma, gender, risk-taking, social norms

*We are grateful to Dirk Engelmann, Klaus Gründler, Dorothea Kübler, Levent Neyse, and Chen Sun for their valuable feedback as well as to conference and seminar participants at the annual meeting of the social science chapter of the VfS 2023 (Nikolskoe), MCC Berlin, the CRC TRR 190 meeting 2022 (Schwanenwerder), the BCCP forum 2022, the IIPF 2022 (Innsbruck), the Annual conference of the Society for the Advancement of Behavioral Economics 2022 (Lake Tahoe), and the International Conference on Social Dilemmas (Copenhagen) for their helpful comments. We thank Anna Balleyer for her contributions during the early stages of the project and specifically for programming the surveys. We gratefully acknowledge financial support from the Dr. Hans Riegel Foundation (grant number AH-047-2009) and from the German Science Foundation (DFG) through CRC TRR 190 (project number 280092119). This study is preregistered as Süer, M., Balleyer, A.H., Cerutti, N., Friedrichsen, J. & Seres, G. (2021, March 26) "Gender and Compliance" in the OSF registry, `https://doi.org/10.17605/OSF.IO/4QKZD` and received IRB approval from the German Association for Experimental Economic Research `https://gfew.de/ethik/TN4SUgad`.

[†]The Halle Institute for Economic Research (IWH). Corresponding author. Email: muege.sueer@iwh-halle.de

[‡]Oasis Loss Modelling Framework Ltd.

[§]Kiel University, Institute of Economics, and CESifo.

[¶]National University of Singapore, N.1 Institute for Health and Institute for Digital Medicine

1 Introduction

Social dilemma situations are ubiquitous in modern societies and their management is a frequent challenge for policymakers. A typical social dilemma features an individually beneficial action that is in conflict with the maximization of social welfare. Thus, public policies often aim to restrict potentially harmful individual behavior below its individually optimal level because this improves social welfare. Policy interventions are particularly relevant in cases where harmful effects are uncertain or delayed, making it difficult for individuals to behave optimally. Examples of such situations are the excessive use of common-pool resources, emission-intensive lifestyles, smoking, or causing fire hazards. Often, individual behavior in these situations is governed by norms, which may sustain outcomes that improve upon the non-cooperative outcome where each individual behaves selfishly (Akfirat et al., 2023; Yamin et al., 2019; Thøgersen, 2014). Interventions can leverage such norms for the social good, which is particularly attractive for problems where formal regulation is difficult or infeasible.[1] For the optimal design of cost-effective and equitable policies, one would want to understand who responds to a normative intervention and who bears its cost. In this paper, we, therefore, complement the literature with a study on gender heterogeneity in the responses to a normative recommendation and in the reaction to social information in an abstract social dilemma situation where harm occurs probabilistically.

Previous research suggests that women are more sensitive to social and environmental cues of a decision situation (Croson and Gneezy, 2009), which could explain why they are found to be more pro-social than men in several studies (Kamas et al., 2008; Soutschek et al., 2017) but not so in others.[2] Consequently, one might expect women to react more strongly to normative cues or recommendations and contribute more to social welfare than men. In line with this argument, Galasso et al. (2020) found that women were more likely to adhere to precautionary measures during the COVID-19 pandemic. More generally, however, the evidence on gender differences in norm-conforming behavior using observational data is mixed. The results vary across the different contexts in which potential gender differences in normative behaviors or conformity with rules and recommendations have been studied, ranging from (in chronological order) sports competitions (Silva, 1983), education (Tibbetts, 1999), tax reporting (Alm et al., 2009; Dulleck et al., 2016; D'attoma et al., 2020), traffic rules (Tom and Granié, 2011), workplace safety (Ucho and Gbande, 2012), public libraries (Apesteguia et al., 2013), peer-to-peer lending (Du et al., 2020),

[1] Norm-nudges, which provide information on what most people do or find appropriate to do (Bicchieri and Dimant, 2022), are tested for instance in an abstract context (Chaudhuri et al., 2006), in the context of energy (Bonan et al., 2020) and water conservation (Brent et al., 2017), in health (Stok et al., 2014; Gelfand et al., 2022), and sustainability (Chen et al., 2009; Loschelder et al., 2019).

[2] The evidence on gender differences in pro-sociality is mixed. The higher sensitivity of women to social and environmental cues offers a potential explanation for these contradictory findings. See Croson and Gneezy (2009) and Sent and van Staveren (2019) for detailed discussion and literature reviews.

and health (Galasso et al., 2020, 2021; Müller and Rau, 2021). Part of the explanation for the inconclusive evidence may be that women are more responsive to social information (Croson and Gneezy, 2009), which may vary across situations and moderate the effect of rules or normative interventions.

In light of the conflicting evidence, we contribute to a better understanding of the role of gender in the reaction to a normative recommendation with evidence from two incentivized online experiments on representative samples of the German adult population. Experiment 1 employs a novel game, the *Probabilistic Dilemma Game (PDG)*, that captures crucial aspects of social dilemma situations with probabilistic welfare losses. In the PDG, individuals, forming groups of 100, decide to what extent they want to engage in a privately beneficial action that carries the risk of causing harm not only to themselves but also to the 99 other participants of their group. Treatments vary depending on whether participants are given a recommendation about behavior that would be optimal for their group and whether they are informed about the behavior of past participants who faced the recommendation but did not follow it. Experiment 2 tests whether behaving in line with the recommendation that favors the utilitarian optimal action in the PDG is an injunctive or a descriptive norm and whether the norms differ by gender.

Experiment 1 consists of three treatment conditions in which participants play the PDG either without a recommendation (Baseline), with a normative recommendation (Rec), or with this recommendation and information about non-compliant behavior of others (Rec&Info).[3] In all treatments, participants play the PDG in fixed groups over ten consecutive rounds. While the first round is identical across treatments, participants in the Rec and the Rec&Info treatments, see the recommendation before every decision from round two to ten. In the Rec&Info treatment, the recommendation is complemented with information about the behavior of four participants who took part in the Rec treatment and deviated from the recommended decision in the majority of rounds. This set of treatment conditions allows us to test the supposedly positive influence of the recommendation (by comparing Rec to Baseline) and the expected erosive effect of observing non-compliant behavior (by comparing Rec&Info to Rec) as well as potential gender difference in the two effects.

In Experiment 2, we use two prominent norm elicitation designs by Bicchieri and Chavez (2010) and Krupka and Weber (2013) and adjust them to analyze subgroup-specific norms. Specifically, we collect information in four treatment conditions using a between-subjects design. Participants in all four conditions are provided with an exact description of the decision situation in the Rec treatment of Experiment 1. Based on this, we ask participants how they believe one should decide in that situation in a treatment

[3]The recommendation informs participants that to maximize their group's total payoff, we recommend a particular decision in the PDG, which is the one that maximizes utilitarian social welfare if taken by every group member.

called Personal Beliefs, thereby gathering data to incentivize answers in the following two treatments. Using the elicitation method by Bicchieri and Chavez (2010), we ask participants to guess which decision the majority of respondents in Personal Beliefs of Experiment 2 stated one should take (treatment Normative Expectations), and which decision they think participants actually took in treatment Rec of Experiment 1 (Empirical Expectations). Finally, we elicit the perceived social appropriateness of different decisions in the Rec treatment of Experiment 1 using the elicitation method by Krupka and Weber (2013). To investigate gender differences in the relevant norms, we elicit normative and empirical observations twice, once with respect to the full sample and then again concerning their expectations only about women. By adjusting the norm elicitation method to the subgroup of females, we can not only test whether participants view compliance with the recommendation as an injunctive norm for society at large but also whether it differs by gender and whether they expect women to be more likely to follow the recommendation than men.

The main insights from Experiment 1 are twofold. First, we find that participants of Experiment 1 clearly follow the recommendation. While behavior is initially remarkably close to the non-cooperative selfish solution of the problem in the Baseline treatment, behavior shifts downward toward the utilitarian optimal choice when this is given as a recommendation in Rec. This effect persists if we take into account that individuals learn to behave more pro-socially over time even in the Baseline treatment leading to a clear downward trend over time. When participants in the Rec&Info treatment are additionally informed that others did not (always) follow the recommendation, we observe an erosion: participants' behavior shifts away from the socially optimal action upward toward the non-cooperative choice, thereby reducing the effectiveness of the recommendation. Second, we find that female participants react more strongly to the recommendation. Starting from only about 16 percent of participants whose behavior in round 1 would be consistent with the recommendation, the rate of those adhering to the recommendation increases by 44 percentage points for men and by 56 percentage points for women when the recommendation is introduced. This result suggests that women are indeed more compliant than men in that they react more often to the recommendation. The picture is less clear for the effect of non-compliant behavior. Comparing behavior in the Rec&Info treatment with that from Rec, we do not find a statistically significant difference in the way men and women react to observing non-compliant choices.

Experiment 2 yields two complementary findings. First, following the recommendation constitutes an injunctive norm according to the method by Bicchieri and Chavez (2010): the majority of participants believe that deciding in accordance with the recommendation is what others think one should do in the Rec treatment. In line with this finding, behavior that conforms with the recommendation is considered socially appropriate according to the elicitation method of Krupka and Weber (2013). Second, while

normative expectations regarding men and women are similar, empirical expectations differ significantly. Specifically, participants expect women to follow the recommendation more than they expect men to follow it but their guess of what a female participant thinks one should do does not significantly differ from their guess about men's beliefs. These results suggest that the recommendation establishes an injunctive norm for both genders but it only constitutes a descriptive norm for women, while men are expected to partially deviate from the recommendation.

Our research makes three contributions to the literature. First, we show that a normative intervention in the from of a non-binding recommendation is highly effective, but the effect differs systematically by gender and the observed differences relate to gender-specific norms that individuals perceive in the relevant decision situation. Specifically, the descriptive norm to conform with the recommendation is stronger for women than for men while the injunctive norm to do so exists for both in the same way. In light of the results from (Bicchieri and Xiao, 2009) that the descriptive norm dominates when in conflict with the injunctive norm, this may explain our finding that women are more likely to follow the recommendation. Second, we document that observing non-compliant behavior in others partially erodes the effect of the recommendation. Thus in addition to potential beneficial effects from observing compliant peers (see e.g., Ehrlich and Szech, 2022), transparency about the behavior of others may also weaken compliance if deviant behavior is already prevalent. Third, we make two methodological contributions. For one, we propose a new game, the Probabilistic Dilemma Game (PDG), as an online implementable tool to examine behavior in social dilemmas with probabilistic harm. The PDG is easy to implement and affords a lot of flexibility to be tailored to various follow-up research questions. In addition, we show that the norm elicitation method introduced by Bicchieri and Chavez (2010) can be fruitfully applied to study subgroup-specific norms, in our case norms that vary with gender. This approach can be easily utilized to explore further social norms and how these may vary across subgroups in a society.

Our results are also informative for policy design. The understanding that men and women differ in their response to normative recommendations at least partly because the corresponding norms differ by gender, may help improve the design of interventions intended to improve compliance with rules and recommendations. Our results suggest that the differential effect of the recommendation between men and women is not rooted in different perceptions of what would be appropriate but in the self-fulfilling expectation that men are less likely to follow the recommendation. Thus, interventions might profit from finding leading examples or policing deviant behavior in men to strengthen the male norm in order to equally distribute the burden from compliance between men and women. Health and safety are potential fields of applications where certain behaviors may be perceived as (myopically) attractive by the individual but carry the risk of leading to high social costs, as is the case with tobacco and alcohol consumption (Bouchery et al.,

2011), unhealthy diets (Candari et al., 2017), unprotected sexual intercourse (Schnitzler et al., 2021; Bahk et al., 2015), drunk driving (Sloan et al., 2014), workplace safety (DeJoy, 2005; Thedell, 2016), or wildfires (Howe et al., 2018).

The rest of the paper is structured as follows. Section 2 explains the details of the probabilistic dilemma game, the theoretical background of the decision setting, and our hypotheses. Section 3 describes the experimental design of the two online experiments and their implementation. Section 4 introduces the results of the two experiments in sequence, which we discuss in more detail in Section 6 and conclude.

2 Theoretical Background and Hypotheses

We are interested in the possibly gendered effect of an unenforceable recommendation and its interaction with information about the non-compliant behavior of others in a social dilemma situation with probabilistic harm. To address our research question, we designed a simple computerized ball-drawing game, hereafter referred to as PDG for *Probabilistic Dilemma Game*, that captures the key features common to the examples mentioned above. In the PDG, individuals belong to groups, and their decisions to engage in a privately beneficial activity, drawing balls from an urn, carry a risk of harm both to the individual and to the group as a whole. In this game, we assume that the marginal expected loss imposed on the group from an individual drawing a ball is always larger than the marginal expected private loss from that ball, which a rational purely selfish individual aims to equate to her marginal benefit. Thus, the individual will draw too many balls, and choose too much of the privately beneficial activity, as compared to the number that would maximize the group's welfare.

In the PDG, which is the core of Experiment 1, a participant repeatedly encounters a virtual urn containing 100 balls, 95 of which are white and 5 of which are black. Participants are matched in groups of 100, where everyone individually and independently decides how many balls they want to draw with replacement. They know that each white ball that they draw yields a private benefit of 0.10 Euro. However, drawing at least one black ball destroys all the benefits of the individually collected white balls, and additionally reduces the value of the group account by 1 Euro per black ball. The group account starts with a value of 100 Euros at the beginning of each round. What is left at the end of the round is divided equally among the 100 people in the group. For example, if a person draws 3 black balls in a round, they lose all of their private earnings from their white balls and also cause a 3 Euro loss to the group account, thus lowering everyone's group earnings by 0.03 Euro in that particular round.

2.1 Theoretical background

To derive the optimal decision of a participant, let us consider a more general mathematical version of the problem. Every round, the participants face an urn containing n balls. Before any ball is drawn, the group account has a value of $\theta > 0$ per person, hence θN for N individuals in a group. A randomly drawn ball is black with probability $(1-p)$ and white with probability p where $0 < p < 1$. Each white ball that an individual draws provides a *private benefit* of $\gamma > 0$. However, an individual who draws at least one black ball loses everything (*private loss*). Additionally, each black ball drawn reduces the group payoff by θ (*group loss*). White balls have no effect on the group account.

Self-interested optimum Suppose first that an individual ignores their own and others' influences on the group payoff through the total number of balls drawn by the group completely and only cares about the payoff from their individual urn. In this case, the individual payoff from the group account is perceived as a constant and, thus, omitted in the individual optimization problem. When deciding about the integer number of balls k to draw, the individual implicitly decides which of the following $k+1$ lotteries to choose:

$$L(k) = \begin{cases} \gamma k & \text{with probability } p^k \\ 0 & \text{with probability } 1-p^k \end{cases} \tag{1}$$

The expected value of lottery $L(k)$ is $\mathbb{E}[L(k)] = p^k \gamma k$, and its value is maximized at $k^* = -\frac{1}{\log(p)}$. The private benefit of a white ball, γ, does not affect the optimal choice of balls for an expected-payoff maximizing individual. For the parameter values of the experiment, the number of balls that maximizes an individual's expected payoff from the private urn is $k^* \in \{19, 20\}$.[4] Risk-averse individuals would choose a k lower than the risk-neutral optimum, and risk-seeking individuals would choose a larger k.

We consider several variations of the self-interested optimum. First, we show that an individual who is self-interested but anticipates the effect on their utility through the group urn continues to draw 19 balls. Additionally, we show that social concerns lower the number of balls drawn. Specifically, we show that individuals, who have Kantian preferences (Roemer, 2015; Alger and Weibull, 2013), i.e. they choose an action that maximizes the group outcome when chosen by everyone, or who experience a cold shiver from drawing balls, i.e. they experience a disutility from an action that may harm the group (Bruvoll and Nyborg, 2004; Brekke et al., 2010), will draw fewer than 19 balls. Details can be found in Appendix A.

[4]The solution to the expected utility maximization problem is $k = 19.49$ but, respecting the integer constraint, 19 and 20 balls yield a utility level that is identical up to the fifth digit.

Social planner optimum To maximize the utilitarian social welfare of the group, we need to take into account the effect of an individual's behavior on the payoff of all N group members. Assuming symmetry, the utilitarian welfare function to be maximized is

$$W(k) = N[\gamma p^k k + \theta - \frac{\theta}{N}(1-p)Nk], \tag{2}$$

where $N\theta$ is the total payoff from the group account when no harm is incurred and $N\theta(1-p)k$ is the expected harm. This expression assumes its maximum at $k = 6$ for the experimental parameter ($p = 0.95, \gamma = 0.1, \theta = 1, N = 100$). This result is the basis for the REC treatment, where we tell individuals that, to maximize the expected total group payoff (=utilitarian welfare), they should draw no more than 6 balls.

