Experimental Evidence on Individual and Collective Decision-Making in Ethical Dilemmas and Unfair Contests

Universität Hamburg
Fakultät Wirtschafts- und Sozialwissenschaften

Kumulative Dissertation

Zur Erlangung der Würde eines Doktors
der Wirtschafts- und Sozialwissenschaften
„Dr. rer. pol.“
(gemäß der PromO vom 18. Januar 2017)

vorgelegt von

Timo Promann
aus Hamburg, Deutschland

Hamburg, 22. April 2024

Vorsitzende: Prof. Dr. Anke Gerber

Erstgutachter: Prof. Dr. Gerd Mühlheußer

Zweitgutachter: Prof. Dr. Andreas Lange

Drittgutachterin: Prof. Dr. Dr. Lydia Mechtenberg

Datum der Disputation: 05.08.2024

Acknowledgements

First and foremost I want to express my gratitude to Gerd Mühlheußer. Thank you, for giving me so much freedom where I work, when I work and how I work. Thank you for always supporting me with kind words, with valuable feedback and, in general, with your time. Thank you for guiding me on this path to knowledge and wisdom.

I am very thankful for the mental support of Niklas Wallmeier. I could always complain about dysfunctional code and felt always more calm and collected after you listened to my rants.

I strongly appreciate my fellowship in the graduate program "Collective Decision-Making" and am very grateful for seemingly unlimited financial support and very constructive feedback by various attendees of our seminars. Special thanks to Anke Gerber for making all this possible and working tirelessly to solve each and every problem.

I was blessed with fantastic co-authors and I enjoyed working with all of you, Christoph Huber, Christos Litsios, Jan-Patrick Mayer, Gerd Mühlheußer, Annika Nieper, Andreas Roider, Eugen Tereschenko and Niklas Wallmeier.

Before joining the graduate program, I was financially supported by the WiSo Graduate School, the GfeW (Gemeinschaft für experimentelle Wirtschaftsforschung) and, once again, Gerd Mühlheußer.

I am overwhelmingly grateful for my social network, that was not directly connected to this work, but did always accept and appreciate me for who I am without any second thought.

In no comparison stands the contribution of my parents, Mechthild Thier and Reiner Promann. With all my heart, I want to thank you for raising me the way you did, giving me every possible opportunity to shape my life and always supporting me with your best intentions in heart.

Lastly, I am incredibly thankful for Hanna Hopf. Thank you for sharing my life, for reading my mind, for making me laugh and for causing my happiness every single day.

Contents

1 Introduction **1**

2 Unfair Contests - Experimental Evidence for Individuals and Teams **8**

2.1 Introduction 8

2.2 Literature 9

2.3 Hypotheses 11

2.4 Experimental Design 12

2.4.1 Experimental Treatments and Procedure 13

2.4.2 Experimental Implementation 16

2.5 Results 18

2.5.1 Main Results 19

2.5.2 Enjoyment 22

2.5.3 Fairness Perception 23

2.5.4 Robustness Checks 25

WiSo-Lab vs. Prolific Teams 25

Pre-Study on Prolific 26

Answer Ranks 26

2.6 Discussion 28

2.7 Conclusion 29

2.A Additional Tables and Figures 30

2.B Data Graphs for Robustness Checks 31

2.B.1 WiSo-Lab vs. Prolific Teams 31

2.B.2 Pre-Study on Prolific 32

2.B.3 Answer Ranks 33

2.C *Family Feud* Game Questions and Answer Distributions 35

2.D Experimental Instructions 38

3 Honesty of groups: Effects of size and gender composition **46**

3.1 Introduction 46

3.2 Related literature 51

3.3 Experiment 53

3.3.1 Design 53

3.3.2 Implementation 54

3.4 Results 57

3.4.1 The effect of group size on lying behavior 57

3.4.2 The effect of the group gender composition on lying behavior 59

3.4.3 Decision times and talking times 62

3.4.4 The effect of individual honesty on group behavior 64

3.4.5 Robustness 66

3.4.6 Groups with no agreement 68

3.5 Conclusion 69

3.A Instructions 71

3.B Additional figures and tables 86

4 ChaTree - video chat integration in oTree 89

4.1 Introduction 89

4.2 Features 90

4.3 Requirements and Implementation 91

5 On Social Norms and Observability in (Dis)honest Behavior 94

5.1 Introduction 94

5.2 Related Literature and Hypotheses 97

5.3 Experimental Design and Implementation 100

5.3.1 Experimental Treatments and Procedure 101

5.3.2 Experimental Implementation 104

5.4 Results 105

5.4.1 Social Norm Nudge 105

5.4.2 Observability 108

5.4.3 Honest Behavior 109

5.4.4 Robustness: Separating Social Norms from Behavior 112

5.5 Discussion and Conclusion 114

5.A Additional Tables and Figures 115

5.B Experimental Instructions 120

List of Figures 141

List of Tables 143

Bibliography 144

Chapter 1.

Introduction

Each and every day is shaped by decision-making. Optimally, each individual can structure their own day solely based on their own decisions. However, as soon as we enter work environments, the minority of decisions affects only the person in charge of deciding. Firms decide whether to hire new personnel, managers distribute tasks to their team members and a co-worker may decide to show up sick to work, infecting the whole division.

As decision-making is omnipresent, an investigation of reasons for certain decision behavior can only focus on very small aspects thereof. To gain a decent understanding of a behavioral mechanism behind specific decisions, many of those tiny aspects have to be assembled. To contribute at least some of these small puzzle pieces, this dissertation examines decision-making of individuals and teams, that are confronted with ethical dilemmas and unfair contests. With this contribution, more light is shed on the performance of individuals and teams that enter an unbalanced competition (Chapter 2). In addition, effects of group size and group gender composition on honesty behavior of groups are investigated (Chapter 3). This dissertation further extends the understanding of honesty behavior of individuals under differing social norms of honesty and examines a potential interaction with social image concerns (Chapter 5). Finally, to improve the examination of face-to-face communication in online experiments, this dissertation presents a newly developed video chat tool (Chapter 4). Methodologically, Chapter 2, 3, and 5 are fairly similar as they all apply an online experiment to elicit data on participant behavior. In addition, the projects discussed in Chapter 2 and 3 apply the newly developed video chat tool that is presented in Chapter 4. In the following, I describe each chapter's contribution and how it is linked to the economics literature.

Chapter 2 investigates the behavior of individuals and teams in unfair contests. The conventional wisdom in contest theory suggests that effort investment in a contest can be maximized by designing the contest as fairly as possible (e.g., Feess et al., 2008; Franke et al., 2018; Zhu, 2021). In the last years, the universality of these findings has been challenged by Drugov and Ryvkin (2017, 2022) who find multiple equilibria in which a biased contest design maximizes aggregated effort. To test the theoretic findings empirically, I investigate performance of individuals and teams in an online experiment. In a two-party contest, participants compete in a cognitively challenging task (see Coffman et al., 2021). The contest can be biased by a small/large handicap for one party. In an additional treatment, both parties know about the potential handicaps, but not whether one of them is applied to either party. Standard contest theory predicts that participants perform best in the fairly designed contest treatment without any handicaps. However, in my experiment, individuals in this treatment perform the worst. They are significantly outperformed by individuals who receive a large head start at the beginning of the contest and by individuals who are unaware of the true starting situation. As these results stand in stark contrast to theoretic predictions, performance in biased contests might not be solely generated by monetary incentives (as assumed by theoretic literature on contest design, e.g., Franke et al., 2018). Contestants may also experience positive reciprocity, such that they exert more effort when their expected payoff significantly increases through a large head start (see Heyman and Ariely, 2004). The high performance in the blind treatment might be caused by hedging against an unfavorable starting position due to risk aversion (see Holt and Laury, 2002). Lastly, participants in the treatment without handicaps might choke under the pressure of a close competition (see, e.g., DeCaro et al., 2011). The exact interplay of these multiple drivers that might affect performance in a contest has to be investigated in future research.[1]

Interestingly, teams remain virtually unaffected by the application of handicaps in a contest. The only marginally significant difference between team performances exists between teams that start with a large disadvantage and teams that start with a large advantage. However, in opposite to the individuals, the teams *suffering* from a large handicap outperform their counterparts. As this difference is only marginally significant, I focus on the finding that no team treatment significantly differs from the average performance. In addition, teams strongly outperform individuals across treatments. This finding was expected, as the task was designed such that deliberation within teams might benefit their performance. Moreover, teams are found to outperform individuals in domains like information processing (Kugler et al., 2012) and analytical problem-solving (Charness and Sutter, 2012), which are both highly relevant for the given task. Finally, also social image concerns in front of their team partner could lead the subjects participating as teams to maximize their performance such that they outperform individuals and eradicate all team treatment effects (Mas and Moretti, 2009). This leads to the conclusion that, to maximize performance and avoid effects of head starts or handicaps in a contest situation, cognitively challenging tasks should rather be endowed to teams than individuals.