Optimum with perceived norms Various studies support the idea that individuals adjust their behavior to perceived norms, not only but very prominently in the context of common-pool resource and social dilemma situations (see, e.g., Cardenas, 2011; Capraro and Rand, 2018; Farrow et al., 2017; Nyborg et al., 2016). Such norms could be *injunctive* (what one should do) or *descriptive* (what others do). To include these two aspects in the utility function, denote by k_P the perceived norm, which can be purely injunctive ($k_P = k_I$), or descriptive ($k_P = k_D$), or a weighted average of the two ($k_P = \alpha k_I + (1-\alpha)k_D$) for some $\alpha \in (0,1)$[5]). We assume that individuals concerned with adherence to the norm experience a disutility when they exceed the perceived norm of ball draws; we do not allow for positive utility from over-compliance, i.e. from drawing fewer balls than perceived as the norm, but utility does not decrease from over-compliance either.[6] The utility is then given by

$$U(k) = \gamma p^k k + (\theta - \frac{\theta}{N}(1-p)(k + k_{-i})) - \beta[k - k_P]^+, \tag{3}$$

where $\beta > 0$ is the marginal disutility from exceeding the perceived norm k_P and k_{-i} is the total number of balls drawn by other group members. Thus, we expect that individuals with $\beta > 0$ will adjust their behavior toward the perceived norm and they will do so more the larger their β (cf. Kimbrough and Vostroknutov, 2016, for the idea that varying levels of pro-social behavior may result from heterogeneous sensitivity to norms).

The perceived norm k_P may be considered exogenous in a static environment. However, in our setting, the treatments are expected to affect the perceived norm and, thereby, behavior. Specifically, suppose that individuals who play the PDG without additional in-

[5]The parameter α may vary between individuals: Some may experience a strong norm of how one ought to behave (high α) whereas others may care more about what others are doing (low α)

[6]A similar formulation of norm-based utility is used in Herweg and Schmidt (2022). Alternative views in the literature posit that individuals experience disutility from over-compliance as well as from non-compliance, see e.g. Brekke et al. (2003).

formation, have some prior about the number of balls one ought to draw in this game $\tilde{k}_I$, some prior about the number of balls that others will draw in this game $\tilde{k}_D$, which is their perceived descriptive norm that we assume for simplicity to coincide with an individual's own behavior.[7] Previous research further suggests that the injunctive norm $\tilde{k}_I$ will be inferred from the descriptive norms in the absence of any further information (Eriksson et al., 2015) so that the prior perceived norm will equal the individual's draws, $k^P = k$. When given the recommendation to draw not more than six balls to maximize the expected group payoff, the prior $\tilde{k}_I$ is updated to six, which has two effects. First, this update decreases an individual's perceived norm k_P for any given prior of the descriptive norm as long as the prior $\tilde{k_I}$ exceeds six. It is easy to see that the introduction of a recommendation of six, therefore, leads norm-compliant individuals who would otherwise draw more than six balls to reduce the number of balls drawn to avoid the disutility from norm violation. Second, this decrease is reinforced by an induced change in the descriptive norm, which will adjust downward in line with the injunctive norm, in the absence of any additional information on what others do (Eriksson et al., 2015). The total effect on behavior will be stronger the larger β is.[8] Conversely, when individuals observe the behavior of others, this does not affect the injunctive norm but the descriptive norm. Specifically, the observation that others do not follow the recommendation leads individuals to update their expectations of the descriptive norm in the direction of the observed average behavior. Based on the findings of Bicchieri and Xiao (2009), we hypothesize that observing non-compliant behavior weakens the effect of the recommendation by leading to a direct increase in the perceived descriptive norm, which counteracts the indirect decrease implied through the recommendation. Thus, the perceived norm weakens and we expect a smaller decrease in balls drawn when both a recommendation and violations thereof are observed than with the recommendation alone.[9]

2.2 Hypotheses

We use the theoretical framework of the PDG to derive hypotheses about the main effects of our treatment variations. Using results from the literature, we specify additional hy-

[7] This assumption is consistent with experimental results on the false consensus effect, which refers to the observation that subjects often expect others to behave in the same ways as themselves (Ross et al., 1977). See also Engelmann and Strobel (2000) as an early experimental economics study on the false consensus effect and Blanco et al. (2014) on the relevance of false consensus in explaining behavior in social dilemmas.

[8] An alternative conceptualization of the effect of the recommendation is that it focuses attention on the norms relevant in the particular situation so that these become behaviorally relevant even if the treatment does not shift the norms. See Krupka and Weber (2009) for empirical evidence for such a focusing effect of making people think about their own or others' behavior.

[9] Observing behavior that violates the norm may alternatively reduce the importance attached to non-compliance, i.e., lead to a decrease in β, which is assumed to be unaffected by the behavior of others in our model.

potheses related to expected gender differences in the effects of the treatments. We discuss in the text where we deviate from the wording of our hypotheses in the preregistration.

The first hypothesis follows from the decision problem of an individual who cares about following a perceived injunctive norm. It was not preregistered. While there is no salient norm or behavioral guidance in the BASELINE treatment, the recommendation in the REC treatment suggests a specific behavior and makes it salient that the injunctive norm is to draw at most six balls.[10] In treatment REC&INFO, this recommendation is complemented with empirical information on the behavior of others who often violate this injunctive norm, so that the descriptive norm does not align with the recommended behavior but tolerates a larger number of draws. As long as people experience disutility from exceeding the perceived norm, which depends on both the perceived injunctive and perceived descriptive norm, which we expect to hold based on previous studies (e.g. Bicchieri and Xiao, 2009, and references above), our theoretical framework yields the following hypotheses regarding our main treatment effects:

Hypothesis 1.A (Recommendation effect)**.** *The share of individuals drawing at most six balls is larger in* REC *than in* BASELINE*. The number of balls drawn is on average closer to the social optimum in* REC *than in* BASELINE*.*

Hypothesis 1.B (Negative information effect)**.** *The share of individuals drawing at most six balls is smaller in* REC&INFO *than in* REC*. The number of balls drawn is on average further from the social optimum in* REC&INFO *than in* REC*.*

Our next hypothesis is based on the finding that women are more communal, collective, and participatory, while men are more agency-driven, indicating individualistic decision-making and autonomy (Diekman and Goodfriend, 2006; Eagly, 2009). Communal individuals tend to conform to perceived group norms, even if these do not favor the most pro-social action.[11] Thus, the two sub-hypotheses that we test represent different variants of the hypothesis that, on average, women are more likely to adjust their behavior in the direction of perceived norms than men.[12] In the REC treatment, the recommendation to draw no more than six balls establishes an injunctive norm, and in the REC&INFO treatment, the perceived norm is a mixture of this injunctive norm and the less rigorous descriptive norm that results from observing others deviating from the recommended action (empirical information).

[10]Experiment 2 explicitly tests whether and confirms that the recommendation establishes an injunctive norm.

[11]General characteristics attributed to men and women are often misinterpreted as pro-sociality expectations. The social role theory proposed by Eagly (2009) states that being communal suggests being pro-community, but this does not necessarily imply pro-sociality. We take up the question of pro-sociality and compliance in later hypotheses.

[12]Hypothesis 2.A rephrases the preregistered Hypothesis 1 "Women are on average more compliant than men". Hypothesis 2.B rephrases the preregistered Hypothesis 3 "Women's compliance behavior erodes when they learn that others comply less." We decided to change the wording to make it clearer that we are interested in compliance as the change in behavior in response to the recommendation.

Hypothesis 2.A. *The treatment difference between* Rec *and* Baseline *(Recommendation effect), is larger for women than for men.*

This first part of the hypothesis postulates that women comply more than men: we expect women to follow the recommendation more often and reduce their draws to at most six balls than men.

Hypothesis 2.B. *The treatment difference between* Rec&Info *and* Rec *(Negative information effect) is larger for women than for men.*

The second part states that the observation that others deviate from the recommended action changes women's behavior more than that of men. We originally preregistered a hypothesis only for women,[13] but given the non preregistered Hypothesis 1.A on the total effect of observing deviant behavior, this new hypothesis captures the preregistered idea that women react more strongly to observing it. The reasoning behind both parts of the hypothesis is that women are expected to react more strongly to normative cues, which in the theoretical framework corresponds to a greater marginal disutility β from exceeding the perceived norm.

Our theoretical framework above uses an objective probability p but we understand individuals as making their decisions using their perceived risk, i.e. based on a subjective probability $\hat{p}$ that need not coincide with p. Our framework immediately yields the prediction that for $p' > p''$, *ceteris paribus*, an individual with a higher perceived risk $1 - p''$ will draw fewer balls than a more optimistic individual with a lower perceived risk p'. Optimism pushes an individual's selfishly optimal action upward, which makes it more attractive to draw additional balls, whereas higher perceived risk will reduce the individual's desire to draw balls. Therefore, we postulate that:

Hypothesis 3. *Individuals who perceive the decision situation as riskier, draw fewer balls and are more likely to draw at most six balls.*

Hypothesis 3 replaces the preregistered Hypothesis 2 "Compliance increases with the increased risk perception". During the analysis, it became clear that "Compliance" must mean something different here than in the preceding hypotheses because compliance is not policed therefore deviating from the recommendation is not riskier than the same behavior in the absence of the recommendation. Further, an individual's risk perception may not only influence individual decisions but may also change over time depending on their number of draws and previous experience. We therefore rephrased the hypothesis to clarify that here we are interested in the relationship between the subjective riskiness of the decision situation and the extent to which individual behavior conforms with the socially optimal number of draws.

[13] Specifically, we preregistered "Women's compliance behavior erodes when they learn that others comply less." as Hypothesis 3.

Our theoretical framework further allows us to make predictions about how behavior in the PDG will adjust to changes in individual risk aversion, social preferences, and beliefs about others, Hypotheses 4.A, 4.B, and 4.C, respectively. These hypotheses were not preregistered. Participants in Experiment 1 report their own risk preferences and preferences for personal and group profit maximization. They also report their beliefs about others' preferences for personal and group profit maximization. We expect the effect of risk aversion on compliance to be positive, of a selfish profit preference to be negative, and of a pro-group profit preference to be positive. Additionally, we expect the effect of beliefs about others' pro-sociality to be also positive. Essentially, an individual who believes that others are more (less) pro-social than themself might adjust their behavior to be more (less) pro-social and draw fewer (more) balls, *ceteris paribus* as in Krupka and Weber (2009).[14] This leads to the following sub-hypotheses:

Hypothesis 4.A. *The more risk-averse individuals are, the smaller is the number of balls they draw and the more likely they are to draw at most six balls.*

Hypothesis 4.B. *The more individuals care about others, the smaller is the number of balls they draw and the more likely they are to draw at most six balls.*

Hypothesis 4.C. *Individuals who believe others to be pro-social draw fewer balls and are more likely to draw at most six balls than those who believe others to be rather selfish.*

The final hypothesis addresses the question of whether the recommendation establishes a gendered social norm. We follow the norm model proposed by Bicchieri (2016) who define injunctive norms as behaviors that are prescribed by normative expectations, descriptive norms as behaviors that are prescribed by empirical expectations, and social norms as the mutual coincidence of normative and empirical expectations. Individuals are more likely to engage in a certain behavior when they think it is commonly practiced (i.e. it aligns with their empirical expectation), and when they expect others to approve of it (i.e. it aligns with their normative expectation). We hypothesize that the empirical and normative expectations align both for women and men but on different behaviors. Specifically, we expect a female social norm to exist in the sense that women are expected to follow the recommendation and also do so, whereas, for men, we predict deviations from the recommendation to be the social norm. To avoid misunderstandings, we rephrase the preregistered Hypothesis 4 "Compliance is a female social norm." as follows:

Hypothesis 5. *Following the recommendation and drawing at most six balls in* Rec *is a social norm that applies to females but not to males.*

[14]However, it is important to note that an individual may also engage in compensatory behavior, where they feel entitled to behave more selfishly when they expect others to be relatively pro-social, or *vice versa* feel obliged to behave more pro-socially when they expect others to be relatively selfish as observed in Fischbacher and Gächter (2010).

3 Experimental Design and Procedures

This study consists of two online experiments: Experiment 1 uses the *PDG* to study actual decision-making. Experiment 2 elicits different beliefs about behavior in Experiment 1 to study the norms that govern behavior. In this section, we first explain the experimental designs of Experiments 1 and 2 and then summarize the procedural details of their implementation and data collection.

3.1 Experiment 1: PDG

In Experiment 1, we study how individuals behave in the PDG. The experiment employs a between-subjects design with three treatments described below. Each treatment consists of three blocks. The treatment variation is implemented in the second block of Experiment 1. The first and the third blocks are identical in all three treatments. In the first block, participants receive detailed instructions about the PDG es explained for the Baseline treatment. They also play one round of it as a test that is not paid to familiarize themselves with the decision situation. In the second block, participants play ten rounds of the PDG in groups of 100. Between rounds, each participant receives feedback about the number of black balls they have drawn in the previous round but not about the decisions or outcomes of their group. In the third block, participants answer an exit questionnaire, which collects amongst others demographic information, willingness to take risks, preference for own versus group payoff maximization, and beliefs about others' preferences for own versus group payoff maximization. Participants were paid for one randomly drawn round from the second block and could receive a bonus from several incentivized belief questions. Appendix G contains the screenshots of the main decision screens of Experiment 1 as well as a transcription of the on-screen instructions.

In the second block, we implement the following three treatments.

1. Baseline: The participants play ten rounds of the PDG as described in Section 2 without any intervention. To recap, participants are matched in groups of 100 and each group is endowed with a group account worth 100 Euros. In every round, each participant decides how many balls to draw with replacement from an urn containing 95 white and 5 black balls. Each white ball is worth 0.1 Euro but drawing any black ball destroys not only the individual's entire private earnings but also reduces the group account by 1 Euro. The remaining amount in the group account at the end of a round would be equally distributed over all group members if this round is drawn for payment.

2. Rec: In round 1, participants face the same decision as in round 1 of the Baseline treatment. Before making their decision in round 2, participants receive the

following non-binding recommendation: "Please note: To maximize the payment to all participants, we recommend that you draw no more than six balls". The recommendation appears on a separate screen after the feedback page of round 1 and is repeated in the upper left corner of each of the subsequent nine decision screens.

3. Rec&Info: In round 1, participants again face the same decision as in round 1 of the Baseline treatment. In rounds 2 to 10, participants receive a combination of the recommendation used in the Rec treatment and empirical information about the past behavior of participants in the Rec treatment. This information is introduced together with the recommendation on a separate screen after the feedback screen from round 1. Specifically, they see the following message in addition to the recommendation: "In the following rounds, you will also be shown the sum of the balls drawn by 4 participants (players A, B, C, and D) from the same round. These values are based on an earlier, identical experiment." The behavior shown to participants is randomly chosen from a subset of participants of the Rec treatment who deviate from the recommended action, i.e. participants who always drew more than and only occasionally equal to 6 balls.[15] All participants in the Rec&Info treatment see the same data in a given round, which changes from round to round in the same way in which the behavior of the chosen players changes in the real data. See Appendix B for details of the provided information on past behaviors.

3.2 Experiment 2: Norms in the PDG

Experiment 2 aims to better understand whether following the non-binding recommendation in Rec is perceived as an injunctive or descriptive norm. To do so, experiment 2 employs a between-subjects design with four treatments, which all consist of three parts. First, participants in all treatments are provided with a description of the Rec treatment of PDG, i.e. they see the instructions of Experiment 1 without actually taking part in it. Second, they are asked to answer a set of treatment-specific questions about their beliefs and expectations. Third, participants in all treatments fill out an exit survey as in Experiment 1.

The treatments were the following, were we build on Bicchieri and Chavez (2010) to elicit normative and empirical expectations and on Krupka and Weber (2013) for the perceptions of social appropriateness. As the responses in these treatments may depend

[15]We use the behavior of participants who often but not always deviate from the recommendation to expose participants to negative empirical information through observing non-compliance and at the same time mitigate demand effects by letting the observed behavior vary and follow the recommendation occasionally. It is important to note that the participants in Rec&Info were not informed about the exact selection of the data they were shown. This allows us to study whether behavior is negatively affected through the information even though the treated participants do not know whether the observed behavior is common or not.

heavily on the participants' understanding of the instructions, the experiment includes two comprehension questions immediately after the instructions. Any participant who fails to answer both correctly is excluded from the experiment without payment.

1. Personal Beliefs: Participants report in an unincentivized way what they think one should do in the PDG, i.e. their personal beliefs, by answering the following question: "Think about the experiment described above. Please indicate how you think one should behave in a decision round. One should draw the following number of balls: _____". The Personal Beliefs treatment is used to incentivize answers in the Normative Expectations treatment.

2. Normative Expectations: Participants report their expectations of what others think one should do. Specifically, participants make an incentivized guess of what a) the majority of all participants and b) the majority of female participants had answered in the Personal Beliefs treatment by answering the following statements: "Indicate which statement you think is correct: The majority of all [all female] respondents in the survey answered that the following number of balls should be drawn: _____".[16] Each correct answer is rewarded with 0.10 Euro.

3. Empirical Expectations: Participants report their expectations of what others actually do. In this treatment, participants make an incentivized guess of what a) the majority of all participants and b) the majority of all female participants chose in the Rec treatment of the PDG by answering the following statement: "Imagine all the (female) participants in the experiment. What do you think the majority of participants did? They drew the following number of balls: _____". Each correct answer is rewarded with 0.10 Euro.

4. KW Method: Participants make an incentivized guess of other respondents' beliefs about how appropriate a certain behavior in the PDG is. Specifically, participants separately rate the social appropriateness of each possible number of draws between 0 and 31 on a 4-point Likert scale (ranging from "very socially appropriate" to "very socially inappropriate"). Out of the 20 answers, two are randomly drawn and rewarded with 0.10 Euro each if these are correct.

Our norm elicitation setting differs from Bicchieri and Chavez (2010) in eliciting the norms also for females separately. To do so, the incentivized questions of Normative Expectations and Empirical Expectations are asked once for the whole group and once for women to capture the existence of any gendered social norm. A social norm, that is likely to affect behavior, is said to exist for a certain population, when participants'

[16]The reason why the questions were not asked for men and women separately but rather for the whole participants and for the female participants is twofold: i) to stick to the original elicitation for the whole group and determine the general norms, ii) to not cause further demand effect.

normative and empirical expectations coincide for this population (Bicchieri, 2016). Thus, if the recommendation establishes a social norm to draw at most six balls only for women, participants should have the according, coinciding normative and empirical expectations for women, but not for men.

3.3 Procedures

Both Experiment 1 and Experiment 2 inclusive of the accompanying survey were programmed in Qualtrics and were administered to a German adult sample with the help of Respondi as a professional survey provider, targeting representativeness with respect to gender, education level (three levels), age, and state (Bundesland). Details on the collected samples are contained in Appendix D.

Experiment 1: PDG The first wave of data was collected on June 10–25, 2021. In this wave, each participant was randomly assigned to one of two treatment conditions: Baseline ($N = 514$) or Rec ($N = 586$). The second wave of data collection took place on July 14–28, 2021. In this wave, all participants were allocated to the Rec&Info treatment, in which participants received information about the behavior of selected participants from the Rec treatment in the first wave of data collection.[17] The median earning, including a fixed fee of 1.00 Euro plus earnings from one randomly selected round, was 1.52 Euros in Baseline, 2.14 Euros in Rec, and 2.11 Euros in Rec&Info. The median completion time in Experiment 1 was 18 minutes and 6 seconds.

Experiment 2: Norm elicitation experiment Data was collected between October 18 and November 15, 2021, using a separate sample ($N = 522$). Each participant in this experiment was randomly assigned to one of the four treatments, Personal Beliefs ($N = 133$), Normative Expectations ($N = 126$), Empirical Expectations ($N = 136$), and K&W method ($N = 127$). The earnings consisted of a fixed fee of 0.50 Euro and additional earnings from the elicitation questions for the relevant treatments. The median earning was 0.50 Euro with a mean of 0.53 Euro. The median completion time in Experiment 2 was 6 minutes and 17 seconds.

[17] Before the data collection of PDG, we ran two pilot studies with a total of 77 students at Humboldt-Universität zu Berlin to calibrate the objective probability parameter. Pilots differed in the probability of drawing a black ball, low risk using 5% as in the actual experiment, and high risk using 15%. The participant behavior did not differ between low-risk 5% and high-risk 15% versions. The pilot data is not included in the main data.

4 Results

4.1 Experiment 1: Gender Differences in Compliance

We designed the PDG to investigate how individuals adjust their behavior in response to an unenforceable recommendation as well as to the observation of non-compliant peer behavior and to see whether the response differs systematically between men and women. The analysis follows the order of hypotheses laid out in Section 2. We start by estimating the overall effect of the two interventions, then progress toward analyzing potential gender differences in the effects, and finally, address the relationship between baseline behavior and individual traits like risk attitudes and social preferences.

In round 1, which was identical for all treatments, the average number of balls drawn is similar across treatments (Table 1). None of the three pairwise comparisons between treatments yields a statistically significant difference.[18] The second column of Table 1 as compared to the first illustrates three points we investigate in more detail below. First, individuals reduced the number of balls they drew over time even absent any treatment: in the Baseline treatment, the average number of balls drawn in rounds 2 to 10 is significantly lower than in round 1 (14.28 versus 17.92, Wilcoxon signed-rank test, $p < 0.0001$). Second, on average, individuals reduced the number of balls drawn by much more in Rec and in Rec&Info: In both of these treatments, the average number of balls drawn in rounds 2 to 10 is significantly lower than in Baseline despite being similar in round 1 (7.19 (Rec) or 8.59 (Rec&Info) vs. 14.28, Wilcoxon rank sum test, $p < 0.0001$ for each compared to the baseline). Third, in rounds 2 to 10, individuals drew significantly more balls in Rec&Info than in Rec on average (Wilcoxon rank sum test, $p < 0.0001$).

The decreasing number of draws suggests that individuals learned over time to draw fewer balls. The data disaggregated by rounds in Figure 1 confirms this idea. The blue line

Table 1: Average number of draws and rates of behavior consistent with recommended action (Max6=$I(\#draws \leq 6)$) in round 1 versus rounds 2-10 by treatment.

	Draws round 1	Draws round 2 to 10	Max6 (%) round 1	Max6 (%) round 2 to 10	#obs.
Baseline	17.92	14.28	15.95	28.06	514
	(9.94)	(8.46)	(36.66)	(35.48)	
Rec	18.07	7.19	16.38	72.01	586
	(9.94)	(5.18)	(37.04)	(34.54)	
Rec&Info	18.45	8.59	16.20	58.58	500
	(9.95)	(6.04)	(36.88)	(37.31)	

Notes: Max6 takes the value 1 for a participant in a given round if they has drawn 6 or fewer balls in that round, 0 otherwise. Standard deviations in parentheses.

[18] Wilcoxon rank sum tests. Baseline vs. Rec: $p = 0.9278$; Baseline vs. Rec&Info: $p = 0.4678$; Rec vs. Rec&Info: $p = 0.5187$.

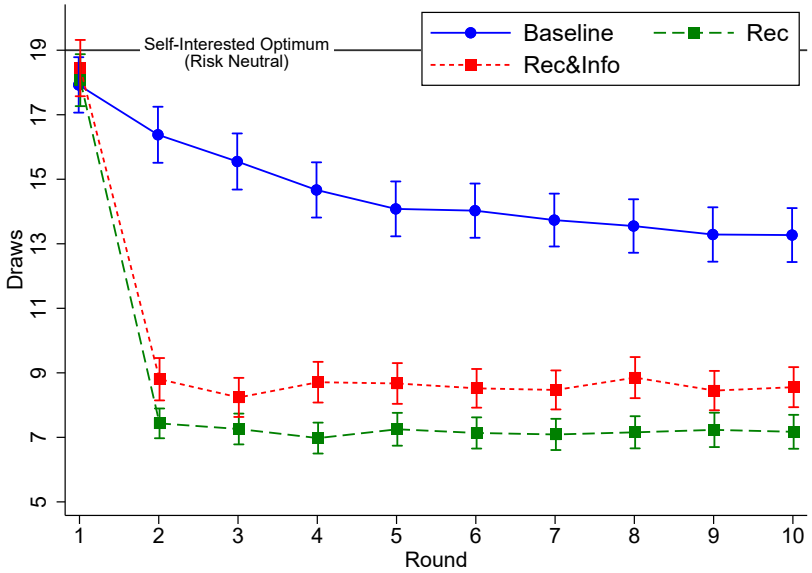


Figure 1: Draws in the different treatments, by round.

Notes: total sample = 1600 individuals (BASELINE = 514, REC = 586, REC&INFO = 500) with 10 observations per individual.

plots the average number of balls drawn in the BASELINE treatment and reveals a clear downward trend. It is evident from the figure that there is no such time trend in the two treatment conditions. Instead, the average number of balls drawn jumps down with the implementation of the recommendation in round 2 of REC and REC&INFO treatments. It stays relatively constant slightly above 7 in REC and at about 8.5 in REC&INFO.

The indicator *Max6*, which takes a value of one if an individual draws no more than six balls, is used in our analysis to investigate to what extent behavior follows the recommendation. In round 1, where treatments are identical, it does not differ significantly across treatments[19] with an average rate of draws not exceeding six of 16%. Furthermore, *Max6* mirrors the pattern observed for draws in reaction to the treatments. First, the mean value of *Max6* in rounds 2 to 10 is much higher than in round 1 already in BASELINE (Wilcoxon signed-rank test, $p < 0.0001$). Second, the mean value of *Max6* in rounds 2 to 10 is much higher in the two treatment conditions with 72% and 59% in REC and REC&INFO, respectively than in BASELINE with only 28% (Wilcoxon rank sum test, $p < 0.0001$ for each compared to the baseline). Third, the mean value of *Max6* when information about non-compliant behavior is shown is lower than with only the recommendation (Wilcoxon rank sum test, $p < 0.0001$).

For testing our hypotheses, we focus on the indicator *Max6* as a primary outcome and the absolute number of balls drawn per round by an individual (*Draws*) as a sec-

[19]None of the three pairwise comparisons yields a statistically significant difference. Wilcoxon rank sum test, BASELINE vs. REC: $p = 0.8472$; BASELINE vs. REC&INFO: $p = 0.9149$; REC vs. REC&INFO: $p = 0.9354$. The results remain unchanged when comparisons are made using paired and two-sample t-tests.

ondary outcome. Table 2 contains the results from a traditional differences-in-differences approach to investigate the main treatment effects. The *Post* dummy takes a value of 1 for observations from rounds 2 to 10, treatment dummy variables *Rec* and *Rec&Info* take a value of one for observations from the respective treatments, and *Round* is a linear time trend. The regression results show the following.

First, using observations from the Baseline and Rec treatments, the results in columns 1 and 3 of Table 2 confirm that the recommendation strongly affects behavior. With the introduction of the recommendation in round 2, the share of participants who draw at most six balls increases substantially by 49.8 percentage points (column 1) and the number of balls drawn and thereby the difference to the socially optimal number of draws decreases by almost nine (column 3). Both coefficients are significant at the 1% level. These results support Hypothesis 1.A. In addition, the significant coefficient of *Round* in both models shows that the number of balls drawn decreases significantly over time in the Baseline treatment, while the similarly sized coefficient with opposite sign on the interaction *Round*×*Rec* indicates that there is no further increase in the share of participants drawing at most six balls or a further decrease in balls drawn over time in the recommendation treatment beyond the change in round 2.

Result 1.A. *The introduction of the recommendation has a significantly positive effect on the share of individuals drawing at most six balls and a significantly negative effect on the difference between individual behavior and the socially optimal action.*

Second, using observations from the Rec and Rec&Info treatments, the results in columns 2 and 4 of Table 2 confirm the erosion of behavior in line with the recommendation when individuals are informed about non-compliant behavior of others. While the average of *Max6* increases by 56.4 percentage points between rounds 1 and 2 in the Rec treatment, the highly significant interaction term reveals that the increase is significantly smaller in the Rec&Info treatment (column 2). While the joint effect of recommendation and information remains highly significant,[20] the increase in compliance is 10.2 percentage points lower than the effect of the recommendation alone. The analogous result is obtained when we look at the average number of balls drawn: the total effect of the recommendation is reduced significantly from about nine balls to less than eight if participants receive information on non-compliant others in addition to the recommendation, as indicated by the significant coefficient on Rec&Info×Post.

Result 1.B. *In the presence of the recommendation, information on the non-compliant behavior of others significantly decreases the share of individuals drawing at most six balls and significantly increases the difference between individual behavior and the socially optimal action.*

[20] A Wald test confirms that the sum of the coefficients of *Post* and *Info*×*Post* is significantly positive ($p < 0.0001$) in column 2 and significantly negative ($p < 0.0001$) in column 4 of Table 2.

Table 2: Diff-in-diff of REC to BASELINE and REC&INFO with *Max6* and *Distance6*

	(1) Max6	(2) Max6	(3) Distance6	(4) Distance6
Post	0.0664***	0.5644***	-1.9392***	-10.9101***
	(0.0173)	(0.0176)	(0.2799)	(0.2458)
Recommendation	0.0131		-0.1427	
	(0.0277)		(0.5099)	
Rec×Post	0.4980***		-8.9709***	
	(0.0237)		(0.3831)	
Rec&Info		-0.0029		0.5006
		(0.0307)		(0.4670)
Info×Post		-0.1018***		1.1262**
		(0.0260)		(0.3631)
Round	0.0117***	-0.0011	-0.3534***	-0.0145
	(0.0018)	(0.0018)	(0.0292)	(0.0257)
Round×Rec	-0.0128***		0.3389***	
	(0.0025)		(0.0400)	
Round×Rec&Info		-0.0067*		-0.0020
		(0.0027)		(0.0379)
Constant	0.3351**	0.0403	7.7438***	13.7860***
	(0.1077)	(0.1144)	(2.1048)	(1.8335)
Observations	10470	10320	10470	10320
R^2	0.2332	0.1239	0.2340	0.1957

Notes: GLS random effects model with robust standard errors. *Max6* is an indicator for drawing no more than six balls in a given round. *Distance6* is the absolute number of balls drawn minus six (=Draws−6). *Post* is an indicator for rounds 2 to 10. Columns 1 and 3 use observations from the BASELINE and REC treatment. Columns 2 and 4 use observations from the REC and REC&INFO treatment. All estimations include as controls: female; age; education levels; log income; indicators for being a parent, being single, being working; political affiliation, and *Bundesland* (state). Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

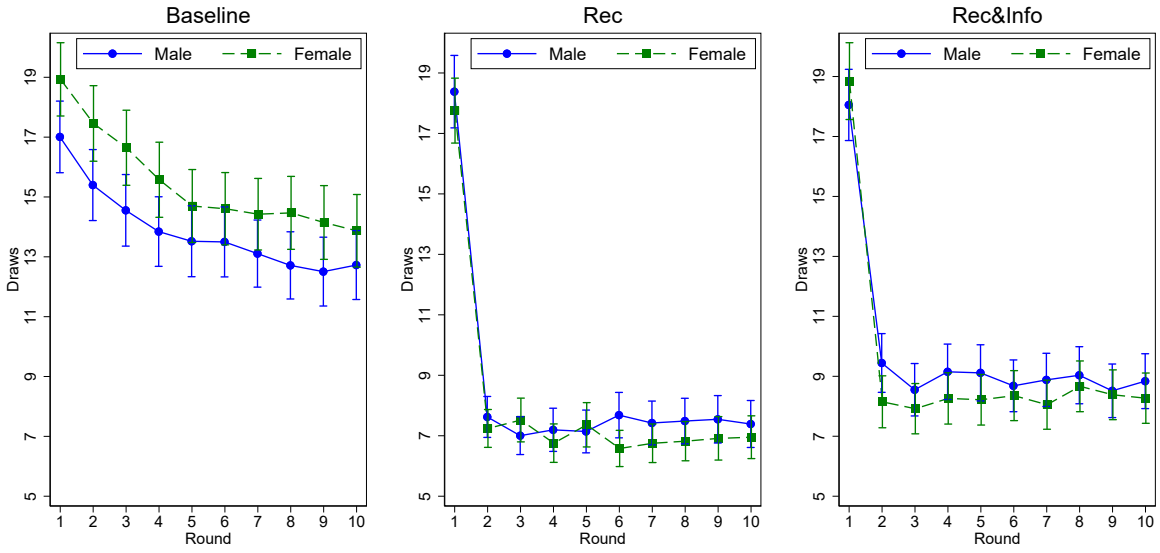


Figure 2: Number of draws in the different treatments, by round and gender.

Notes: total sample = 1600 individuals (Baseline = 514, REC = 586, REC&INFO = 500) with 10 observations per individual.

Next, we turn to Hypotheses 2.A and 2.B concerning gender differences in the observed treatment effects. Recall that we expected women to react more strongly both to the recommendation and to the negative peer information. Figure 2 plots the average number of balls drawn over time separately for male and female participants in each of the three treatments. While we observe a gender difference in the number of balls drawn in the absence of a recommendation, we do not find a significant gender difference in the share of individuals drawing no more than six balls in round 1 in any of the three treatments.[21]

To investigate potential gender differences in response to our two treatments and address our second set of hypotheses, we use a triple differences (TD) approach (Gruber, 1994). TD is an important tool for providing causal evidence without the need to prove the parallel trends assumption required by the regular difference-in-differences approach (Olden and Møen, 2022). As the TD approach takes the difference between difference-in-differences, it is also called *difference-in-difference-in-differences* (DDD). In addition to the simple difference-in-differences model above, we now include a dummy *Female*, which takes the value 1 if a participant is female, and the corresponding interaction terms. As in our previous analysis, we also include a linear time trend *Round* and its interaction with *Rec* to allow the time trend to be different in the REC treatment than in the BASELINE.