Beside intellective decisions, teams often face judgmental decisions which do not require them to solve a cognitively challenging problem, but to find a solution that depicts the group mem-

[1] In addition, effort investment in contests designed as all-pay auctions is usually measured by simple effort tasks (e.g., Niederle and Vesterlund, 2007; Gill and Prowse, 2012; Charness et al., 2018). I use a cognitively challenging task (see Coffman et al., 2021) to see whether deliberation in teams can increase performance in a contest. Hence, simple effort investment and the performance in my task are not completely comparable and, thus, differences to theoretic predictions are, to some extend, expectable.

bers collective preferences as close as possible in a mutual decision. One domain of judgmental decisions is unethical behavior and, as high-profile cases like Volkswagen or Enron show, corporate fraud is definitely a topical issue. Also apart from these prominent and media-effective cases, unethical corporate behavior is wide-spread. Kroll (2016) finds that 75% of surveyed senior executives state that their company had become a fraud victim in the previous year. This renders the overall investigation of unethical behavior a current and important topic. For individuals, Abeler et al. (2019) provide a comprehensive overview of 90 experimental studies in economics, psychology and sociology. The existing literature for groups is rather scarce but growing. For example, Muehlheusser et al. (2015) examine honesty behavior of dyads while Kocher et al. (2018) study triads. However, to the best of my knowledge, larger group sizes and the overall effect of group size on honesty behavior have not been studied yet. Therefore, Chapter 3 (co-authored by Gerd Muelheusser, Andreas Roider, and Niklas Wallmeier) aims to fill this gap by comparing honesty behavior of groups with the size of 2, 3, 4, and 5, as empirical studies of top management teams in the U.S. find the average team size to be around 3.4 with standard deviations of 1.2-1.5 (see Haleblian and Finkelstein, 1993; Amason and Sapienza, 1997). In addition, not only the size of a group but also its gender composition might be relevant for the group's honesty behavior. According to a survey by the Association of Certified Fraud Examiners (2022), 73% of cases of corporate fraud are committed by men, and 27% by women. Hence, we want to examine whether this behavior translates to group decisions, especially, when the group decision has to be unanimous.

To investigate how group size and group gender composition affect the honesty behavior of a group, we set up an online experiment. We use an adapted version of the Fischbacher and Föllmi-Heusi (2013)-die roll paradigm to elicit honesty and apply a newly developed video chat tool (Chapter 4) to enable face-to-face communication within groups. In a total of 18 treatments (all possible gender compositions for group sizes 2, 3, 4, and 5) we collect data for 1677 participants in 477 groups. As our first result, we find that dishonesty significantly increases with group size. In particular, groups of 5 lie more than twice as much as groups of 2. Secondly, we find that the gender composition of a group has a significant effect on its honesty. Remarkably, this finding is solely driven by all-male groups, who lie significantly more compared to every other group gender composition within all group sizes (except groups of three, where all-male groups still lie the most, but not by a significant margin). However, adding one female participant to a group completely mitigates the pronounced unethical behavior of all-male groups. Almost all-male groups (i.e., groups with one female participant) do not lie significantly more than groups with a higher female percentage. In fact, if all-male groups are excluded, there are no significant differences in lying between any two gender compositions within each group size.

Hence, as all-male groups distinctly differ from all other gender compositions, we try to find reasons why one female group member is enough to significantly influence a group's decision. One viable explanation would be that the female group member convinces the males in her group that being honest is the right way to go or just stubbornly insists on behaving honestly until the male group members give in, as the group decision has to be unanimous. However, we do not find longer discussion times in almost all-male groups compared to all-male groups. In addition, related to findings of Karpowitz et al. (2024), the single female group members have a maximally equal but most often lower than average share of total talking time in the group. This refutes both proposed explanations in which the female group member has a rather active role in changing the group decision. Hence, the mere presence of a female seems to change the honesty behavior of male group members. We will investigate in a subsequent study whether males have stronger image concerns in front of a female and want to appear more honest in her presence, or whether males *expect* the female to have strong honesty preferences and want to avoid long discussions by adapting their behavior and make their decision in line with these expected preferences.

We can learn from Chapter 3 that creating all-male groups should be avoided at all costs if the group is expected to make ethical decisions. Also, increasing a group's size has detrimental effects on it's honesty as the probability to get at least one bad apple, who proposes unethical behavior, increases. In fact, Dimmock et al. (2018) find that misconduct in the work place is contagious, such that the probability of misconduct for one individual increases with the number of bad apples in their team. Hence, another way to achieve more ethical group decisions is to improve the individual honesty of each group member. Abeler et al. (2019) conclude that the main drivers for individual honesty are *to be honest* and *to be seen as honest*. Therefore, if an individual is dishonest by nature, increased observability of their actions can improve their honesty. As an additional channel, previous research finds that *norms* play an important role in people's prosocial behavior (e.g., Bicchieri, 2016; Bicchieri et al., 2022; Cialdini et al., 1990; Krupka and Weber, 2013). Bicchieri and Muldoon (2011) remark aptly: "Whenever individuals believe they are expected by their group ... to behave according to a given standard, and also expect the norm to be generally followed, they usually comply."

Now, the question remains whether the two channels of increased social image concerns through the observability of actions and the application of social norms can be combined to even further mitigate dishonest behavior. Chapter 5 (co-authored by Christoph Huber, Christos Litsios, and Annika Nieper) investigates this question in an online experiment. Also in this chapter, an adapted version of the Fischbacher and Föllmi-Heusi (2013)-die roll paradigm is applied to elicit honesty behavior. Previous studies have shown that participants cheat less in an experimental die-rolling task if they are observed by others (Gneezy et al., 2018; Fries et al., 2021; Bašić and Quercia, 2022). However, increasing people's concern for how others view them can also have no effect on honesty; it can even increase dishonesty if the observer benefits from a dishonest decision (see Weisel and Shalvi, 2015; Kocher et al., 2018). Hence, we hypothesize that under an honest norm, social image concerns increase honesty, while under a dishonest norm, social image concerns decrease honest behavior.

As a first result, we find that social norms can have a significant effect on honesty. Participants who observed other participants being dishonest, lie by almost 20 percentage points more than participants who observed others being honest. Our results show that normative and empirical expectations were both significantly affected by the induced norm nudge: Compared to the honest social norm, participants who observed liars regard lying as more socially appropriate and expect the overall percentage of liars to be higher. However, social image concerns have no effect on honesty behavior in our particular setup. As a consequence, also the interaction between social norm and social image concerns remains ineffective. We suspect the digital anonymity of our experimental setting to be the main reason for the lack of social image concerns, as, according to Cohn et al. (2022), the physical presence of humans can enhance such concerns. As now, after the COVID-19 pandemic, experiments with physically present participants are possible again, our hypothesis could be tested in a subsequent study with stronger social image concerns. Still, it remains as a major takeaway, that social norms can strongly affect honesty behavior. Therefore, managers of organizations are advised to increase the observability of actions when they assess the majority of team members as honest. Conversely, overall observability should be decreased, if a lot of malfeasance is suspected.

Beside the replication of the experiment discussed in Chapter 5 in physical presence, adding a video chat could also increase social image concerns to a certain amount. Several studies show that the percentage of people working from home strongly increased during the COVID-19 pandemic and that this movement did not fully reverse after the pandemic was over (e.g., Von Gaudecker et al., 2020; Bick et al., 2023). Hence, studying online face-to-face communication becomes more and more relevant. In addition, Bos et al. (2001) find that the effect of face-to-face communication in a trust game does not significantly differ between online and lab environments. Brosig and Weimann (2003) confirm this finding for a cooperation game, while Grözinger et al. (2020) reach similar findings for creative collaboration. Therefore, Chapter 4 (co-authored by Jan-Patrick Mayer, Gerd Muehlheusser, Andreas Roider, Eugen Tereschenko, and Niklas Wallmeier) introduces a newly developed video chat tool that can be easily integrated in oTree (Chen et al., 2016) experiments. The video chat tool was already applied in the projects presented in Chapter 2 and Chapter 3, successfully hosting a video conference for up to 5 people. The code and a comprehensible instruction how to implement it are deposited on GitHub. This deposition includes the code for capturing audio levels of participants, which is applied in Chapter 3.

The main contributions of this dissertation concern individual and collective decision-making in ethical dilemmas and unfair contests. Chapter 2 shows that individuals who receive a large head start perform best in biased contests. In contrast to theoretic predictions, they perform the worst when the contest is designed fairly. Competing teams remain virtually unaffected by head starts and handicaps and seem to maximize performance independent of their starting position. Teams significantly outperform individuals in a cognitively challenging task. Chapter 3 adds to the understanding of group behavior in the domain of honesty. Honesty significantly decreases with an increase of group size. In respect of a group's gender composition, all-male groups are to be avoided at all costs as they are significantly more dishonest than all other group gender compositions across almost all group sizes. Chapter 4 provides a video chat tool to further study behavior of groups in online environments. Finally, Chapter 5 continues to study honesty behavior and shows that a social norm of honesty or dishonesty can significantly shift a person's honesty behavior.

Chapter 2.

Unfair Contests - Experimental Evidence for Individuals and Teams

Abstract

Competitive business environments oftentimes do not feature equal starting positions for each competitor. Unequal starting positions result in unequal winning chances. In theory, greater inequality should lead to less effort investment and thereby a weaker performance in a competition, as winning chances become less dependent on effort if the inequality of the starting positions increases. To test these theoretic predictions, this paper investigates how competitors perform in a contest with different handicaps being applied. Furthermore, potential changes of performance when teams instead of individuals compete will be examined. Lastly, individual and team behavior under unknown handicaps will be explored. Surprisingly, performance of individuals significantly *increases* for highly unequal starting positions, in particular for the advantaged participants. Meanwhile, teams heavily outperform individuals but remain virtually unaffected by handicaps.

Keywords: Contests, asymmetric starting situations, online experiments, teams

JEL Classification: D44, C91, C92

2.1 Introduction

Competition with others is omnipresent. Throughout our whole lives, we fight with others for a variety of prizes. We might compete for a job, a promotion, a bonus or even more recognition by our superior (e.g., see Konrad, 2009; Vojnović, 2015). Within these types of contests, opponents regularly face unequal starting positions. Sometimes, the inequality may arise unintentionally. Better looking applicants can have a higher chance of being hired (e.g., Mobius and Rosenblat, 2006) and evaluation committees can be subconsciously biased against one gender (Bagues and Esteve-Volart, 2010; Bagues et al., 2017). Other times, unequal starting chances may be present by design. Gender quotas, for example, seem to give women a favored position in hiring decisions. However, as these quotas were introduced to *fight* inequality, there might have been unequal hiring chances for men and women in the first place. Depending on the situation, the unfairness of a contest can be rather hidden. A gender quota is communicated openly, while a supervisor might not even know herself that she has a personal bias against one candidate. The inequality of a starting position in a contest can also have different levels of inequality. If the response time on a call for tenders is limited by a fixed deadline, the time for preparation might vary drastically, depending on when an applicant learns of this call.

Based on the vast theoretic literature on contest design (e.g., Feess et al., 2008; Franke et al., 2018; Zhu, 2021), contests should be designed as fairly as possible to induce the highest amount of effort investment. Effort investment yields the strongest influence on winning chances when a competition is really close. With decreasing chances of winning/losing a contest, a competitor's performance is expected to decline as well. Although the bulk of theoretic literature on contest design concerns individuals, the main predictions can be transferred to teams, as a team's behavior, in general, is found to be even closer to theoretic predictions than an individual's (Kugler et al., 2012). However, the main factors affecting performance in the standard theoretic framework are the, most often monetary, incentive of winning the contest and the cost of exerting effort. Influences like contestants choking under the pressure of a close competition (e.g., DeCaro et al., 2011; Byrne et al., 2015; Ariely et al., 2009; Dohmen, 2008) or feeling obligated to perform well if the contest is biased in their favor (see Heyman and Ariely, 2004) are neglected.

To closer investigate actual behavior of individuals and teams in a contest under various handicaps, I designed an online experiment. The experiment features a cognitively challenging task (see Coffman et al., 2021) that is solved either individually or in teams of two. Every individual (team) enters a contest with another individual (team). Participants in the team treatments are able to communicate via video chat with their team partner but not with the other team. Depending on the treatment, either a small or a large handicap is applied to one of the two contesting parties. As a consequence, one party starts with a small (large) disadvantage in the contest, while the other party benefits from their opponent's handicap. In a baseline treatment, none of the two parties is endowed with a handicap inducing a contest with equal starting situations. Lastly, in a blind treatment, both parties remain uninformed whether one of them starts the contest with a handicap.

The remainder of this paper is organized as follows. In Section 2.2, relevant literature related to my research question is reviewed. Section 2.3 discusses the hypotheses that are to be tested while Section 2.4 describes the experimental design and procedure. In Section 2.5, the main results are presented and in Section 2.6 they are discussed. Section 2.7 concludes.

2.2 Literature

Overall, there is a vast amount of literature on contest theory investigating how individuals are expected to behave when they are competing for a prize. Corchón et al. (2018) provide the most recent overview discussing how effort translates to winning probabilities in several contest success functions and how the basic model can be extended. Chowdhury et al. (2023) add a summary of theoretical results concerning affirmative actions in contests.

As I focus specifically on *unfair* contests, the theoretic background on biased starting situations has the most relevance for this paper. The closest related theoretical paper on unfair contests is provided by Feess et al. (2008) who analyze a two-player discriminatory contest. They find that the optimal strategy for a heavily disadvantaged player is to reduce his costly effort investment. As a consequence, the advantaged player can also reduce her effort investment without significantly decreasing her winning probability. The result of this strategy is that contests with equal chances for both competitors usually yield the highest effort investment. Franke et al. (2018) add a mechanism for contest designers such that equal chances in an uneven contest can be achieved by giving the weaker player a head start. Zhu (2021) confirms that "leveling the playing field" by supporting the weaker contestant should, theoretically, induce the highest aggregated effort. Drugov and Ryvkin (2017) challenge this commonly acknowledged finding that, for symmetric players, designing a contest fairly does always induce the highest effort levels. They find multiple equilibria in which a biased contest design maximizes aggregated effort. In a second paper, Drugov and Ryvkin (2022) further challenge that the "discouragement effect" (i.e., decreased motivation of disadvantaged contestants) is always present in heterogeneous contests.

Investigating contests with heterogeneous players, Siegel (2010) examines the impact of head starts on effort investment. He finds a scenario in which weaker players invest effort more aggressively than stronger players, which further opposes standard findings from all-pay auction theory (for extensions of this finding see Siegel, 2014a and Siegel, 2014b). Kirkegaard (2012) proposes that giving the weaker player a head start while simultaneously discounting her effort can be the optimal employment of external instruments in a contest. Franke et al. (2013) also bias an unfair contest in favor of the weaker player to induce a closer competition. However, to maximize the aggregated effort investment in this scenario, the stronger player has to maintain the higher winning probability. Finally, Mealem and Nitzan (2016) provide a survey on discriminatory contests, concluding that in all but one of the examined cases, "reduction of the asymmetry between contestants enhances competition and, in turn, exerted efforts." Head starts in a simple lottery constitute the only exception.

Experimentally, applying affirmative action to purposefully bias a contest in favor of the weaker contestant has been found to be effective. Schotter and Weigelt (1992) find that equalizing opportunities in a tournament increases effort levels of all contestants, but only if the previous disadvantage was severe. Affirmative action has the opposite effect when the initial disadvantage was only small. Calsamiglia et al. (2013) confirm these findings in a field experiment. They create a competition for children from two schools who receive a different amount of training on a specific task. To equalize winning chances, they bias the contest in favor of the less trained children. This enhances their performance without significantly reducing the performance of the others. Niederle et al. (2013) and Balafoutas and Sutter (2012) investigate the effect of gender quotas. They find that gender quotas increase the probability of women entering competitions without an overall decrease in performance. For a broader overview on experimental findings concerning contests, tournaments and all-pay auctions, Dechenaux et al. (2015) provide a comprehensive survey.