[21]In round 1, the shares of individuals with *Draws*$\leq$ 6 is 16.36% (male) and 15.51% (female) in BASELINE ($p = 0.7937$), 17.91% (male) and 14.83% (female) in REC ($p = 0.3146$), and 14.68% (male) and 17.74% (female) in REC&INFO ($p = 0.3537$), where the p-values are from Wilcoxon rank sum tests. The values are virtually identical in two-sample tests of equal proportions.

Again, we use data from BASELINE and REC to investigate potential gender differences in the reaction to the recommendation. We estimate the following equation,

$$
\begin{aligned}
(4) \qquad \text{Max6}_{ict} = \; & \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Post} + \beta_3 \text{Post} \times \text{Female} + \beta_4 \text{Rec} \\
& + \beta_5 \text{Rec} \times \text{Post} + \beta_6 \text{Rec} \times \text{Female} \\
& + \beta_7 \text{Rec} \times \text{Female} \times \text{Post} \\
& + \beta_8 \text{Round} + \beta_9 \text{Round} \times \text{Rec} + \varepsilon_{ict},
\end{aligned}
$$

where we are mainly interested in the coefficient β_7 on the triple interaction variable *Rec*×*Female*×*Post*. If β_7 is positive and significant, this indicates that the recommendation is significantly more effective for women than for men in that it moves more female than male actions from above to weakly below the recommended action, which would be in line with Hypothesis 2.A. In a second step, we use data only from the REC and REC&INFO treatments to speak to the potentially gendered effect of information from Hypothesis 2.B using an appropriately adjusted equation, where *Rec* is substituted for with *Rec&Info* and only observations from REC and REC&INFO are included. We conduct analogous estimations with the difference between individual draws and the socially optimal action. The estimation results are collected in Table 3.

In line with the previous difference-in-difference estimations, we see in column 1 of Table 3 that the share of individuals who draw at most six balls increases over time but much more so if the recommendation is introduced in round 2. The coefficients of the interactions *Post*×*Female* and *Rec*×*Female* show that women are more likely than men to continue drawing more than six balls in rounds 2 to 10 in the BASELINE treatment, i.e. in the absence of a recommendation. But women react more strongly than men to the introduction of the recommendation as indicated by the significant coefficient of the triple interaction term *Rec*×*Post*×*Female*: in reaction to the recommendation, the share of females drawing no more than six balls increases by another 11.5 percentage points beyond the increase of 44.3 percentage points observed for male participants. This additional increase in female behavior following the recommendation is partly but not only due to the observed gender differences in the baseline treatment. Women reduced the number of balls they drew after round 1 less than men in BASELINE so that there is more scope for the recommendation to affect behavior. But the sum of coefficients on *Post*×*Female* and the triple interaction term leaves a significantly positive effect of about 6 percentage points indicating that the recommendation was more effective for women than men even if we correct for the difference in potential effectiveness.

When looking at the difference between absolute draws and the socially optimal action as the dependent variable, we do not find a significant gender difference in the treatment effect of the recommendation. This may relate to the fact that female participants in

Table 3: Triple-diff of Rec to Baseline and Rec&Info with *Max6* and *Distance6*

	(1) Max6	(2) Max6	(3) Max6	(4) Distance6	(5) Distance6	(6) Distance6
Female	-0.0242	-0.0213	-0.0204	2.4105**	-0.3365	2.4053**
	(0.0406)	(0.0396)	(0.0426)	(0.7481)	(0.5999)	(0.7947)
Post	0.0916***	0.5347***	0.0916***	-1.7232***	-10.9615***	-1.7232***
	(0.0221)	(0.0229)	(0.0233)	(0.3566)	(0.3202)	(0.3734)
Post×Female	-0.0547°	0.0606*	-0.0547°	-0.4684	0.1049	-0.4684
	(0.0296)	(0.0300)	(0.0313)	(0.4789)	(0.4191)	(0.5016)
Rec	0.0199			0.9317		
	(0.0380)			(0.7002)		
Rec×Post	0.4431***			-9.2383***		
	(0.0305)			(0.4933)		
Rec×Female	-0.0134			-2.3460*		
	(0.0552)			(1.0173)		
Rec×Post×Female	0.1152**			0.5733		
	(0.0405)			(0.6546)		
Rec&Info		-0.0266	0.0099		-0.2617	0.4729
		(0.0418)	(0.0428)		(0.6338)	(0.8020)
Rec&Info×post		-0.0671*	0.3760***		2.0210***	-7.2173***
		(0.0337)	(0.0335)		(0.4717)	(0.5373)
Info.*female		0.0484	0.0402		1.5735°	-0.7990
		(0.0583)	(0.0603)		(0.8823)	(1.1241)
Rec&Info×Post×Female		-0.0708	0.0444		-1.8468**	-1.2735°
		(0.0443)	(0.0446)		(0.6191)	(0.7141)
Round	0.0117***	-0.0011	0.0117***	-0.3534***	-0.0145	-0.3534***
	(0.0018)	(0.0018)	(0.0019)	(0.0292)	(0.0257)	(0.0306)
Round×Rec	-0.0128***			0.3389***		
	(0.0025)			(0.0400)		
Round×Rec&Info		-0.0067*	-0.0195***		-0.0020	0.3369***
		(0.0027)	(0.0027)		(0.0379)	(0.0437)
Constant	0.3329**	0.0638	0.0144	6.0671**	13.8136***	10.7764***
	(0.1086)	(0.1155)	(0.1171)	(2.1601)	(1.8493)	(2.3844)
Observations	10470	10320	9610	10470	10320	9610
R^2	0.2354	0.1241	0.1255	0.2392	0.1962	0.1769

Notes: GLS random effects model with robust standard errors. *Max6* is an indicator for drawing no more than six balls in a given round. *Distance6* is the absolute number of balls drawn minus six (=Draws−6). *Post* is an indicator for rounds 2 to 10. Columns 1 and 4 use observations from the Baseline and Rec treatment. Columns 2 and 5 use observations from the Rec and Rec&Info treatment. Columns 3 and 6 use observations from the Baseline and Rec&Info treatment. All estimations include as controls: age; education levels; log income; indicators for being a parent, being single, being working; political affiliation, and *Bundesland* (state). Standard errors in parentheses. ° $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

the Rec treatment in round 1 drew significantly fewer balls than female participants in round 1 of the baseline treatment and are therefore already closer to the social optimum even though there is no significant gender difference in the indicator *Max6* in round 1. These gender differences originate in a range of ball draws far above six and likely reflect differences in preferences that correlate with gender instead of attitudes toward compliance (see column 4 in Table 3 and Figure 2 for details and the results on preferences and beliefs below).

Result 2.A. *The effect of the recommendation is larger for women than for men.*

Comparing behavior in Rec&Info and Rec, we find that the effect of observing non-compliant behavior in others is negative but male and female participants react similarly to observing non-compliant behavior in others (see column 2 of Table 3). In Figure 2, middle and the right panel, we see that the number of balls drawn is higher when information on non-compliant behavior is given (Rec&Info) than when it is not (Rec) for both male and female participants. This observation remains true when we look at the extent to which behavior follows the recommendation as shown in column 2 of Table 3. The share of decisions to draw at most six balls decreases by about 7 percentage points once information on deviant others is shown to participants and the triple interaction with gender does not gain statistical significance at conventional levels (column 2, $p = 0.110$).

Since this result was unexpected (cf. Hypothesis 2.B), we investigate it in more detail. Indeed, when we estimate a simple difference-in-differences model separately for men and women, the coefficient on *Rec&Info*×*Post* is about -0.1 in the male and -0.17 in the female subsample, suggesting that women tend to react more strongly to observing non-compliant behavior. However, the effect size seems to be too small for our study to pick it up even though the absolute difference appears large (see Table 7 in Appendix C for details). When it comes to the difference between the number of draws per round and the social optimum, the triple interaction term is significantly negative. In combination with the previous result, this indicates that women's behavior reacts differently to observing non-compliant behavior but these differential changes are inframarginal and affect the number of balls drawn without shifting behavior below the recommended threshold level of draws (see column 5 in Table 3 and for results separated by gender Table 8 in Appendix C).

Result 2.B. *The effect of observing non-compliant behavior of others in the presence of the recommendation does not differ significantly between women and men.*

Despite this negative result, we have one additional piece of evidence suggesting that female behavior might still be more malleable than that of men's when exposed to observing non-compliant behavior in others: We estimate the joint effect of the recommendation

(where women react more strongly) and negative peer information by comparing behavior only in the Rec&Info and Baseline treatments. In these estimations, shown in columns 3 (for *Max6*) and 6 (for *Distance6*) of Table 3, when compared to columns 1 and 4, we find no significant gender difference in the treatment effect for *Max6* anymore and a negative treatment effect for the difference between draws and the social optimum, significant at 10% level. This result is consistent with women generally following the recommendation more often than men but also slacking off more often when observing that others do not follow it. In consequence, the joint effect of the two, i.e. the difference between Rec&Info and Baseline is similar for both genders. Note that this result also contradicts the specific part of our hypothesis stating that men do not react to observing empirical information.

Next, we are interested in better understanding how the perceived riskiness of the decision situation affects behavior as specified in Hypothesis 3. To analyze the relationship, we generate two variables from the subjective assessments that participants answer after each decision round that shed light on the subjective risk assessment. The indicator *Optimistic* takes a value of one if an individual expects not to have drawn any black ball in the given round even though the actual expected value of black balls for their number of draws is one or larger.[22] The variable *Risk Perception* elicits a subject's perceived risk from drawing another ball, conditional on their draws in that round.[23] To mitigate concerns of endogeneity, we employ a panel regression to study the relationship between optimism and perceived risk on *next* round decisions, controlling for current decisions. The regressions partially confirm our hypotheses. Individuals who are optimistic are less likely to draw no more than six balls but optimism does not correlate significantly with the number of balls drawn. Further, individuals who assess their number of draws as more risky are more likely to draw at most six balls and they generally draw fewer balls per round on average. The results are collected in Table 4 and summarized below.

Result 3. *Individuals who perceive the situation as less risky, measured as less optimistic or by their direct statement, are more likely to draw no more than six balls. The correlation with the absolute number of draws is only significantly negative for stated riskiness.*

While the set of hypotheses 4.A, 4.B, and 4.C cannot be investigated using exogenous variation and were not preregistered, we still find it instructive to analyze the correlation of the specified individual characteristics and beliefs with individual behavior in the experiment as hypothesized above. To do so, we use data only from the Baseline treatment.

[22] After every decision round, participants were reminded how many balls they had drawn and then had to answer the incentivized question: "How many of the [number of draws] balls you have drawn you think were black?" This is compared with the expected value of black balls conditional on the individuals *Draws* in that round.

[23] The question used for this variable is "Recall how many balls you have just drawn. How risky do you feel it would have been to draw another ball?" and was also asked after every round.

Table 4: Relationship of optimism and perceived risk with *Max6* and *Draws*

	(1) Max6	(2) Max6	(3) Draws	(4) Draws
L.Optimistic	-0.0577***		-0.0853	
	(0.0131)		(0.2162)	
L.Risk Perception		0.0193***		-0.4035***
		(0.0029)		(0.0507)
Take risk	-0.0065**	-0.0081***	0.1122**	0.1542***
	(0.0022)	(0.0022)	(0.0380)	(0.0381)
L.Max6	0.6169***	0.6274***		
	(0.0128)	(0.0121)		
L.Draws			0.7596***	0.7483***
			(0.0097)	(0.0097)
Observations	4392	4392	4392	4392
R^2	0.4184	0.4216	0.6321	0.6373

Notes: GLS random effects model with robust standard errors using observations only from Baseline. *Max6* is an indicator for drawing no more than six balls in a given round. *Draws* is the number of balls drawn in a given round. *Optimistic* takes a value of one if an individual expects not to have drawn any black ball in the given round even though the actual expected value of black balls for their number of draws is one or larger. *Risk Perception* is subjective perceived risk from drawing another ball, conditional on their draws in that round. *Take risk* contains answers to the question about general willingness to take risks used in SOEP and other surveys. *L.* indicates lagged values. All regressions control for gender as well as age; education levels; log income; indicators for being a parent, being single, being working; political affiliation, and *Bundesland* (state). Standard errors in parentheses. $^{\circ}$ $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

As the dependent variable, we use an individual's average behavior over rounds 1 to 10 in the PDG. We start by analyzing the correlation between the participant's willingness to take risks and their decisions in the PDG. The willingness to take risk was elicited in a post-experimental questionnaire using an established survey question. It is coded in *Take risk* where higher values indicate a larger willingness to take risks.[24] As this variable is constant at the individual level over time, we use it in a regression with the individual's average behavior over rounds 2 to 10 as the dependent variable. In line with hypothesis 4.A, we find a significantly negative correlation with the share of rounds in which the individual drew at most six balls in Baseline and a significantly positive one with the average number of draws per round and a (column 1 and 4, Table 5).

Next, we turn to the question of how social preferences relate to observed behavior using both a participants' tendency to put group welfare ahead of their individual payoff (*Pro-social*) and their expectation of the extent to which others do so (*Others pro-social*) according to their self reports and expectations about other participants.[25] In line with

[24]The question we used is taken from the individual questionnaire used in the German Socioeconomic Panel (SOEP) and reads as "Are you generally willing to take risks or do you try to avoid risks?" where answer options ranged from 0 (not at all willing to take risk) to 10 (very much willing to take risks).

[25]In the questionnaire, we asked how important on a scale from 0 (not at all important) to 10 (very important) it was for the respondent to a) maximize their individual bonus and b) not reduce the group

Hypothesis 4.B, we find a weakly significant negative correlation between a subject's pro-group orientation and their average number of draws per round (column 5, Table 5) but the hypothesized positive correlation with an individual's propensity to draw no more than six balls is not significant at conventional levels (column 2). With respect to Hypothesis 4.C, we find no significant correlation between expectations of others' pro-sociality and average behavior (columns 3 and 6, Table 5).

Table 5: Relationship of willingness to take risk, own and other social preferences with *Max6* and *Draws*

	(1) Avg.Max6	(2) Avg.Max6	(3) Avg.Max6	(4) Avg.Draws	(5) Avg.Draws	(6) Avg.Draws
Take risk	-0.0159*			0.4296**		
	(0.0064)			(0.1492)		
Pro-social		0.0213			-0.7008°	
		(0.0168)			(0.3937)	
Others pro-social			-0.0053			-0.2786
			(0.0164)			(0.3847)
Observations	488	488	488	488	488	488
R^2	0.0732	0.0638	0.0607	0.1535	0.1440	0.1391

Notes: OLS model with robust standard errors using observations only from Baseline. *Avg.Max6* (*Avg.Draws*) is the individual-level average over rounds 1 to 10 for *Max6* (*Draws*) and *Max6* is an indicator for drawing no more than six balls in a given round. *Take risk* contains answers to the question about the general willingness to take risks used in SOEP and other surveys. *Pro-social* is constructed as the reported importance of the group bonus minus the reported importance of own bonus and standardized to have mean zero and standard deviation of one. *Others pro-social* is constructed as the expected importance of the group bonus for others minus the expected importance of the own bonus for others and standardized to have mean zero and standard deviation of one. All regressions control for gender as well as age; education levels; log income; indicators for being a parent, being single, being working; political affiliation, and *Bundesland* (state). Standard errors in parentheses. $^{\circ}$ $p < 0.10$ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.2 Experiment 2: Norm elicitation

In Personal Beliefs, we asked participants to report their beliefs regarding the number of balls one should draw in the Rec treatment of Experiment 1. Male and female participants expressed similar personal beliefs regarding the number of draws (mean of draws being 5.79 (SD=3.62) for men and 5.16 (SD=1.90) for women, Wilcoxon rank sum test, $z = 0.731$, $p = 0.4650$). That is, the two genders have the same personal beliefs

bonus. From this, we construct the variable *Pro-social* as the reported importance of the group bonus (as a proxy for group orientation) minus the reported importance of the own bonus (as a proxy for selfish behavior). Respondents further stated how important on a scale from 0 (not at all important) to 10 (very important) they thought it was for other group members to a) maximize their own bonuses and b) not reduce the group bonus. We use these to construct *Others pro-social* as the expected importance of the group bonus for others (as a proxy for others' group orientation) minus the expected importance of the own bonus for others (as a proxy for others' selfish behavior). Both variables are standardized to have a mean of zero and a standard deviation of one.

about how many balls one should draw in the presence of a recommendation. The results from the KW Method, following Krupka and Weber (2013) in analyzing social norms, are fully in line with these results. When asked to grade the appropriateness of each number of draws in the Rec treatment on a Likert scale ranging from 1 (very inappropriate) to 4 (very appropriate), we did not find any significant gender difference in this treatment. All the p-values and z-scores from the Wilcoxon rank-sum (Mann-Whitney) tests are reported in Appendix E.1.