2.3 Hypotheses

The following hypotheses[1] reflect behavioral expectations based on contest theory. These expectations are to be tested and the results will be discussed critically.

H1a Participants in the 'no handicap' treatment will perform best.

H1b Participants in the 'small handicap' treatment will perform worse than participants in the 'no handicap' treatment. They will perform better than participants in the 'large handicap' treatment.

H1c Participants in the 'large handicap' treatment will perform worst.

Hypotheses H1a, H1b, and H1c reflect the relationship between chances of winning the contest and performance based on effort investment. In theory, more effort to increase individual performance is invested when a competition is close, because a strong performance in a close competition has the highest chance to change the contest's outcome and win the proclaimed prize (Feess et al., 2008). The contest in this experiment is closest in the 'no handicap' treatment, less close in the 'small handicap' treatment, and practically a foregone conclusion in the 'large handicap' treatment. Hence, if participants invest more effort when their investment has the highest chance to change the outcome of the contest and investing effort translates into a stronger performance, participants in the 'no handicap' treatment should perform best and participants in the 'large handicap' treatment should perform worst.

H2 Participants in the 'blind handicap' treatment will not perform significantly worse than participants in the (best performing) 'no handicap' treatment.

Participants in the 'blind handicap' treatment have an equal chance to either suffer from a small handicap, suffer from a large handicap, benefit from a small handicap for their opponent, benefit from a large handicap for their opponent or end up in an equal starting situation. Due to risk aversion, participants in the 'blind handicap' treatment are expected to exert too much rather than too little effort and behave similarly to how they would if they were in the situation in which effort investment is most effective: the equal starting situation (see Holt

[1] The hypotheses were pre-registered with a slightly different wording solely referring to teams but not individuals. The investigation of individual behavior was included in this study after data collection for teams started, but before any data was analyzed. As the hypotheses for teams are based on theoretic predictions for individual behavior, they also apply for individuals.

and Laury, 2002). Therefore, participants in the ‘blind handicap’ treatment are expected to perform slightly but not significantly worse than participants in the ‘no handicap’ treatment (Hypothesis H2).

H3a Participants in the ‘small handicap’ treatment who suffer from the handicap will not perform significantly different to participants in the ‘small handicap’ treatment who benefit from the handicap.

H3b Participants in the ‘large handicap’ treatment who suffer from the handicap will not perform significantly different to participants in the ‘large handicap’ treatment who benefit from the handicap.

After one party in the ‘small (large) handicap’ treatment has received a handicap, the chances of winning a contest decrease in exactly the same amount for the disadvantaged party as they increase for the advantaged party. Hence, if performance is primarily related to effort investment and if effort investment depends on its necessity to win a contest, advantaged and disadvantaged parties should not differ in their performance (Hypotheses H3a and H3b).

2.4 Experimental Design

To investigate the given hypotheses, I used an experimental framework closely related to the popular game show *Family Feud*.[2] This experimental design is based on Coffman et al. (2021), who were the first to apply this task.[3] To build the necessary foundation for the main study, a pre-study, in which 100 participants were asked to answer 50 simple questions, was conducted.[4] From these 50 questions, ten questions were selected for the main study.[5] Participants in the main study had to answer these ten selected questions and gained points, depending on how frequently their answer was given in the pre-study. To further illustrate, Figure 2.1 displays the answer distribution for one of the questions which were used in the main study.

[2] In this game show, two competing parties gain points by answering questions. All questions were previously answered by exactly 100 people in a survey. Both parties try to find answers similar to those given in the previous survey. A party gains points by answering as similarly to as many participants from the previous survey as possible. If a party comes up with an answer that was given by 43 people in the previous survey, this party is awarded 43 points for this answer. The objective is to maximize points.

[3] This experimental design was chosen over classic effort tasks (e.g., Niederle and Vesterlund, 2007; Gill and Prowse, 2012; Charness et al., 2018) such that collective deliberation in teams can have a significant impact on performance.

[4] These 50 questions were partly picked from a web database, `https://www.familyfeudfriends.com/answers/`, the same source that Coffman et al. (2021) used. The web database did not cover the targeted number of questions completely. Therefore, the selection was completed by questions I created, aiming for a similar question format as the original family feud questions.

[5] The questions were sorted by the Gini-coefficient of their answer distribution in ascending order. Out of each bracket of 5 consecutive questions, one questions was picked for the main study.

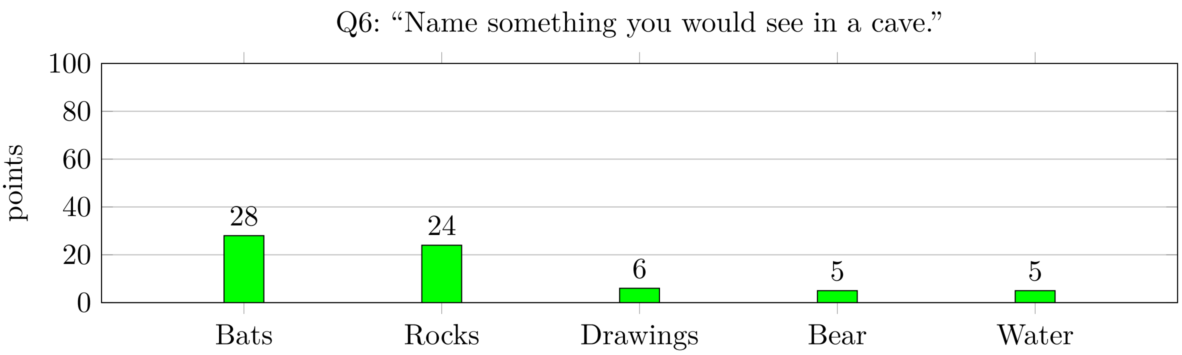


Figure 2.1: **Answer Distribution for Example Question.**

This figure displays the answer distribution of the five most frequently given answers on the question "Name something you would see in a cave." of participants in the pre-study. 28 participants in the pre-study answered "Bats", 24 participants answered "Rocks", 6 participants answered "Drawings", and 5 participants answered "Bear", respectively "Water". Answers that were given less often than five times are not displayed but still awarded points.

If a party in the main study answered the question "Name something you would see in a cave." with "Bats", 28 points were awarded. If participants answered "Rocks", they gained only 24 points for this question. The answer "drawings" would award 6 points, and so on.[6] A party's objective in the main study was to maximize the sum of points gained from all ten questions by answering as similarly to participants in the pre-study as possible.

2.4.1 Experimental Treatments and Procedure

The main study employed a 2x6 between-subjects design. The first dimension distinguished between subjects who participate individually and subjects who are matched in teams of two. Team members were able to communicate via video chat. The second dimension distinguished between different starting situations of contestants. Table 2.1 provides an overview of the six different starting situation treatments.

[6] When the given answer was similar enough to a listed answer from the pre-study, the corresponding points were awarded as well. For example, if a party in the main study answered "stones" on the example question in Figure 2.1, 24 points were awarded, as "stones" is similar enough to "rocks".

Treatment	Starting Situation		
	Circumstances	Your party	Opponent(s)
No handicap	No handicaps for both parties	0 points	0 points
Small handicap (advantaged)	Small handicap for other party	0 points	-40 points
Small handicap (disadvantaged)	Small handicap for your party	-40 points	0 points
Large handicap (advantaged)	Large handicap for other party	0 points	-200 points
Large handicap (disadvantaged)	Large handicap for your party	-200 points	0 points
Blind handicap	Handicaps unknown for both p.	?	?

Table 2.1: **Treatment Overview.**

Depending on the treatment, you/your team or your opponent(s) may start the contest with a handicap. In the treatment ‘no handicap’, both parties start without handicaps. In the treatment ‘small handicap (advantaged)’, your opponent(s) will start the contest with a small handicap of -40 points. In the treatment ‘small handicap (disadvantaged)’, you/your team will start the contest with a small handicap of -40 points. In the treatment ‘large handicap (advantaged)’, your opponent(s) will start the contest with a large handicap of -200 points. In the treatment ‘large handicap (disadvantaged)’, you/your team will start the contest with a large handicap of -200 points. In the treatment ‘blind handicap’, handicaps are unknown for both parties.