In the Normative Expectations treatment, participants received two incentivized questions aimed at measuring general and gendered normative expectations. The first one asked participants to guess which answer was given by the majority of participants in the Personal Beliefs treatment.[26] Based on the original answers, which related to the number of balls to be drawn, we code an indicator analogous to the variable *Max6* in Experiment 1. We find that 82.5% of the participants in Normative Expectations expect the majority of the participants in the Personal Beliefs treatment to report that one should draw a number of balls smaller than or equal to six when it is recommended to do so. In other words, a large majority believe that others consider it the right thing to do to follow the recommendation. This result is the first proof that complying with the recommendation in Experiment 1 is an injunctive norm. The next question elicited normative expectations with respect to the female participants in Personal Beliefs. Using the same recoding as before, we find that 81% of the participants in Normative Expectations expect that the majority of the female participants in Personal Beliefs stated that one should draw at most six balls. Thus, we have two measures of normative expectations, one for the mixed-gender sample and a separate one only for women. The McNemar's chi-square test revealed no significant difference between the participants' expectations in the Normative Expectations treatment about women's personal beliefs and the personal beliefs of a mixed-gender sample (McNemar's chi-square test, McNemar's $\chi^2 = 0.25$, $p = 0.6171$).

In the third treatment, we elicited participants' empirical expectations about behavior in Experiment 1. We asked them to separately report how many balls they expected the majority of all participants and the majority of all female participants in Rec, respectively, to have drawn. The answers to these two questions were incentivized using the actual results of the treatment Rec. We recode the answers again into an indicator that takes a value of one for answers of at most six balls. We find that the empirical expectations about female behavior differ from those about the behavior of the mixed-gender participants in Rec (McNemar's chi-square test, McNemar's $\chi^2 = 10.71$, $p = 0.0011$). Specifically, 83.8% (SD=0.03) of the participants expected women to follow the recommendation and draw at most six balls, whereas this share is only 72.8% (SD=0.04) when

[26]The actual results of Personal Beliefs were used to incentivize these guesses.

asked about the mixed-gender sample (see Appendix E.2 for the distributions of expected draws by gender for relevant treatments).

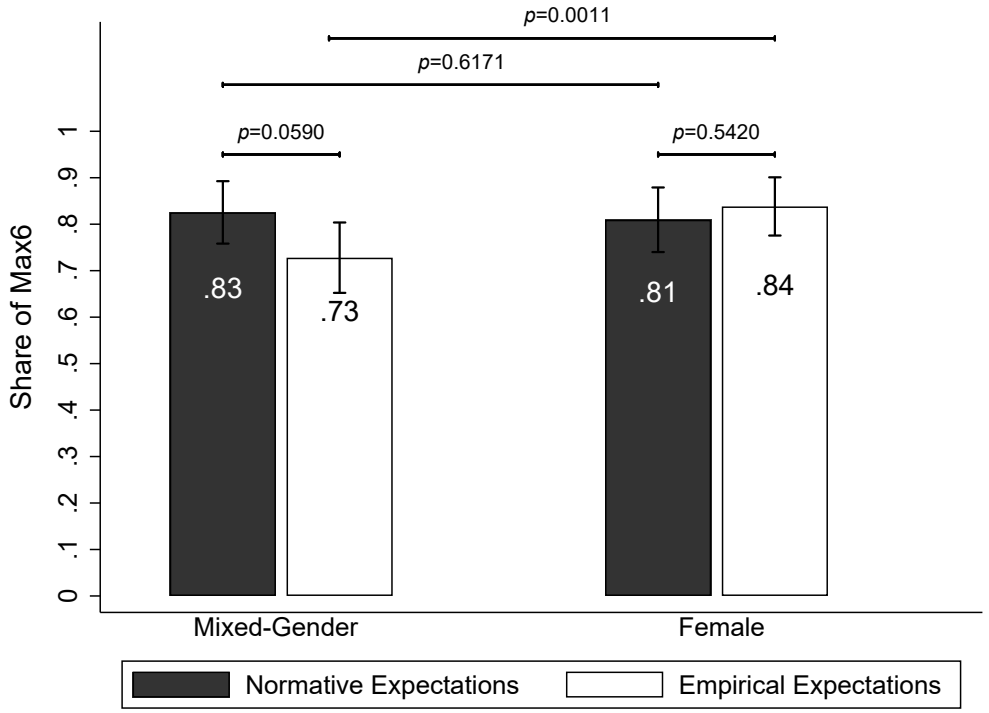


Figure 3: Social norm elicitation.

Notes: The figure illustrates the expectations of all participants regarding the share of *Max6*. The left panel represents expectations for a mixed-gender group, while the right panel focuses on expectations for a female group. Bars represent means, and error bars represent 95% confidence intervals. Proportions were compared across treatments using Pearson's and McNeamer's chi-squared tests, with p-values reported where appropriate.

In the next step, we check whether the normative expectation with respect to behavior in the presence of the recommendation differs from the empirical expectation about a) female and b) mixed-gender participants. We find no significant difference in these two types of expectations when they are elicited about females (Pearson's chi-squared test, Pearson's $\chi^2 = 0.3725$, $p = 0.542$). We conclude that the Normative Expectations and the Empirical Expectations concerning women coincide in our study. However, the difference in expectations referring to all participants, i.e. a mixed-gender sample, is significant at 10%-level (Pearson's chi-squared test, Pearson's $\chi^2 = 3.5602$, $p = 0.059$). As stated above, 72.8% of respondents expect the majority of the mixed-gender sample to comply with the recommendation and draw at most 6 balls, whereas 82.5% of respondents expect the majority of participants to believe that one should not draw more than six balls. Thus, the expectation of actual compliance falls short of the expectation of personal beliefs. When we impute the expectations about male behavior from those concerning the mixed-gender and female samples, the result becomes even clearer: men are significantly less often expected to actually follow the recommendation (59.6%) than they are expected

to believe one should do so (72.2%) (Pearson's chi-squared test, Pearson's $\chi^2 = 4.6509$, $p = 0.031$).

When we look at the results from our Normative Expectations and Empirical Expectations treatments in light of the social norm definition by Bicchieri (2016), we conclude that there exists a social norm for women to follow the recommendation and choose a number of balls consistent with it because the normative and empirical expectations about women coincide. The presence of such a social norm can be expected to increase the extent to which women actually follow the recommendation. In contrast, the same analysis for the mixed-gender sample shows that men are expected to follow the recommendation significantly less often than women and to also follow it significantly less often than personal beliefs would suggest. This result suggests that there is not a clear social norm for men to follow the recommendation. Hence, our results support Hypothesis 5 that a social norm to follow the recommended action applies to women but rather not to men (see Figure 3).

Result 4. *Following the recommendation is a social norm for women, but not clearly so for men.*

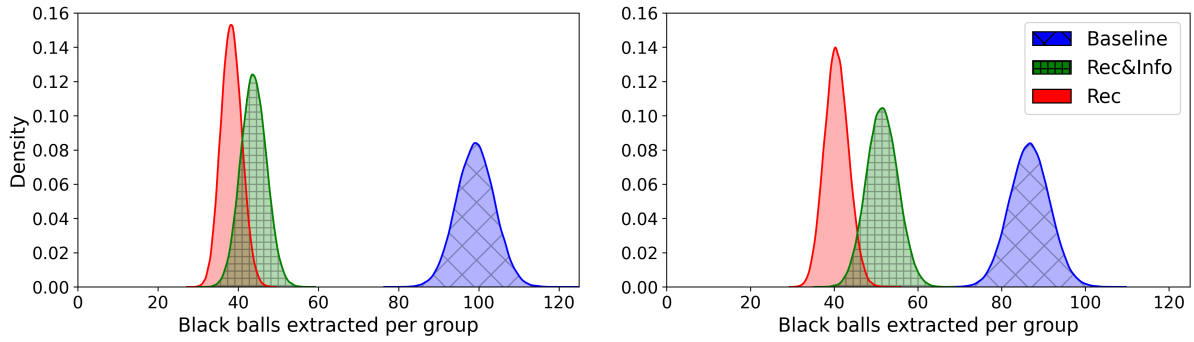
In light of the results from the four norm-related treatments, we conclude that there exists a female social norm that prescribes following the recommendation and drawing no more than six balls, whereas men are expected to deviate from it more often. Empirical expectations fit the observed behavior, as men were less likely to follow the recommendation than women in Experiment 1.

5 Discussion of welfare implications

Our results on the main effect of the recommendation and its interaction with observing non-compliant behavior in others suggest that the recommendation is an effective tool to improve social welfare even when some deviant behavior is observed. The estimated main effects translate into drastic changes in the total harm incurred at the group level as any reduction in the number of balls drawn per individual leads to a stark reduction in the number of black balls drawn within each group. To illustrate the welfare implications, we use our experimental data to simulate behavior for 100,000 groups of 100 participants each.[27] Figure 4, illustrating simulated harm in round 2, which is the first round where treatment effects apply, shows that the recommendation in Rec affects men

[27]For each round and each condition, we randomly sample 100 participants and use their behavior to compute the expected group harm. Considering that the total harm is in part stochastic (two individuals could draw the same number of balls but extract a different number of black balls), we consider the expected amount of harm to the public good given the balls drawn by the individual, independent of the actual number of balls drawn in the experiment. This also allows us to verify the robustness of our results independent of the specific group combinations.

and women differently (Wilcoxon rank-sum test, z=2.439, p-value=0.015) with women being responsible for significantly lower harm than men. On average, women are simulated to draw 0.44 black balls and men 0.52. This difference is especially striking when comparing it with the baseline treatment, where women inflict significantly more harm on the public good than men in that, on average, women are simulated to draw 0.99 black balls but men only 0.87 (Wilcoxon rank-sum test, z=-2.349, p-value=0.019). When the recommendation and information about non-compliant behavior of others come together, both genders harm the public good to a similar extent (Wilcoxon rank-sum test, z=0.433, p-value=0.665). As is clearly visible from Figure 4, total harm will still be substantially lower in Rec&Info than in Baseline, documenting the effectiveness of the recommendation even it is known not to be followed by everyone. In the simulations, the average number of black balls drawn in round 2 in the Baseline treatment is 92.64 (SD=5.49), resulting in an average round payoff of €68.90 per group. In the Rec and Rec&Info treatments, the average number of black balls drawn drops to 39.44 (SD=3.04) and 47.73 (SD=4.03), respectively, as individuals strongly adjust their behavior toward the social optimum as is recommended to them. This aggregate behavior change leads to substantial increases in average round 2 payoffs to €105.55 in Rec and €98.83 in Rec&Info per group. Further figures, describing the social harm generated in all other rounds, can be found in Appendix F.



(a) Female participants only

(b) Male participants only

Figure 4: Simulation results for social harm by treatment and gender

Notes: The figure shows the simulated distribution of total black balls drawn in round 2 per group of 100 participants, where each group consists of 100 randomly drawn female or male participants. The simulation uses observed behavior in Experiment 1 from the randomly drawn participants to calculate the expected number of black balls drawn per group. The figure illustrates the results for 100,000 simulations.

The substantial welfare gains highlight the effectiveness of simple normative interventions in promoting socially beneficial behavior. They can be interpreted as the result of an information nudge in the form of the recommendation that is designed to align individual behavior with the social optimum. In this sense our findings relate to the theoretical analysis by Mariotti et al. (2023), who analyze how information nudges can help individuals overcome self-control problems. In their model, individuals face immediate positive gains

versus a heterogeneous risk of incurring harm in the future. Their information nudges inform individuals, who initially know only the distribution of risks but not their individual risk, whether their risk realization is above or below a threshold value. Even though not binding, the provided information is shown to improve decision-making and thereby welfare. In our experiment, the recommendation instead provides information about socially optimal behavior in a complex dynamic setting, where immediate benefits are traded off against uncertain future harm to the group.

While we see a clear positive effect on monetary payoffs, which is the best proxy for individual welfare in our experiment, previous studies suggest that nudges or interventions that push individuals to behave more prosocially through social pressure or reference to norms may have non-monetary side effects (see e.g. DellaVigna et al., 2012, and Allcott and Kessler, 2019). Such side effects may also occur in response to the recommendation in our experiment because individuals cannot avoid the information we give them about socially desirable behavior. Serra-Garcia and Szech (2022) find that individuals indeed often prefer to remain ignorant about the moral implications of their actions to choose more individually advantageous behavior that comes at a social cost. They cannot uphold their ignorance in the presence of the recommendation in our experiment but may feel they have to adjust their behavior in light of the new information. At the same time, the recommendation and implies change in behavior may lead to negative feelings as in the studies cited above. Finally, our results highlight the potential ambiguity of revealing information about peer behavior on outcomes. While Ehrlich and Szech (2022) show that informing individuals about a prosocial act of another one encourages the observing individual to engage in that same act (downloading the Corona tracing app in that context), i.e. observing socially desirable behavior reinforces the desirable behavior, our results suggest that observing undesirable behavior likewise fosters the prevalence of undesirable behavior. Specifically, our evidence suggests that individuals who observe others to deviate from a recommended action are more likely to deviate themselves.

6 Conclusion

Our study addresses a common but understudied claim: women are on average more compliant with rules, norms, or recommendations than men and more generally more susceptible to social cues, whether positive or negative. To test this assertion systematically, we implement a novel experimental design, the Probabilistic Dilemma Game (PDG), where participants are matched in groups and individually decide how many balls to draw. Each ball carries a private benefit alongside the risk of private and social harm. Using two online experiments with representative samples of the German adult population, we analyze how behavior in this game changes in response to two exogenous treatment ma-

nipulations, how it relates to individual characteristics like perceived risk, optimism, or pro-social attitudes, and how norms may help explain the observed behaviors.

In Experiment 1, the subjects directly play the PDG in groups of 100 in one of three treatments. In addition to a baseline without any intervention, the treatment Rec introduces a recommendation that suggests to participants that they should draw at most six balls – the number of balls maximizing the expected group payoff. The treatment Rec&Info introduces the same recommendation in combination with information about the behavior of four past participants who did not adhere to this recommendation in most rounds. We first establish the main effects: The recommendation is highly effective and significantly increases the share of participants drawing at most six balls by 44.3 percentage points. The additional provision of information on non-compliant behavior in others decreases the effect of the recommendation significantly by 10.2 percentage points but the joint effect as compared to the baseline remains substantially positive and significant. The subsequent analysis of gender differences in the treatment effects yields differentiated results. First, we find that, as expected, women react more strongly to the recommendation than men. Second, and somewhat surprisingly, we find no robust evidence that women react more strongly than men to observing non-compliant behavior, even though the point estimate of the information effect is larger for women. We further find that individual preferences and beliefs correlate with baseline behavior in an intuitive way. Stronger pro-social preferences tend to reduce the number of balls an individual draws whereas a lower perceived risk and larger willingness to take risk increase the number of balls drawn and reduce the likelihood that an individual's behavior conforms to the social optimum in the absence of the recommendation.

The contribution of Experiment 2 to understanding compliance and gender differences therein is twofold. First, the results demonstrate that complying with the recommendation is considered an injunctive norm in the studied sample. Second, the results provide evidence that for women, compliance is considered not only an injunctive but also a descriptive norm for women, making compliance a social norm for women. In contrast, we do not find this match of empirical and normative expectations for men: while men are expected to perceive the same injunctive norm as women, they are not believed to behave accordingly, implying a differing descriptive norm of less compliance for men, which will imply if anything a weaker social norm of compliance for men. This difference in the social norm for men and women may rationalize the gender difference in compliance from Experiment 1, where women comply with the recommendation significantly more often than men.

From a policy perspective, our results and the discussion of the implied welfare effects suggest that interventions providing clear normative guidance have the potential to significantly enhance social welfare in settings characterized by probabilistic harm because

individuals will follow it. Caution needs to be taken insofar as such interventions are most effective if individuals only observe behavior that conforms with the recommendation. Any observed deviations may trigger an erosion of compliance as such observations reduce the strength of the normative signal contained in the recommendation.

Our results thereby contribute to the literature on social norms and their role in promoting desirable behavior. Previous research has shown that social norms are more effective in influencing behavior if injunctive and descriptive norms align with each other. In particular, an injunctive norm can become ineffective if observed behavior as a cue for the descriptive norm is in conflict with it. We show that the recommendation establishes an injunctive norm that affects behavior more strongly for women, for whom the perceived descriptive norm aligns with the recommended behavior, whereas the effect is weaker for male participants for whom injuntive and descriptive norm do not align. In this respect, public policy plays a crucial role in reducing such normative conflict which will in turn improve compliance with an existing injunctive norm and thereby improve social welfare. For example, during Covid-19 pandemic, public health messages from authorities reached the public in much of the world indeed before they learned about the behavior of others, which allowed the public recommendations to shape both normative and empirical expectations and thereby maximize compliance in community settings. In this and similar cases, it is plausible to conclude that injunctive norms can be shaped by policymakers and supported by examples of norm-compliant behavior. Later, it will be complemented by a descriptive norm that emanates from communities and the actually observed behavior. From a policy point of view, it is important to focus on the development of this descriptive norm because small deviations in behavior may trigger an erosion of norm-guided behavior. Our experiment captures some basic features of this complex dynamics – injunctive norms are strong but observing non-compliant peer behavior can drive the descriptive norms down. We further document non-negligible gender differences in compliance with a recommendation but not in reaction to non-compliant others. These results show on the one hand that policies that aim at distributing the burden of compliance equally on men and women need to take into account the stronger reaction of women to normative cues. But one the other hand, the results suggest that gender differences in compliance with a recommendation may diminish over time as participants necessarily learn about the behavior of others and female participants tend to exhibit stronger erosion when observing non-compliant others. All of these findings would benefit from further research.