In these treatments, one of two contesting parties could receive a handicap in the form of negative points. The other party benefited from their opponent’s handicap, as the handicap was deducted from the final sum of points and the contest was won by the party which achieved more points. The handicaps varied in size. The ‘large handicap’ entailed a deficit of -200 points, while the ‘small handicap’ was -40 points. In a baseline treatment, both parties received ‘no handicap’. The last possible starting situation was the ‘blind treatment’ in which neither party knew if they or the other party started with a handicap. Figure 2.2 displays a diagram of all treatments. The colors of the starting situations match with the upcoming bar diagram in section 2.5, which displays the average performance of participants in each treatment.

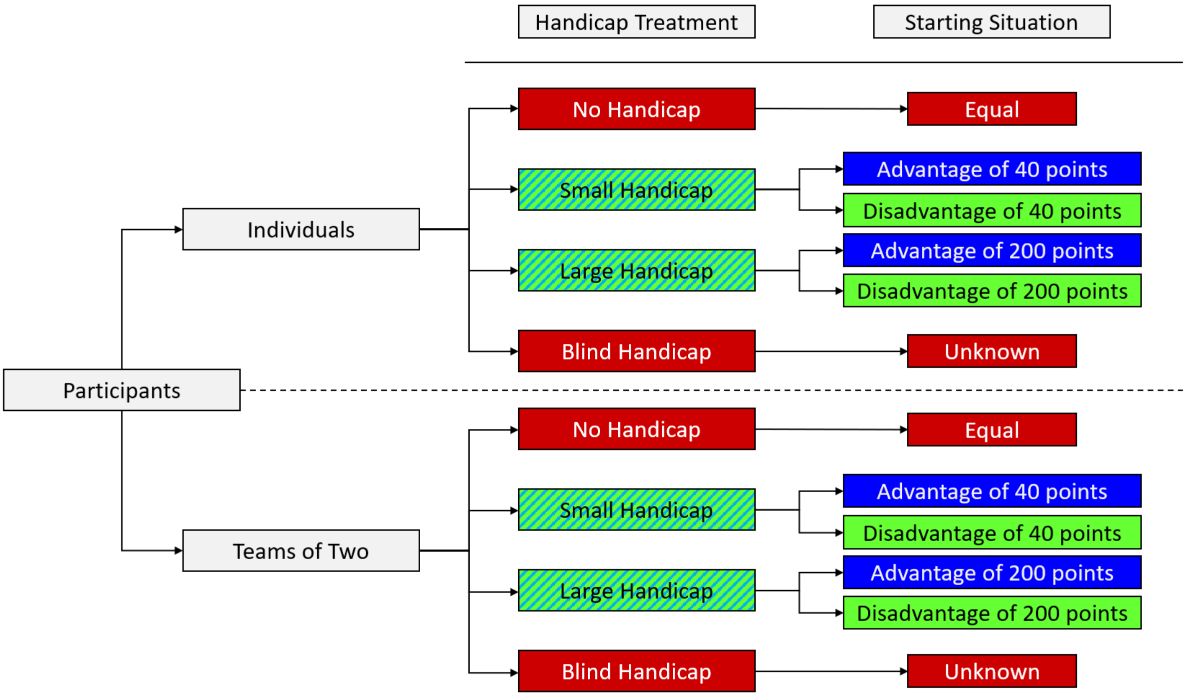


Figure 2.2: **Treatment Overview.**

This figure displays all treatments. Participants participated either individually or in teams of two. The handicap treatment is assigned randomly. The starting situation of a party in the contest depended on the assigned handicap treatment.

Individuals and teams were randomly assigned to one treatment. Every individual participant competed against one other individual participant while every team competed against one other team. The party who won the contest by earning the higher sum of points was rewarded with a bonus payment that doubled their initial payoff. The maximal sum of points was 434. The points of all second-best answers summed up to 218. Individuals and teams passed through similar screens during the experiment. However, as teams communicated via video chat, participants in the team treatments had to pass a functional check of their camera and microphone individually before they could proceed to a video chat test screen. As soon as both team members indicated flawless communication on the test screen, they proceeded to a waiting room, ready to be matched with an opposing team. Each individual (team) chose a representing avatar consisting of a color and an animal (e.g., “red elephant(s)”). They had twelve colors and twelve animals to choose from.[7] After an avatar was chosen and an opposing individual (team) was found, the participants proceeded to an instruction page that introduced them to the main task. The main task followed. Participants were obligated to stay for at least two minutes on the page of the main task; their upper time limit to answer all ten questions on this page was 15 minutes. If an individual (team) was still on the page of the main task after the 15 minutes were exceeded, the page auto-submitted, leading the participant(s) automatically to the subsequent questionnaire.

[7] This measure was used to increase the credibility of participants competing against other, real participants. Feedback from other online experiments rendered this measure appropriate.

Teams had to agree upon one joint answer on every question. They could discuss freely via video chat, but after one team member entered an answer, both team members had to accept this answer by clicking an acceptance button. Participants were informed that an answer that was not mutually accepted would award no points. After the main task, participants were asked to rate their overall enjoyment of the study as well as their fairness perception of the starting situation in their treatment. Figure 2.3 displays a flow diagram for the participation process of individuals and teams. Every single screen can be reviewed in Appendix 2.D.

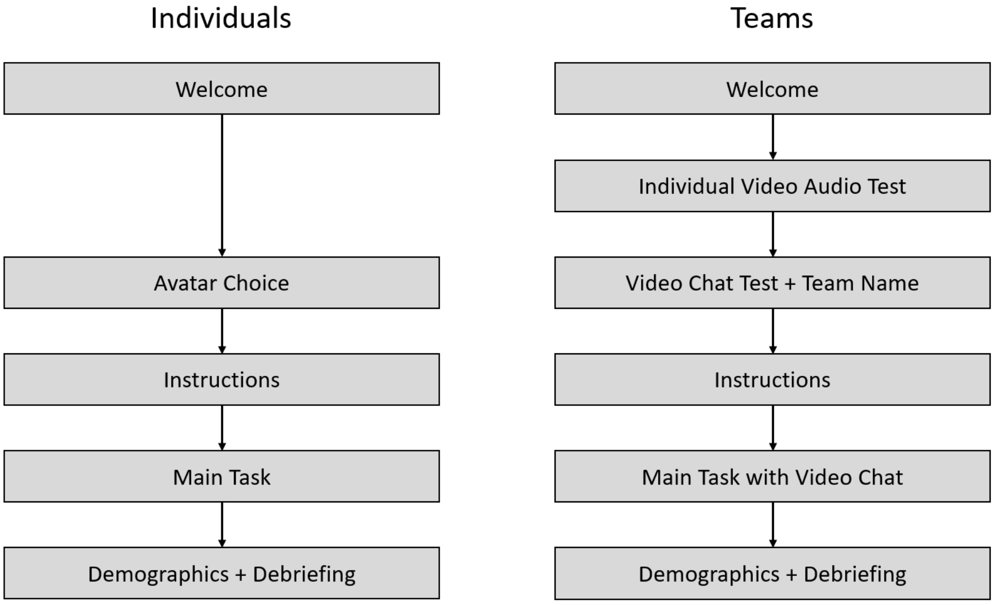

Individuals
Teams
Welcome
Avatar Choice
Instructions
Main Task
Demographics + Debriefing
Welcome
Individual Video Audio Test
Video Chat Test + Team Name
Instructions
Main Task with Video Chat
Demographics + Debriefing

Figure 2.3: **Flow Diagram.**

This figure displays the participation process of individuals and teams in the form of a flow diagram. Each step does not necessarily represent only one screen.

Participants were not informed about their own or their opponent's performance at any point in the experiment. This was because points for submitted answers were calculated manually after each experimental session. This way, typos in submitted answers or the use of synonyms could be accounted for. Participants knew whether they had won or lost the contest after they received their payment as the contest winner received twice as much.