Our study further makes a methodological contribution by proposing a new online implementable tool to examine behavior in probabilistic social dilemma settings, the PDG. The PDG affords a lot of flexibility and could for instance be adjusted to different group sizes, to be run with or without a shared identity, and to vary the levels of individual or social harm over time or in relation to each other. While the PDG shares similarities with the bomb risk elicitation task (BRET) by Crosetto and Filippin (2013), a method to

measure risk preferences in a real-time box opening game, there are important differences. In the BRET, participants decide about the number of boxes they want to open, knowing that one contains a bomb which would make them lose their entire earnings. In the PDG, in contrast, players participate in groups and the decision of how many balls to draw does not only influence an individual's expected payoff from a private account, which becomes zero as soon as the individual draws a black ball, but it also affects the expected payoff of all members of their group through its effect on a group account that is equally shared within the group. This group account decreases with the sum of black balls drawn in the group so that choices matter to the group even if the individual believes to have lost their private earning already. The PDG relates to a small literature on social dilemmas with probabilistic losses, where more selfish behavior of a group increases the probability that the group experiences a loss. In contrast to Blanco et al. (2017), where the probability of a group loss depends on the aggregate behavior of the group and is not tied to a separate private loss, the PDG couples the two types of losses to each other and keeps the probabilistic component at the individual: every ball drawn may be black with a certain probability and if so triggers a sure loss for the individual and in addition for the group.

References

Akfırat, S., Bayrak, F., Üzümçeker, E., Ergiyen, T., Yurtbakan, T., and Uysal, M. S. (2023). The roles of social norms and leadership in health communication in the context of covid-19. *Social Science & Medicine*, 323:115868.

Alger, I. and Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302.

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):236–276.

Alm, J., Jackson, B. R., and McKee, M. (2009). Getting the word out: Enforcement information dissemination and compliance behavior. *Journal of Public Economics*, 93(3-4):392–402.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477.

Apesteguia, J., Funk, P., and Iriberri, N. (2013). Promoting rule compliance in daily-life: Evidence from a randomized field experiment in the public libraries of barcelona. *European Economic Review*, 64:266–284.

Bahk, J., Yun, S.-C., Kim, Y.-m., and Khang, Y.-H. (2015). Impact of unintended pregnancy on maternal mental health: a causal analysis using follow up data of the panel study on korean children (pskc). *BMC Pregnancy and Childbirth*, 15(1):1–12.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bicchieri, C. and Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, 23(2):161–178.

Bicchieri, C. and Dimant, E. (2022). Nudging with care: The risks and benefits of social information. *Public Choice*, 191(3-4):443–464.

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.

Blanco, E., Haller, T., and Walker, J. M. (2017). Externalities in appropriation: responses to probabilistic losses. *Experimental Economics*, 20(4):793–808.

Blanco, M., Engelmann, D., Koch, A. K., and Normann, H.-T. (2014). Preferences and beliefs in a sequential social dilemma: a within-subjects analysis. *Games and Economic Behavior*, 87:122–135.

Bonan, J., Cattaneo, C., d'Adda, G., and Tavoni, M. (2020). The interaction of descriptive and injunctive social norms in promoting energy conservation. *Nature Energy*, 5(11):900–909.

Bouchery, E. E., Harwood, H. J., Sacks, J. J., Simon, C. J., and Brewer, R. D. (2011). Economic costs of excessive alcohol consumption in the us, 2006. *American Journal of Preventive Medicine*, 41(5):516–524.

Brekke, K. A., Kipperberg, G., and Nyborg, K. (2010). Social interaction in responsibility ascription: The case of household recycling. *Land Economics*, 86(4):766–784.

Brekke, K. A., Kverndokk, S., and Nyborg, K. (2003). An economic model of moral motivation. *Journal of Public Economics*, 87(9-10):1967–1983.

Brent, D. A., Lott, C., Taylor, M., Cook, J., Rollins, K., Stoddard, S., Brent, D., Lott, C., Taylor, M., and Cook, J. (2017). Are normative appeals moral taxes? Evidence from a field experiment on water conservation. *Louisiana State Department of Economics Working Papers Series*.

Bruvoll, A. and Nyborg, K. (2004). The cold shiver of not giving enough: on the social cost of recycling campaigns. *Land Economics*, 80(4):539–549.

Candari, C. J., Cylus, J., Nolte, E., Organization, W. H., et al. (2017). *Assessing the economic costs of unhealthy diets and low physical acitivity: an evidence review and proposed framework*. World Health Organization. Regional Office for Europe.

Capraro, V. and Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making*, 13(1):99–111.

Cardenas, J. C. (2011). Social norms and behavior in the local commons as seen through the lens of field experiments. *Environmental and Resource Economics*, 48:451–485.

Chaudhuri, A., Graziano, S., and Maitra, P. (2006). Social learning and norms in a public goods experiment with inter-generational advice. *The Review of Economic Studies*, 73(2):357–380.

Chen, X., Lupi, F., He, G., and Liu, J. (2009). Linking social norms to efficient conservation investment in payments for ecosystem services. *Proceedings of the National Academy of Sciences*, 106(28):11812–11817.

Crosetto, P. and Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, 47(1):31–65.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74.

DeJoy, D. M. (2005). Behavior change versus culture change: Divergent approaches to managing workplace safety. *Safety Science*, 43(2):105–129.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, 127(1):1–56.

Diekman, A. B. and Goodfriend, W. (2006). Rolling with the changes: A role congruity perspective on gender norms. *Psychology of Women Quarterly*, 30(4):369–383.

Du, N., Li, L., Lu, T., and Lu, X. (2020). Prosocial compliance in p2p lending: A natural field experiment. *Management Science*, 66(1):315–333.

Dulleck, U., Fooken, J., Newton, C., Ristl, A., Schaffner, M., and Torgler, B. (2016). Tax compliance and psychic costs: Behavioral experimental evidence using a physiological marker. *Journal of Public Economics*, 134:9–18.

D'attoma, J. W., Volintiru, C., and Malézieux, A. (2020). Gender, social value orientation, and tax compliance. *CESifo Economic Studies*, 66(3):265–284.

Eagly, A. H. (2009). The his and hers of prosocial behavior: an examination of the social psychology of gender. *American Psychologist*, 64(8):644.

Ehrlich, D. and Szech, N. (2022). How to start a grassroots movement. *CESifo working paper*.

Engelmann, D. and Strobel, M. (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics*, 3(3):241–260.

Eriksson, K., Strimling, P., and Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, 129:59–69.

Farrow, K., Grolleau, G., and Ibanez, L. (2017). Social norms and pro-environmental behavior: A review of the evidence. *Ecological Economics*, 140:1–13.

Fischbacher, U. and Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–556.

Galasso, V., Pons, V., Profeta, P., Becher, M., Brouard, S., and Foucault, M. (2020). Gender differences in covid-19 attitudes and behavior: Panel evidence from eight countries. *Proceedings of the National Academy of Sciences*, 117(44):27285–27291.

Galasso, V., Profeta, P., Foucault, M., and Pons, V. (2021). Covid-19 vaccine's gender paradox. *medRxiv*.

Gelfand, M., Li, R., Stamkou, E., Pieper, D., Denison, E., Fernandez, J., Choi, V., Chatman, J., Jackson, J., and Dimant, E. (2022). Persuading republicans and democrats to comply with mask wearing: An intervention tournament. *Journal of Experimental Social Psychology*, 101:104299.

Gruber, J. (1994). The incidence of mandated maternity benefits. *The American Economic Review*, pages 622–641.

Herweg, F. and Schmidt, K. M. (2022). How to regulate carbon emissions with climate-conscious consumers. *The Economic Journal*, 132(648):2992–3019.

Howe, P. D., Boldero, J., McNeill, I. M., Vargas-Sáenz, A., and Handmer, J. (2018). Increasing preparedness for wildfires by informing residents of their community's social norms. *Natural Hazards Review*, 19(2):04017029.

Kamas, L., Preston, A., and Baum, S. (2008). Altruism in individual and joint-giving decisions: What's gender got to do with it? *Feminist Economics*, 14(3):23–50.

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

Krupka, E. and Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, 30(3):307–320.

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.

Loschelder, D. D., Siepelmeyer, H., Fischer, D., and Rubel, J. A. (2019). Dynamic norms drive sustainable consumption: Norm-based nudging helps café customers to avoid disposable to-go-cups. *Journal of Economic Psychology*, 75:102146.

Mariotti, T., Schweizer, N., Szech, N., and von Wangenheim, J. (2023). Information nudges and self-control. *Management Science*, 69(4):2182–2197.

Müller, S. and Rau, H. A. (2021). Economic preferences and compliance in the social stress test of the covid-19 crisis. *Journal of Public Economics*, 194:104322.

Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., Adger, W. N., Arrow, K. J., Barrett, S., Carpenter, S., et al. (2016). Social norms as solutions. *Science*, 354(6308):42–43.

Olden, A. and Møen, J. (2022). The triple difference estimator. *The Econometrics Journal*, 25(3):531–553.

Roemer, J. E. (2015). Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, 127:45–57.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3):279–301.

Schnitzler, L., Jackson, L. J., Paulus, A. T., Roberts, T. E., and Evers, S. M. (2021). Intersectoral costs of sexually transmitted infections (stis) and hiv: a systematic review of cost-of-illness (coi) studies. *BMC Health Services WResearch*, 21(1):1–10.

Sent, E.-M. and van Staveren, I. (2019). A feminist review of behavioral economic research on gender differences. *Feminist Economics*, 25(2):1–35.

Serra-Garcia, M. and Szech, N. (2022). The (in)elasticity of moral ignorance. *Management Science*, 68(7):4815–4834.

Silva, J. M. (1983). The perceived legitimacy of rule violating behavior in sport. *Journal of Sport and Exercise Psychology*, 5(4):438–448.

Sloan, F. A., Eldred, L. M., and Xu, Y. (2014). The behavioral economics of drunk driving. *Journal of Health Economics*, 35:64–81.

Soutschek, A., Burke, C. J., Raja Beharelle, A., Schreiber, R., Weber, S. C., Karipidis, I. I., Ten Velden, J., Weber, B., Haker, H., Kalenscher, T., et al. (2017). The dopaminergic reward system underpins gender differences in social preferences. *Nature Human Behaviour*, 1(11):819–827.

Stok, F. M., De Ridder, D. T., De Vet, E., and De Wit, J. B. (2014). Don't tell me what i should do, but what others do: The influence of descriptive and injunctive peer norms on fruit consumption in adolescents. *British Journal of Health Psychology*, 19(1):52–64.

Thedell, T. D. (2016). Nudging toward safety progress. *Professional Safety*, 61(9):27.

Thøgersen, J. (2014). The mediated influences of perceived norms on pro-environmental behavior. *Revue d'Économie Politique*, (2):179–193.

Tibbetts, S. G. (1999). Differences between women and men regarding decisions to commit test cheating. *Research in Higher Education*, 40(3):323–342.

Tom, A. and Granié, M.-A. (2011). Gender differences in pedestrian rule compliance and visual search at signalized and unsignalized crossroads. *Accident Analysis & Prevention*, 43(5):1794–1801.

Ucho, A. and Gbande, A. (2012). Personality and gender differences in compliance with safety behaviour among factory workers of dangote cement company, gboko. *IFE PsychologIA: An International Journal*, 20(2):134–141.

Yamin, P., Fei, M., Lahlou, S., and Levy, S. (2019). Using social norms to change behavior and increase sustainability in the real world: A systematic review of the literature. *Sustainability*, 11(20):5847.

APPENDIX

A Details on theoretical background

Optimum for a self-interested individual that anticipates effect through group urn Consider a problem where drawing a black ball does not only have individual consequences but, in addition, contributes to the collective harm the total number of balls drawn $K = \sum_i k_i$, and the corresponding expected social harm is $(1-p)K\theta N$. The decision problem of individual i who decides how many balls k_i to draw and takes into account the consequence of the social harm only on her own payoff then takes the following form:

$$L_i(k_i) = \begin{cases} \gamma k_i - \frac{1}{N}(1-p)K\theta & \text{with probability } p^{k_i} \\ -\frac{\theta}{N}(1-p) & \text{with probability } 1-p^{k_i} \end{cases}, \tag{5}$$

where we omit the individual payoff of θ if the social harm is zero and focus on payoff components that depend on actions. Note first that the non-cooperative solution to this problem is the same as above as long as the individuals take the total number of balls K as beyond their control, which will be a typical result if an individual's share is sufficiently small.

In the implementation of the online experiment, groups consist of 100 individuals, which may be small enough for individuals to perceive their own influence on the group outcome, which makes the decision situation strategic. To take the strategic interaction within groups into account, we model $K = k_i + K_{-i}$, where $K_{-i} = \sum_{j \neq i} k_j$ is the sum of balls drawn by all other group members but individual i and everyone simultaneously decide about their number of balls in a non-cooperative game. Still, individuals are unconcerned with their effect on the payoff of others but care only about their own expected payoff, which implies that they only take into account $\frac{1}{N} = 1\%$ of the social harm of a black ball drawn.

$$L_i(k_i, K_{-i}) = \gamma k_i p^{k_i} + \theta - \frac{\theta}{N}(1-p)(k_i + K_{-i}). \tag{6}$$

For the parameters of the experiment, this expression is maximized at 19.23. Thus, respecting the integer constraint, a money-maximizing risk-neutral individual who takes into account their personal consequences from social harm implied through their own behavior but does not care about the externalities imposed on others, will still draw 19 balls, suggesting that N=100 is already a large group leading to almost complete neglect of the social urn. Again, risk aversion will lead to a smaller number of balls to be individually optimal. This leads us to our first hypothesis about behavior in the Baseline treatment.

Optimum for socially minded individuals Let us now consider how the decision of a socially concerned individual is affected by her potential contribution to the social harm that they inflict not only on themselves but also on others. There are different ways to conceptualize social concern. First, an individual with Kantian preferences will decide that—if taken by everyone—will maximize the group's welfare.[28] Thus, it would decide to draw 6 balls. Solving the problem correctly is complicated though and individuals may also have doubts about the behavior of the other group members. Therefore, we complement the Kantian perspective with a more behavioral approach that does not require that individuals fully anticipate and internalize the expected utility consequences of their and others' actions on each other. Instead, we suppose that individuals experience a cold shiver from any ball they draw as this might be a black ball that will not only hurt themselves but also all others. This approach can be understood as the flip side of a "warm glow"-model of impure altruism, where individuals have a positive marginal from giving. We model the cold shiver as being related to the quantity of the harmful action like a standard warm glow model would (Andreoni, 1990), whereas Bruvoll and Nyborg (2004); Brekke et al. (2010) discuss cold shiver as something that affects duty-oriented individuals when their behavior falls short of the perceived duty or norm.

We focus on the risk-neutral case as the qualitative effect of risk aversion will again decrease the optimal number of balls and abstract from an individual's anticipated effect on their own payoff through the group urn as this is negligible above. Then, the expected utility from k_i balls is given by

$$U(k_i) = \gamma k_i p^{k_i} + \theta - \frac{\theta}{N}(1-p)(k_i + K_{-i}) - \mu k_i, \tag{7}$$

where the cold shiver is parameterized by $\mu > 0$. Note that the cold shiver does not depend on whether or not a ball turns out to be black. Instead, disutility arises from doing something that may harm the group. We assume that $\gamma \geq \mu$, that is the private utility from a white ball is higher than the disutility from the cold shiver when drawing a ball.

The resulting optimal number of balls k^* is identical to the self-interested solution for $\mu = 0$ but is decreasing in μ, so that socially minded individuals will draw fewer than 19 balls and their number of draws will be smaller the stronger the cold shiver that they experience.[29]

[28] This follows from Kant's imperative to "act only on the maxim that you would at the same time will to be a universal law". For an economic treatment of how such "Kantian optimization" will help resolve social dilemma situations, see for instance Roemer (2015). Alger and Weibull (2013) model individuals who follow Kant's imperative as *homo Kantiensis* who directly attempts to maximize the outcome that resulted if everyone did the same as herself.

[29] Using the first order condition for utility maximization, we obtain the solution $k^* = \frac{1-\text{ProductLog}(e^{\frac{\mu}{\gamma}})}{\log(p)}$. *ProductLog* is the Lambert W function and cannot be expressed in terms of elementary functions.

B Peer information in the Rec&Info treatment

Table 6: Empirical information presented in theRec&Info treatment.

Subjects	**A**	**B**	**C**	**D**
Round 2	14	20	6	20
Round 3	8	6	6	18
Round 4	14	6	19	7
Round 5	10	10	15	14
Round 6	14	10	17	13
Round 7	9	10	15	8
Round 8	11	10	15	13
Round 9	18	10	19	13
Round 10	11	10	17	13

C Additional results

C.1 The effect of information in only male, only female subsamples

In separate analyses for men and women, we investigate the effect of negative peer information on ball drawing behavior. At first sight, it seems to be the case that women react more to the Rec&Info treatment: the coefficient on the interaction of *Rec&Info* and *Post* is larger in absolute size than the corresponding coefficient for males. It must be noted, however, that in the Rec treatment, which serves as the comparison group here, female participants are more likely to draw at most six balls in rounds 2 to 10 than men. This can be seen by comparing the coefficient on *Post* between men and women. The supposedly stronger effect of negative information on female participants compensates for their higher compliance in the Rec treatment. The triple difference estimation in the main text (see Table 3) that takes into account both these differences simultaneously shows that the differentially negative for women is not statistically different from zero.

Table 7: Difference-in-difference results for the effect of information on compliance estimated separately for men and women.

	(1)	(2)	(3)	(4)
	Max6	Max6	Max6	Max6
	Male only		Female only	
Post	0.5285***	0.5290***	0.5847***	0.5896***
	(0.0205)	(0.0211)	(0.0208)	(0.0213)
Rec&Info	-0.0322	-0.0419	0.0291	0.0306
	(0.0398)	(0.0435)	(0.0396)	(0.0443)
Rec&Info×Post	-0.0960**	-0.1005**	-0.1698***	-0.1714***
	(0.0303)	(0.0311)	(0.0307)	(0.0316)
Constant	0.1791***	-0.0835	0.1483***	0.1607
	(0.0270)	(0.1687)	(0.0269)	(0.1594)
Controls	no	yes	no	yes
Observations	5480	5290	5380	5030
R^2 (overall)	0.1034	0.1283	0.1165	0.1361

Notes: GLS random effects model with robust standard errors. The dependent variable *Max6* is an indicator for drawing no more than six balls in a given round. Columns 1 and 3 use observations from the Baseline and Rec treatment. Columns 2 and 4 use observations from the Rec and Rec&Info treatment. Standard errors in parentheses. $^{\circ}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Difference-in-difference results for the effect of information on draws estimated separately for men and women.

	(1) Draws	(2) Draws	(3) Draws	(4) Draws
	Male only		Female only	
Post	-10.9944***	-11.0339***	-10.7667***	-10.9290***
	(0.2993)	(0.3048)	(0.2793)	(0.2864)
Rec&Info	-0.3302	-0.3682	1.0916°	1.3612*
	(0.6218)	(0.6799)	(0.5806)	(0.6519)
Rec&Info×Post	1.8555***	2.0111***	0.1726	0.1644
	(0.4414)	(0.4488)	(0.4113)	(0.4245)
Constant	18.3818***	22.5322***	17.7552***	17.0622***
	(0.4217)	(2.7613)	(0.3942)	(2.4731)
Controls	no	yes	no	yes
Observations	5480	5290	5380	5030
R^2 (overall)	0.1570	0.1857	0.1918	0.2353

Notes: GLS random effects model with robust standard errors. Draws is the number of balls in a given round. Columns 1 and 3 use observations from the Baseline and Rec treatment. Columns 2 and 4 use observations from the Rec and Rec&Info treatment. Standard errors in parentheses. ° $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

D Representativeness

The descriptive details and representativeness comparison for Experiment 1 and 2 can be found in Table 9 and 10.

Table 9: Representativeness check for Experiment 1.

Criteria	Baseline	Rec	Rec&Info	Germany
Sample Size	514	586	500	
Male	52%	51%	50%	51%
Female	48%	49%	50%	49%
Lower	34%	34%	30%	29%
Middle	30%	32%	31%	34%
High Edu	37%	34%	39%	38%
18-29	19%	20%	20%	20%
30-39	20%	18%	19%	19%
40-49	18%	18%	18%	18%
50-59	25%	24%	24%	24%
60-69	17%	19%	18%	18%
Baden Württemberg	14%	13%	13%	13%
Bayern	16%	16%	15%	16%
Berlin	3%	4%	4%	4%
Brandenburg	4%	2%	3%	3%
Bremen	1%	1%	1%	1%
Hamburg	2%	2%	2%	2%
Hessen	7%	8%	7%	7%
Mecklenburg-Vorpommern	3%	2%	2%	2%
Niedersachsen	9%	11%	10%	10%
Nordrhein Westfalen	23%	20%	24%	22%
Rheinland-Pfalz	6%	5%	4%	5%
Saarland	1%	1%	1%	1%
Sachsen	4%	5%	5%	5%
Sachsen-Anhalt	2%	2%	3%	3%
Schleswig-Holstein	3%	4%	3%	3%
Thüringen	3%	3%	3%	3%

Table 10: Representativeness check for Experiment 2.

Criteria	PB	NE	EE	KW	Germany
Sample Size	133	126	136	127	
Male	53%	48%	54%	50%	51%
Female	47%	52%	46%	50%	49%
Lower	29%	30%	27%	29%	29%
Middle	29%	31%	38%	32%	34%
High Edu	41%	39%	35%	39%	38%
18-29	24%	17%	20%	19%	20%
30-39	16%	17%	21%	15%	19%
40-49	15%	25%	15%	20%	18%
50-59	23%	24%	29%	22%	24%
60-69	23%	17%	16%	24%	18%
Baden Württemberg	10%	14%	11%	18%	13%
Bayern	23%	17%	9%	16%	16%
Berlin	5%	2%	5%	2%	4%
Brandenburg	4%	2%	6%	1%	3%
Bremen	0%	2%	1%	1%	1%
Hamburg	3%	1%	2%	2%	2%
Hessen	8%	6%	5%	9%	7%
Mecklenburg-Vorpommern	2%	3%	3%	0%	2%
Niedersachsen	7%	13%	13%	8%	10%
Nordrhein Westfalen	17%	25%	24%	24%	22%
Rheinland-Pfalz	7%	2%	6%	6%	5%
Saarland	1%	2%	0%	2%	1%
Sachsen	5%	6%	4%	5%	5%
Sachsen-Anhalt	3%	2%	3%	2%	3%
Schleswig-Holstein	4%	4%	3%	2%	3%
Thüringen	3%	0%	5%	5%	3%

E Experiment 2

E.1 P-values of the KW Method treatment of Experiment 2 by gender

Table 11: Two-sample Wilcoxon rank-sum (Mann-Whitney) test results.

Number of Draws	P-VALUE	Z-SCORE
0	0.2953	1.047
1	0.8985	-0.128
2	0.8389	-0.203
3	0.8501	-0.189
4	0.9525	-0.060
5	0.4904	-0.690
6	0.4874	-0.694
7	0.7146	0.366
8	0.3946	0.851
9	0.2470	1.158
10	0.3422	0.950
11	0.2852	1.069
12	0.3277	0.979
13	0.3464	0.941
14	0.4231	0.801
15	0.7201	0.358
16-20	0.8417	0.200
21-25	0.6167	0.500
26-30	0.6037	0.519
31-more	0.4615	0.736

Notes: The table shows the p-values and z-scores of the Wilcoxon rank-sum test for each number of draws by gender.

E.2 Distribution of draws

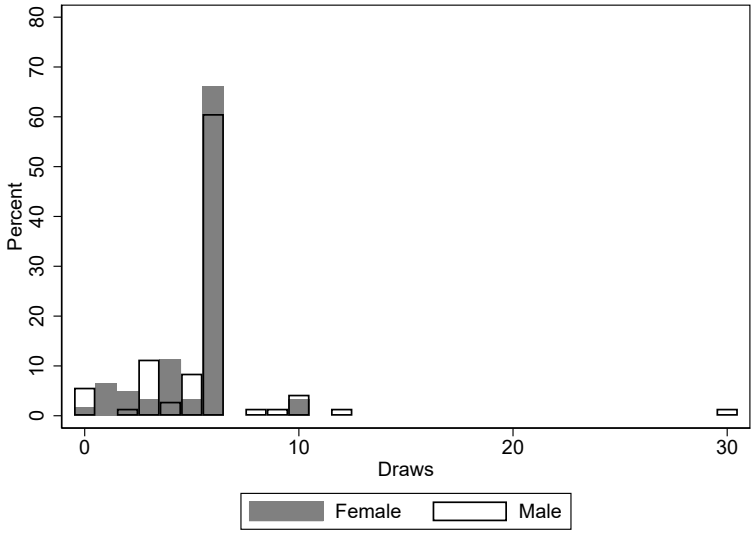


Figure 5: Personal beliefs by gender.

Notes: The figure illustrates the distribution of personal beliefs regarding the *Draws*.

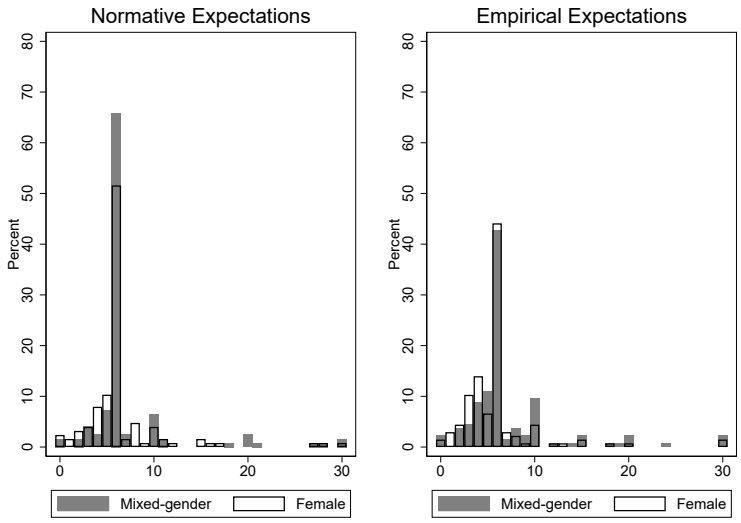
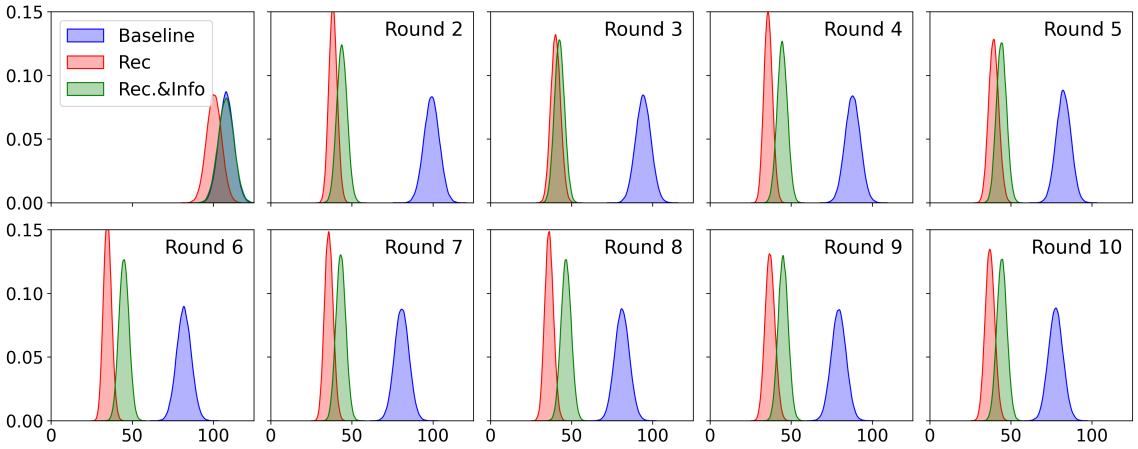


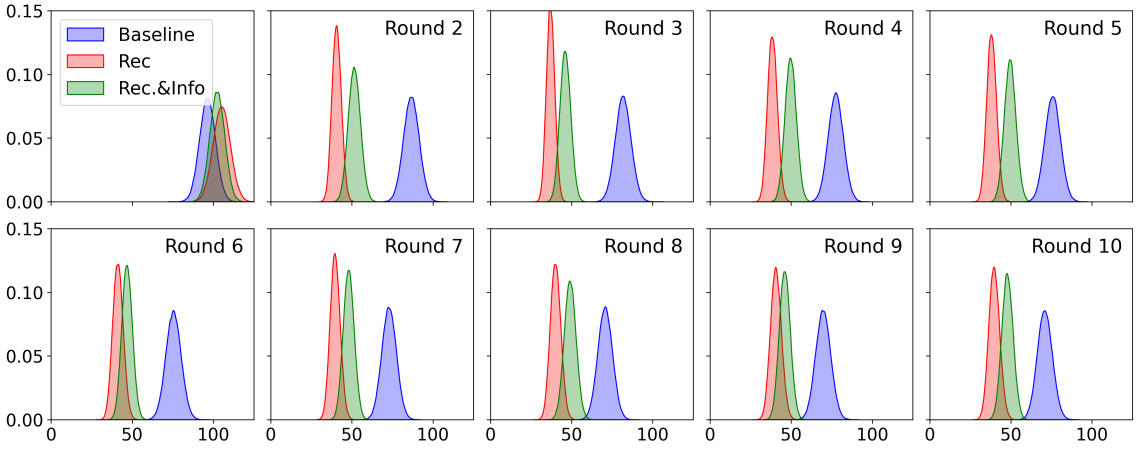
Figure 6: Expectations.

Notes: The figure illustrates the distribution of expectations regarding the *Draws*. The left panel represents normative expectations while the right panel focuses on empirical expectations.

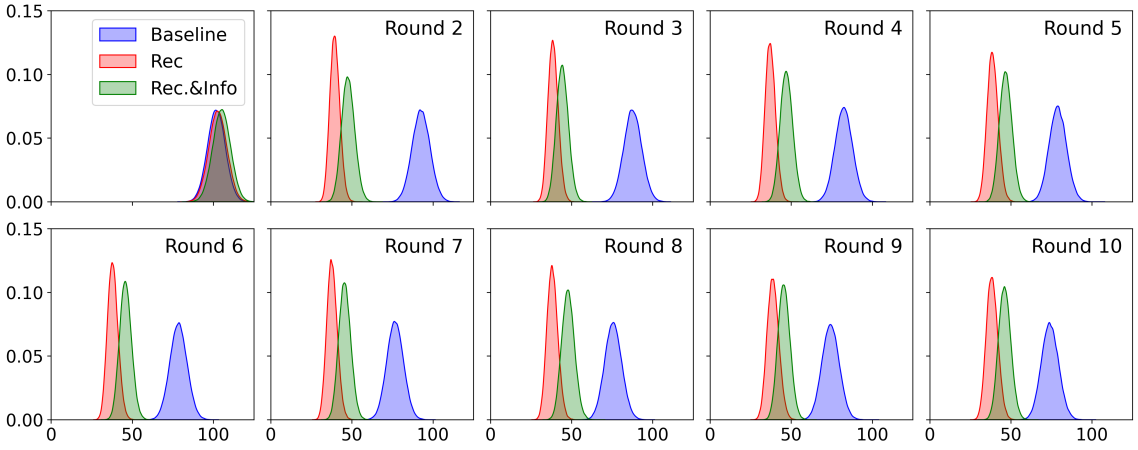
F Welfare implications



(a) Female participants



(b) Male participants



(c) All participants

Figure 7: Simulation results for social harm (rounds 1-10).

Notes: The figure shows the simulated distribution of total black balls drawn in each round per group of 100 participants. Panel (a) represents groups of 100 randomly drawn female participants, panel (b) represents groups of 100 randomly drawn male participants, and panel (c) represents groups of 100 randomly drawn participants of both genders.

G Screen shots of main decision screens in Experiment 1 and full instructions

Figure 8: Experiment 1: Decision screens in Round 1 (all treatments) and Round 2 to 10 in Baseline.

Bitte beachten Sie:

Um den Bonus für alle Teilnehmer zu maximieren, empfehlen wir Ihnen **nicht mehr als 6 Bälle** zu ziehen.

(a) Information shown only in treatment Rec.

Bitte beachten Sie:

Um den Bonus für alle Teilnehmer zu maximieren, empfehlen wir Ihnen **nicht mehr als 6 Bälle** zu ziehen.

In den folgenden Runden werden Sie außerdem **die Summe der gezogenen Bälle von 4 Teilnehmern** (Spieler A, B, C und D) aus der gleichen Runde angezeigt bekommen. Diese Werte basieren auf einem früheren, identischen Experiment.

(b) Information shown only in treatment Rec&Info.

Figure 9: Experiment 1: Information shown before Round 2 in (a) Rec and (b) Rec&Info.

(a) Decision screens in Round 2 to 10 in treatment Rec.

(b) Decision screens in Round 2 to 10 in treatment Rec&Info.

Figure 10: Experiment 1: Decision screens in Round 2 to 10 in (a) Rec and (b) Rec&Info.