2.4.2 Experimental Implementation

The experiment was programmed and conducted using *oTree* (Chen et al., 2016). Participants were recruited via the WiSo-Lab in Hamburg using *hroot* (Bock et al., 2014) and via *Prolific* (Palan and Schitter, 2018) after the participation pool of the former dried out.

The participants in the pre-study were solely recruited via the WiSo-Lab in Hamburg. They received a participation fee of 4€. A bonus payment of 50€ was awarded to the participant who earned the most points.[8] Payoffs for participants in the main study are displayed in Table 2.2.[9]

	Individuals	Teams (Prolific)	Teams (WiSo-Lab)
Base Payoff	£3	£4	5€
Bonus Payoff	£3	£4	5€
Average Payoff	£4.5	£6	7.5€

Table 2.2: **Payoffs.**

The payoffs depend on the first treatment dimension (individuals/teams), performance and participant pool. Every participant received the base payoff. Every participant who won their competition (i.e., every second participant) received the bonus payoff.

All participants from both platforms participated online using instructions in English. Participants needed to be at least 18 years old and speak English fluently. Participants from Prolific needed to have an approval rating of at least 90% and to have previously participated in at least five studies. The participation pool recruited via Prolific was limited to participants from the UK.

In the pre-study, 100 subjects participated over the course of two sessions in May 2023. For the main study, a total of 596 participants was recruited. 396 subjects participated in the team treatments forming teams of two. This resulted in 198 observations evenly divided between the handicap treatments ‘no handicap’, ‘small handicap’, ‘large handicap’ and ‘blind handicap’.[10] Of these 396 team treatment participants, 188 were recruited via the WiSo-Lab in Hamburg within six sessions in June 2023. The remaining 208 team treatment participants were recruited via Prolific in five sessions in August/September 2023. 200 subjects participated in the individual treatments via Prolific, spread over two sessions in September 2023. Across treatments, nine teams and three individual participants were excluded because they answered less than three (mostly zero) questions. This leads to a final sample of 575 participants (197

[8] Pre-study participants earned points by answering similarly to the other 99 pre-study participants. For example, if a participant answered “earth” on the question “Name a planet in this solar system.” and 53 other participants also gave this answer, all these 54 participants got 54 points for this particular question. The pre-study comprised 50 questions.

[9] Payoffs slightly differed between participant pools and within the first treatment dimension (individuals/teams). After the participation pool of the WiSo-Lab in Hamburg dried out, the observations were completed by recruits from Prolific. As these two platforms pay their participants in different currencies and as I wanted to avoid paying very odd payoffs to exactly equalize the payoff’s value, there was a minor difference between WiSo-Lab and Prolific payments. In addition, individuals were expected to finish the experiment faster than teams as they did not face the technical testing pages to ensure video chat functionality and did not need to reconcile their answers with a team partner. Hence, participants in the individual treatments were paid less than participants in the team treatments to reach a similar ratio between payoff and time investment.

[10] Note that observations in the handicap treatments ‘small handicap’ and ‘large handicap’ have to be divided between advantaged and disadvantaged participants, resulting in less observations per sub-treatment.

individuals and 189 teams) in the main study.[11] How the final observations are distributed across treatments is displayed by Table 2.3.

Treatment	no	small (d.)	small (a.)	large (d.)	large (a.)	blind	overall
Individuals	49	25	25	24	25	49	197
Teams (Prolific)	29	11	11	10	10	27	98
Teams (WiSo-Lab)	18	14	14	13	13	19	91
Teams (overall)	47	25	25	23	23	46	189
Overall	96	50	50	47	48	95	386

Table 2.3: **Number of observations per treatment.**

The treatments are abbreviated as follows: 'no' for 'no handicap' treatment, 'small (d.)' for 'small handicap (disadvantaged)' treatment, 'small (a.)' for 'small handicap (advantaged)' treatment, 'large (d.)' for 'large handicap (disadvantaged)' treatment, 'large (a.)' for 'large handicap (advantaged)' treatment, 'blind' for 'blind handicap' treatment.

2.5 Results

Overall, 57% of the participants were female and the average age was 36 years. On average, participants in the individual treatments took 6.5 minutes while participants in the team treatments took a little less than 16 minutes to complete the study. Table 2.6 in Appendix 2.A summarizes the distribution of demographics across treatments.

[11] The participants in the remaining sample answered mainly all ten but at least eight of the ten questions. The few observations with only eight or nine answers were kept while the missing answer(s) were awarded zero points, as, technically, participants were allowed to leave questions unanswered.

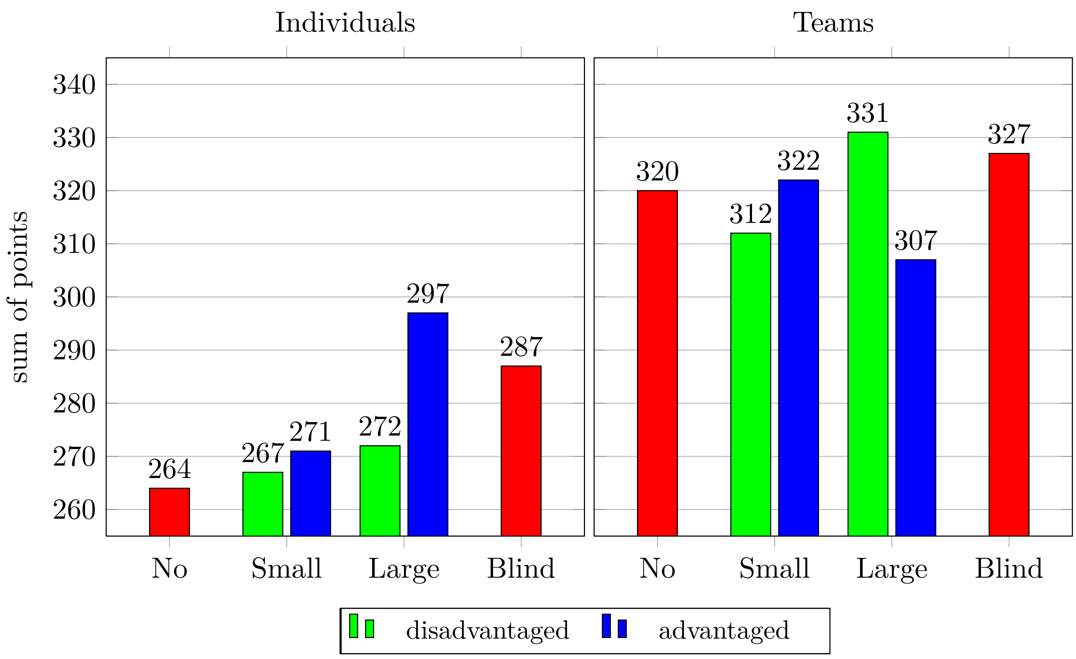


Figure 2.4: **Sum of points for all treatments.**

The left bar graph displays the average sum of points of individuals in the six handicap treatments. The right bar graph displays the average sum of points of teams in the six handicap treatments. The red bars display the average sum of points of participants in the 'no handicap' and 'blind handicap' treatments. The green bars display the average sum of points of *disadvantaged* participants in the 'small handicap' and 'large handicap' treatments. The blue bars display the average sum of points of *advantaged* participants in the 'small handicap' and 'large handicap' treatments.

Figure 2.4 provides an overview of the average sum of points individuals and teams accumulated in the six handicap treatments. Handicaps are only applied for payoff calculations and are not included in the displayed sums of points in Figure 2.4. Hence, Figure 2.4 reflects the actual performance of individuals and teams in each handicap treatment.

2.5.1 Main Results

The first and the most surprising result is

Result 1: *Individuals performed best in the 'large handicap (advantaged)' and 'blind handicap' treatments.*

Following Hypothesis H1c, both advantaged and disadvantaged individuals in the 'large handicap' treatment were anticipated to show the worst performance across all handicap treatments as they either had a very low chance of loosing or a very low chance of winning (see Feess et al., 2008). The small impact of their performance on the contest's outcome was expected to decrease the motivation for a high effort investment and thereby decrease the average performance. Even a very strong performance of disadvantaged participants in the 'large handicap' treatment would only have had a minor chance to overcome the large handicap of -200 points. However, out of all individual handicap treatments, participants in the 'large handicap (advantaged)' treatment performed best and accumulated an average of 297 points. They performed significantly better than participants from all other individual handicap treatments except the 'blind handicap' treatment.