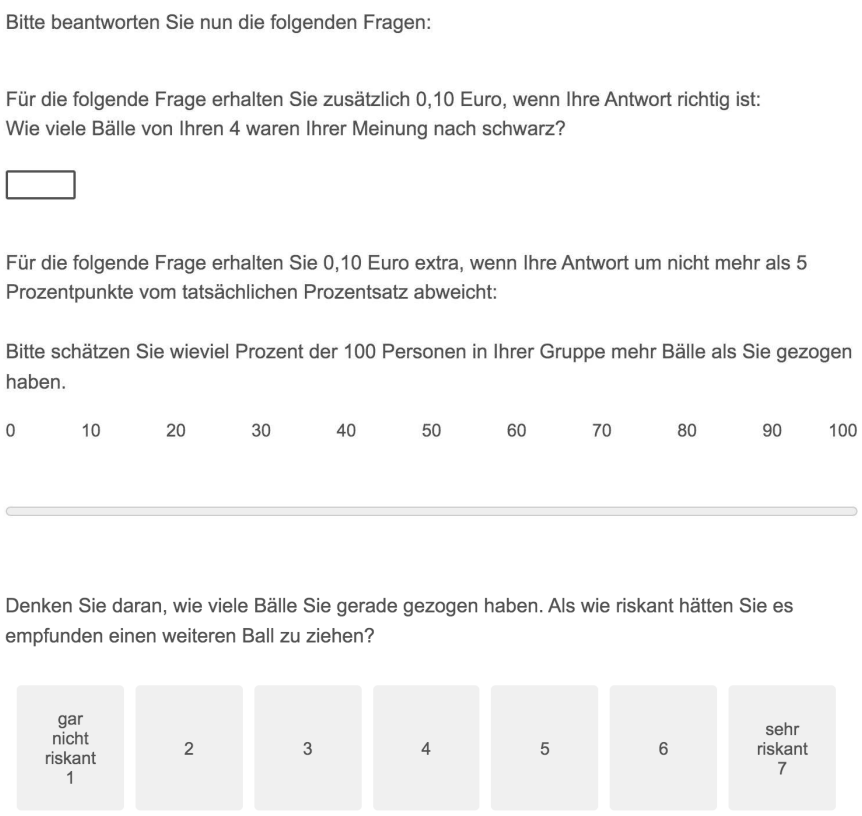


Figure 11: Experiment 1: Elicitation of risk perceptions after each round in all treatments

Instructions

This document is an machine-generated translation of the original German instructions.

Experiment 1 - Baseline, Rec, and Rec&Info

Welcome!

This study aims to help understand how people make decisions. It is funded by various research institutions.

The entire study will take approximately 20 minutes. Please ensure you have this time available.

You must be at least 18 years old to participate in this study.

Your responses will be treated anonymously and confidentially.

In addition to your regular payment through [provider], you may earn a bonus amount, which depends on your decisions in the study. An explanation of how to earn bonuses will follow in the study.

Your participation in the study is voluntary, and you can withdraw at any time. Please note that in such cases, no compensation will be provided.

For any issues or feedback, contact [email of provider]. (For content-related questions: [email of research team].)

I have read the above information. By clicking the red button, I agree to the terms of participation in this study.

We are glad that you would like to participate in this study.
This is an experiment on decision-making. The bonus payment you receive depends on the decisions made by you and other participants.

The rules of the experiment will be explained in detail on the following pages. Afterward, you can try navigating a test screen before the actual experiment begins. Please read the rules carefully.

Please click continue to proceed.

This experiment consists of 10 independent decision rounds. Each round lasts 40 seconds.

In each round, you have three ways to generate bonus payments: an individual bonus, a group bonus, and a bonus from two estimation questions. At the end of the experiment, one of the 10 rounds will be randomly selected. Your final bonus payment will result from the sum of your bonuses in the selected round.

Your personal bonus and the bonus for estimation questions are based on your individual decisions.

Your group bonus is calculated based on your decisions and those of 99 randomly selected participants at the end of the experiment.

At the end of each round, you will only be informed about your personal bonus for that specific round.

Please click continue to proceed.

Experiment Procedure and Bonus Payments: Please Read Carefully!

In each round, you will see an urn on the screen. This urn contains **always 100 balls**. In each round, you have 40 seconds to draw balls from the urn. You decide how many you want to draw. **Of the 100 balls, 95 are white and 5 are black.** Each ball you draw from the urn is immediately replaced with a ball of the same color. At the end of each round, you will be informed about the number of white and black balls you have drawn.

Personal Bonus: For each **white ball**, you receive **€0.10 if you draw only white balls**. Your personal bonus is then the number of white balls multiplied by €0.10. However, if you draw **at least one black ball**, you lose all your white balls. Your personal round bonus is then **€0**.

Group Bonus: In each round, you also have the chance to receive a group bonus. The total **group bonus for each round is €100**. This amount is shared with 99 other participants; this corresponds to **€1 per participant**. If a group member draws **1 black ball from their individual urn, the group bonus decreases by €1**. For example, if the group members draw a total of only 5 black balls in the round, each group member receives €0.95. If a total of 100 black balls are drawn in a round, the group bonus is €0.

Click the red button to familiarize yourself with the process in a test round.

Test Round

Time remaining: _____
Draws (Number of Balls): _____

[see screenshots for a visual impression of the screen]

Please Remember: In each of the 10 rounds, the following applies:

- The urn contains **always 100 balls**, of which **95 are white** and **5 are black**.
- You can receive €0.10 per draw as a personal bonus if you draw a white ball.
- Your personal bonus is €0 if you draw a black ball.
- The group bonus decreases by €1 for each black ball drawn.

At the end of the study, one round will be randomly selected, and your final bonus payment will be calculated.

Please click the red button to start the first round.

Round 1

Time remaining: _____
Draws (Number of Balls): _____

[see screenshots for a visual impression of the decision screens]

In this round, you have drawn a total of _____ balls.

Please now answer the following questions:

For the following question, you will receive an additional €0.10 if your answer is correct:

How many of your _____ drawn balls do you think were black?

For the following question, you will receive an additional €0.10 if your answer deviates by no more than 5 percentage points from the actual percentage:

Please estimate what percentage of the 100 participants in your group drew more balls than you.

[Slider running from 0 to 100]

Think about how many balls you just drew.

How risky would you have considered it to draw another ball?

[Answer options on 7-point Likert scale: Not risky at all 1 – Very risky 7]

Round 1 Summary [identical apart from round number for rounds 2 to 9]

Number of white balls this round: _____
Number of black balls this round: _____

[if participant had drawn any black ball] Since you did not draw any black balls, your personal bonus is _____ Euros.

Your draw had no impact on the group bonus.

[alternative text if participant had drawn at least one black ball] For your _____ white balls, you would receive _____ Euros. Since you drew at least one black ball, your personal bonus for this round is €0. You are responsible for reducing the group bonus by _____ Euros.

Please click the red button to proceed to the next round. You will have 40 seconds again.

[additional screen between Round 1 and Round 2 for REC]

Please Note:

To maximize the bonus for all participants, we recommend drawing **no more than 6 balls.**

[additional screen between Round 1 and Round 2 for REC&INFO]

Please Note:

To maximize the bonus for all participants, we recommend drawing **no more than 6 balls.**

In the following rounds, you will also see **the total number of balls drawn by 4 participants** (players A, B, C, and D) from the same round. These values are based on a previous, identical experiment.

Round 2 [identical apart from round number for rounds 3 to 10] *[see screenshots for a visual impression of the decision screens]*

[only in REC]

Recommended: no more than 6 balls

[only in REC&INFO]

Recommended: no more than 6 balls
Number of balls drawn by player A:_____; B:_____; C:_____; D:_____

[for all]

Time remaining: _____
Draws (Number of Balls): _____

In this round, you have drawn a total of _____ balls.

Please now answer the following questions:

For the following question, you will receive an additional €0.10 if your answer is correct:

How many of your _____ drawn balls do you think were black?

For the following question, you will receive an additional €0.10 if your answer deviates by no more than 5 percentage points from the actual percentage:

Please estimate what percentage of the 100 participants in your group drew more balls than you.

[Slider running from 0 to 100]

Think about how many balls you just drew.

How risky would you have considered it to draw another ball?

[Answer options on 7-point Likert scale: Not risky at all 1 – Very risky 7]

[Attention Check included after Round 5]

What do you draw from the urn?

o Balls

o Bananas

Round 10 Summary

Number of white balls this round: _____
Number of black balls this round: _____

[if participant had drawn any black ball] Since you did not draw any black balls, your personal bonus is _____ Euros.

Your draw had no impact on the group bonus.

[alternative text if participant had drawn at least one black ball] For your _____ white balls, you would receive _____ Euros. Since you drew at least one black ball, your personal bonus for this round is €0. You are responsible for reducing the group bonus by _____ Euros.

Please click the red button to proceed to the final questions.

Winnings

The round randomly selected for payment is round _____. You receive a personal bonus of _____ Euros from this round. Your bonus for correctly answered estimation questions as well as the group bonus from the same round will be calculated after the study is completed. All amounts will be paid out along with your personal bonus.

Please click the red button to proceed to the questionnaire portion of this study. Note that you can only receive payment if you also complete the questionnaire portion.

Please answer the following questions as accurately as possible. Your answers are of great value to this scientific study. Of course, your responses will remain anonymous and be treated with absolute confidentiality. We thank you in advance for your cooperation.

Are you generally a risk-tolerant person, or do you try to avoid risks?

Please check a box on the scale, where 0 means "Not at all risk-tolerant" and 10 means "Very risk-tolerant." With the values inbetween you you can grade your assessment.

[Radio buttons with 11-point Likert scale]

Please indicate your agreement with the following statements:

[for these statements, answers were given through radio buttons with 11-point Likert scale, ranging from *Strongly disagree* to *Strongly agree*]

Generally, people can be trusted.

Nowadays, you cannot rely on anyone anymore.

When dealing with strangers, it is better to be cautious before trusting them.

You can trust the government in this country.

You can trust scientific institutions.

How often did you in the last week:

[for these statements, answers were given through radio buttons with 11-point Likert scale, ranging from *Never* to *Always*]

Wash your hands with soap for at least 20 seconds?

Wear a face mask indoors when recommended?

Wear a face mask outdoors when recommended?

Maintain a distance of at least 150 cm from people not in your household?

Have you been vaccinated against COVID-19? [one of the following had to be ticked]

o Yes, I have already received at least one dose.

o No, but I already have an appointment.

o No, but I plan to schedule an appointment.

o No, I am not eligible for vaccination.

o No, I do not wish to be vaccinated.

o No, I am unsure if I want to be vaccinated.

How stable was your internet during this study?

[radio buttons with 11-point Likert scale, ranging from *Not at all stable* to *Completely stable*]

[The study at this stage included a demographic questionnaire at the end, which is omitted here for brevity]

How important was it ...

[for these statements, answers were given through radio buttons with 11-point Likert scale, ranging from *Not important at all* to *Very important*]

... **for you** to maximize your personal bonus during the experiment?

... **for you** that your behavior did not reduce the group bonus?

... in your opinion, **for the majority of your group members** to maximize their personal bonuses during the experiment?

... in your opinion, **for the majority of your group members** to not reduce the group bonus during the experiment?

End of Experiment 1

Experiment 2

Welcome!

This study aims to help understand how people make decisions. It is funded by various research institutions.

The entire study will take approximately 10 minutes. Please ensure you have this time available.

You must be at least 18 years old to participate in this study.

Your responses will be treated anonymously and confidentially.

Please note: In this study, you will receive a guaranteed payment of €0.50 in mingle points if you correctly answer two comprehension questions about the study content and if you complete the study to the end. The comprehension questions will be presented to you after the study instructions. Only if you correctly answer both comprehension questions can you continue with the study and thus receive payment.

Additionally, you may earn a bonus amount, which depends on your decisions in the study. An explanation of this will follow in the study.

Your participation in the study is voluntary, and you can withdraw at any time. Please note that in such cases, no payment will be provided.

If you have questions about this study, please contact [name and email of contact to research team].

I have read the above information. By clicking the red button to proceed, I agree to the terms of participation in this study.

We are glad that you would like to participate in this study.

This is an experiment on decision-making. On the following pages, the rules of the experiment will be explained in detail. Please read the rules carefully and remember that after the study briefing, you must correctly answer two comprehension questions to proceed with the study.

Please click continue to proceed.

Below we describe a scenario from a previous experiment. The participants' decisions were implemented as described, and the resulting bonus payments were paid to the participants.

In this experiment, participants had to decide how many balls they wanted to draw from an urn. Each of the 10 decision rounds lasted 40 seconds.

In a decision round, participants could earn two types of bonuses: a personal bonus and a group bonus.

The personal bonus was based on an individual participant's decision.

The group bonus was calculated based on the respective individual decision and the decisions of 99 other randomly selected participants.

The more balls a participant drew, the lower their chances of actually receiving the personal and group bonuses.

Please click continue to proceed.

Experiment Procedure and Bonus Payments - Please Read Carefully!

Participants received the following instructions:

"In each decision round, you will see an urn on the screen. This urn contains **always 100 balls**. In each round, you have 40 seconds to draw balls from the urn. You decide how many you want to draw. **Of the 100 balls, 95 are white and 5 are black.** Each ball you draw from the urn is immediately replaced with a ball of the same color. At the end of each round, you will be informed about the number of white and black balls you have drawn.

Personal Bonus: For each **white ball**, you receive **€0.10 if you draw only white balls**. Your personal bonus is then the number of white balls multiplied by €0.10. However, if you draw **at least one black ball**, you lose all your white balls. Your personal round bonus is then **€0**.

Group Bonus: In each round, you also have the chance to receive a group bonus. The total **group bonus for each round is €100**. This amount is shared with 99 other participants; this corresponds to **€1 per participant**. If a group member draws **1 black ball from their individual urn, the group bonus decreases by €1**. For example, if the group members draw a total of only 5 black balls in the round, each group member receives €0.95. If a total of 100 black balls are drawn in a round, the group bonus is €0."

Participants were also advised to draw no more than 6 balls to maximize the total payout for all participants. Please click continue to see an example of what the decision screen looked like for participants.

Participants had 40 seconds to draw balls from the urn. At the beginning of the experiment, participants saw the following screen:

The number of balls drawn was a personal decision of the participants. Participants were not bound to the recommendation to draw 6 balls.

At the end of each decision round, each participant was shown feedback on how many black balls they had drawn and how this had affected the group bonus.

Please now answer the following two comprehension questions to proceed with the study. If you do not answer the questions correctly, you will automatically be redirected to the panel provider and cannot continue with this study.

In the study just described, participants were recommended a maximum number of balls to draw. How many balls were participants recommended to draw? [0 balls / 6 balls / 10 balls]

Please recall the study just described. What happens if a participant draws a black ball? [The participant earns an additional €1. / The group bonus increases by €1 in the round. / The participant loses their personal bonus for the round, and the group bonus decreases by €1 in the round.]

[Treatment Personal beliefs]

Think about the experiment just described. Please indicate how you believe one should behave in a decision round.

One should draw the following number of balls:

[answer from a menu with all integers from 0 to 30]

[Treatment Normative Expectations]

In a separate survey, we asked people how they believe one should behave in the experiment described above.

Please indicate what you think people answered in this survey. You will receive an additional payment of €0.10 for each question in which your estimate is correct.

Indicate which statement you think is correct: The majority of all respondents answered in the survey that one should draw the following number of balls:

[answer from a menu with all integers from 0 to 30]

Indicate now which statement you think is correct: The majority of all female respondents answered in the survey that one should draw the following number of balls:

[answer from a menu with all integers from 0 to 30]

[Treatment Empirical Expectations]

Please indicate how you believe the participants in the described experiment behaved. These participants were selected to be representative of the German population. You will receive an additional payment of €0.10 for each question in which your estimate is correct.

Imagine all participants in the experiment. What do you think the majority of participants did? They drew the following number of balls:

[answer from a menu with all integers from 0 to 30]

Now imagine all female participants in the experiment. What do you think the majority of participants did? They drew the following number of balls:

[answer from menu with all integers from 0 to 30]

[Treatment KW Method]

Please rate in each row how appropriate the respective behavior is. Base your judgment on what you believe other respondents in this survey most frequently selected in this row. All respondents in this survey were given the exact same task as you.

You can receive an additional payout if you identify the responses most frequently selected by other respondents. For two randomly selected rows, you will receive an additional payout of €0.10 each if you selected the response most frequently chosen in that row.

Please now select for each row which of the response options other respondents most frequently selected in this row:

Number of Balls	Very Inappropriate	Rather Inappropriate	Rather Appropriate	Very Appropriate
0 Balls	o	o	o	o
1 Ball	o	o	o	o
...	...	...	...	...
15 Balls	o	o	o	o
16-20 Balls	o	o	o	o
21-25 Balls	o	o	o	o
26-30 Balls	o	o	o	o
31 or more Balls	o	o	o	o

[The study at this stage included a demographic questionnaire as Experiment 1 at the end, which is omitted here for brevity]

End of Experiment 2