Participants in the 'blind handicap' treatment performed second best and accumulated an average of 287 points. They met the expectation to perform not significantly worse than participants from the best performing treatment (compare Hypothesis H2). However, according to Hypothesis H1a, participants in the 'no handicap' treatment were anticipated to perform best—which they did not. In fact, participants in the 'no handicap' treatment performed worst. They were expected to perform best because the close competition within their handicap treatment rendered a strong performance most impactful on their contest's outcome (see, e.g., Zhu, 2021). As individuals in the 'no handicap' treatment accumulated an average of only 264 points, closeness of the contest seemingly does not lead to a strong performance in the given task.

The performance of individuals in the 'blind handicap' treatment, who were second best, partly confirms expectations concerning risk aversion in this treatment. Exerting effort to brace against an unfavorable situation is still a plausible theory to explain a high performance in the 'blind handicap' treatment (see Holt and Laury, 2002). However, the individuals in the 'no handicap' treatment performing worst raises questions regarding the main driver of a strong performance in the given task. Investing a lot of effort to prevail in a close competition can either be ruled out as the main force or is simply not effective enough to significantly increase performance.

The second and most eye-catching result is

Result 2: *Teams outperformed individuals.*

When comparing the left and the right bar graph in Figure 2.4, the difference in performance between individuals and teams across all handicap treatments becomes very apparent. Teams accumulated a higher average sum of points in each handicap treatment. This result was strongly anticipated as the task was designed to benefit joint deliberation, which is only possible for teams. In addition, teams have been found to outperform individuals in domains like information processing (Kugler et al., 2012) and analytical problem-solving (Charness and Sutter, 2012), which are both highly relevant for the given task. The difference between individuals and teams is significant at the 1%-level within each handicap treatment except 'large handicap (advantaged)', in which individuals performed best and teams performed worst—still slightly better than individuals but not significantly (297 points vs. 307 points, p=0.2288, t-test). These behavioral differences between individuals and teams in the 'large handicap (advantaged)' treatment are further illuminated in the discussion of result 3 and result 4.

Result 3: *Advantaged individuals outperformed disadvantaged individuals in the 'large handicap' treatment.*

Advantaged individuals in the 'large handicap' treatment accumulated an average of 297 points, while their disadvantaged counterparts achieved an average of just 272 points. Statistically, this difference is only significant at the 10%-level (p=0.0783, t-test) as the number of observations is quite low (24 vs. 25 observations, see Table 2.3). To shed more light on this result, we will go back to Figure 2.1, which displays the points for each answer on the question "Name something you would see in a cave." The maximum points a participant could get for this question was 28 for the answer "bats", the second most was 24 for "rocks", and the third most was 6 points for "drawings". Even the fifth best answer on this question would award 5 points. Hence, the difference between best and fifth best answer on this question was 23 points. The average difference between advantaged and disadvantaged individuals in the 'large handicap' treatment was 25 points. This means, on average, participants in the 'large handicap (advantaged)' treatment answered at least one question significantly better than their disadvantaged counterparts.

This result was not expected, as advantaged and disadvantaged participants within the 'large handicap' treatment were anticipated to perform equally well (see Hypothesis H3b). Following the discussion of result 1, pure motivation to win the contest seems not to be the main cause of a high performance in the given task, as advantaged and disadvantaged participants had equal chances to change the contest's outcome by investing more effort. A potential explanation is provided by Heyman and Ariely (2004), who find that better paid participants invest more effort. Advantaged participants in the 'large handicap' treatment started with a very high chance of winning their contest and, thereby, had a significantly higher expected payoff than their disadvantaged counterparts. Hence, the prospect of a high payoff could have been one cause for the strong performance of advantaged individuals in the 'large handicap' treatment.

Result 4: *Disadvantaged teams outperformed advantaged teams in the 'large handicap' treatment.*

In the 'large handicap' treatment, advantaged teams accumulated an average of 307 points while their disadvantaged counterparts achieved an average of 331 points. This difference is again only significant on the 10%-level (p=0.0558, t-test) as the number of observations is quite low (23 vs. 23, see Table 2.3). It is of particular interest that the treatment effect of receiving a handicap in the 'large handicap' treatment is exactly the opposite for teams and individuals. Advantaged individuals performed significantly better than their counterparts, while advantaged teams performed significantly worse. Hence, the explanation given for individuals cannot apply for teams or at least is much less effective and overshadowed by other effects.

In general, participants in the team handicap treatments do not differ much in their average performance. Disadvantaged teams in the 'large handicap' treatment perform slightly better than the average of all teams across treatments, while advantaged teams in the 'large handicap' treatment perform slightly worse. Hence, if there is a treatment effect within the team handicap treatments, it is not very strong. This remark is expressed in

Result 5: *Close to no significant differences between team treatments.*

Beside Result 4, the only significant differences between any two team handicap treatments can be found between the treatments 'blind handicap' and 'large handicap (advantaged)', 327 vs. 307 points (p=0.0628, t-test), as well as 'large handicap (disadvantaged)' and 'small handicap (disadvantaged)', 331 vs. 312 points (p=0.0863, t-test). When advantaged and disadvantaged teams in the 'small handicap' and 'large handicap' treatment are considered collectively, there are no significant differences between the remaining four team handicap treatments 'no handicap', 'small handicap', 'large handicap', and 'blind handicap'. As teams, in general, show a strong performance in the given task, it is likely that they maximize their performance unconditional of the treatment, resulting in just minor treatment effects. A potential explanation could be found in social image concerns toward their team partner leading to a higher effort investment and, thereby, to a stronger performance (see Mas and Moretti, 2009).

2.5.2 Enjoyment

After the main task, every participant was asked, "Please rate your enjoyment of this study from 1 to 10, with 1 being not enjoyable at all and 10 being totally enjoyable." This measure was employed as in standard contest theory (e.g., Feess et al., 2008) effort is costly and gets traded with expected payoff. If participants judge the given task as highly enjoyable, investing effort in this task is less costly. Hence, investing effort to increase the likelihood of a higher payoff would be less convincing as the main driver for a strong performance.

The average enjoyment over all treatments was 9.05. Participants in the individual handicap treatments enjoyed the study significantly more than participants in the team handicap treatments (9.40 vs. 8.87, p=0.0000, t-test). However, this difference decreases when only comparing individuals and teams who participated via Prolific (9.40 vs. 9.21, p=0.0391, t-test) as participants from the WiSo-Lab in Hamburg indicated the least average enjoyment (8.51). Statistically, the enjoyment of the study in the 'blind handicap' treatment is significantly worse than in all other treatments. This is (again) primarily caused by participants from the WiSo-Lab in Hamburg, who reported an average enjoyment of 7.42 in the 'blind handicap' treatment. The average enjoyment in all other treatments ranges between 8.53 and 9.63 (see Table 2.4 for more details).

Treatment	no	small (d.)	small (a.)	large (d.)	large (a.)	blind	overall
Individuals	9.63	9.48	9.20	9.46	9.40	9.20	9.40
Teams (Prolific)	9.21	8.86	8.82	9.45	9.55	9.30	9.21
Teams (WiSo-Lab)	8.53	8.86	9.11	8.96	8.61	7.42	8.51
Teams (overall)	8.95	8.86	8.98	9.17	9.02	8.52	8.87
Overall	9.18	9.07	9.05	9.27	9.15	8.76	9.05

Table 2.4: **Average enjoyment over all treatments.**

Participants followed the task, “Please rate your enjoyment of this study from 1 to 10, with 1 being not enjoyable at all, and 10 being totally enjoyable.” This table displays the average answers across all treatments. The treatments are abbreviated as follows: ‘no’ for ‘no handicap’ treatment, ‘small (d.)’ for ‘small handicap (disadvantaged)’ treatment, ‘small (a.)’ for ‘small handicap (advantaged)’ treatment, ‘large (d.)’ for ‘large handicap (disadvantaged)’ treatment, ‘large (a.)’ for ‘large handicap (advantaged)’ treatment, ‘blind’ for ‘blind handicap’ treatment. Due to significant differences, the average answers of participants in the team handicap treatments are displayed separately for Prolific and WiSo-Lab participants.

In conclusion, the enjoyment of the given task was very high and did not differ much between treatments. The only striking difference in enjoyment was found between participants from the WiSo-Lab in Hamburg and participants from Prolific. Hence, it is reasonable to believe that enjoyment of the study was mainly independent of the treatment. What remains is that the overall enjoyment was quite high. This supports the argument that effort was not solely invested to increase the chances of winning the contest but was at least partly invested out of enjoyment of the task.

2.5.3 Fairness Perception

In addition to enjoyment, the participants were asked for their fairness perception of their own starting situation in the contest: “Please rate the fairness of your starting situation in the competition with the other individual (team) from 1 to 10, with 1 being completely unfair and 10 being completely fair.” This question was employed to test whether treatments were perceived as differently fair. Since the handicap treatments differed in fairness by design, this measure was merely used to confirm that the different degrees of fairness are perceived as such. Perceiving one’s own starting situation as unfair could also be a factor that decreases motivation and thereby performance.

In general, between participants in the individual and participants in the team treatments fairness perception did not significantly differ. Within each handicap treatment, the only significant difference in fairness perception between individuals and teams was in the ‘large handicap (advantaged)’ treatment (3.72 vs. 5.17, p=0.0425, t-test). However, between handicap treatments, when pooling individuals and teams as well as advantaged and disadvantaged participants for the ‘small handicap’ and the ‘large handicap’ treatment, all pairwise comparisons between ‘no handicap’, ‘small handicap’, ‘large handicap’, and ‘blind handicap’ are significant at the 1%-level.

When comparing advantaged and disadvantaged participants in the ‘large handicap’ treatment from individual and team treatments jointly, fairness perception also significantly differed (3.43 vs. 4.66, p=0.0160, t-test). This difference is heavily driven by participants in the team handicap treatments (3.48 vs. 5.17, p=0.0137, t-test) as the difference between ‘large handicap (disadvantaged)’ and ‘large handicap (advantaged)’ is not significant for individuals (3.33 vs. 3.72, p=0.3156, t-test). Fairness perception between advantaged and disadvantaged participants in the ‘small handicap’ treatment only differed significantly for teams from the WiSo-Lab participant pool (4.46 vs. 6.00, p=0.0212, t-test) but not for individuals or teams from Prolific. The average fairness perception across all treatments is displayed in Table 2.5.

Treatment	no	small (d.)	small (a.)	large (d.)	large (a.)	blind	overall
Individuals	9.63	5.08	5.24	3.33	3.72	7.47	6.44
Teams (Prolific)	9.32	5.00	4.95	3.90	5.55	8.09	7.06
Teams (WiSo-Lab)	9.17	4.46	6.00	3.15	4.88	6.66	5.96
Teams (overall)	9.26	4.70	5.54	3.48	5.17	7.50	6.53
Overall	9.39	4.83	5.44	3.43	4.66	7.49	6.50

Table 2.5: **Average fairness perception over all treatments.**

Participants followed the task “Please rate the fairness of your starting situation in the competition with the other individual (team) from 1 to 10, with 1 being completely unfair and 10 being completely fair.” This table displays the average answers across all treatments. The treatments are abbreviated as follows: ‘no’ for ‘no handicap’ treatment, ‘small (d.)’ for ‘small handicap (disadvantaged)’ treatment, ‘small (a.)’ for ‘small handicap (advantaged)’ treatment, ‘large (d.)’ for ‘large handicap (disadvantaged)’ treatment, ‘large (a.)’ for ‘large handicap (advantaged)’ treatment, ‘blind’ for ‘blind handicap’ treatment. Due to significant differences in some treatments, the average answers of participants in the team treatments are displayed separately for Prolific and WiSo-Lab participants.

In summary, fairness is perceived as the design intended, with significant differences between ‘no handicap’, ‘small handicap’, ‘large handicap’ and ‘blind handicap’. Interestingly, there are hardly any significant differences between advantaged and disadvantaged participants within the ‘small handicap’ and ‘large handicap’ treatments, suggesting that at least the advantaged participants judge the starting situation very objectively. However, perceiving the starting situation in their treatment as unfair could have different effects on advantaged and disadvantaged participants as the former might feel some pity for their opponents while the latter could be somewhat angry or at least upset about their handicap. Whether and how fairness perception affects performance in the given task is rather unclear as it does not explain any of the main results.

2.5.4 Robustness Checks

To test how robust the given results are, several measures were applied. First, the observations for the six team handicap treatments were split up between the WiSo-Lab in Hamburg and Prolific. As a second measure, to address potential differences between WiSo-Lab and Prolific participants, the pre-study was repeated on Prolific. Thirdly, instead of analyzing points, each given answer received a rank to equalize the weight of top-answers between all ten questions. The visualized data for every robustness check can be found in Appendix 2.B.

WiSo-Lab vs. Prolific Teams

The original pre-study was conducted with online participants from the WiSo-Lab participant pool, while the main objective for participants in the main study was to guess the pre-study participants' answers. Hence, WiSo-Lab participants seem slightly advantaged compared to Prolific participants as the former are expected to be culturally and demographically closer to the participants in the pre-study. And indeed, WiSo-Lab participants outperformed Prolific participants in the team handicap treatments. For the team handicap treatments collectively, WiSo-Lab participants accumulated an average of 336 points while Prolific participants accumulated an average of only 307 points. This difference is highly significant (p=0.0001, t-test). The difference varies only marginally when comparing WiSo-Lab and Prolific participants in the team handicap treatments individually and always stays larger than 20 points. However, Result 2, teams outperforming individuals, still holds when individuals are compared to only those teams that participated on Prolific (276 vs. 307 points, p=0.0001, t-test).[12] When considering WiSo-Lab and Prolific observations separately, Result 4 only holds for the WiSo-Lab participants. In the WiSo-Lab sample, disadvantaged teams in the 'large handicap' treatment collected 31 points more than advantaged teams (p=0.0399, t-test). This difference amounted to only 15 points in the Prolific sample (p=0.2704, t-test). Nevertheless, the effect's direction is still opposite to Result 3 for both samples. When reviewing Result 5, it is noteworthy that the team handicap treatment 'small handicap (advantaged)' features significantly different performances between WiSo-Lab (351 points) and Prolific (286 points) participants. 351 points is the best and 286 points the worst performance for a single team treatment. Hence, for split samples, the 'small handicap (advantaged)' treatment significantly differs from other team handicap treatments. For the combined sample, virtually all of these differences cancel out.

[12] The individual treatments were entirely elicited on Prolific.

Pre-Study on Prolific

As a result of the differences between WiSo-Lab and Prolific participants, a second robustness check measure was applied. The pre-study was repeated on Prolific, such that all given answers in the main study could be re-evaluated with a second data set.[13] When applying the Prolific pre-study data set as the evaluation basis, participants in the individual handicap treatments fared significantly better. Individuals in the 'large handicap' treatment accumulated an average of 336 points when advantaged and disadvantaged participants are evaluated collectively. Evaluated with the Prolific data set, individual participants scored an average of 320 points in the 'no handicap' treatment, 316 points in the 'blind handicap' treatment, and 312 points in the 'small handicap' treatment (advantaged and disadvantaged participants evaluated collectively). Hence, Result 1 only holds for the 'large handicap' treatment as individuals in the 'blind handicap' treatment performed worst when evaluated with the new data set.

Teams still outperformed individuals when using the Prolific pre-study as the evaluation basis (333 vs. 321 points, p=0.0345, t-test), which confirms Result 2. However, when using the Prolific pre-study data set, the average performance of WiSo-Lab teams (314) and Prolific teams (351) differs distinctly. This confirms the assumption from the beginning of the previous subsection that main study participants are likely to perform better if they are culturally and demographically closer to the pre-study participants. Result 3 becomes statistically insignificant when evaluated with the Prolific pre-study data. Advantaged individuals in the 'large handicap' treatment scored an average of 346 points, while disadvantaged individuals in the 'large handicap' treatment scored an average of 325 points (p=0.1353, t-test). Hence, the difference in performance between advantaged and disadvantaged individuals in the 'large handicap' treatment is still visible but not statistically significant anymore. Result 4, however, also holds when applying the Prolific pre-study answers as the evaluation basis. Advantaged teams in the 'large handicap' treatment collected an average of 317 points whereas disadvantaged teams in the 'large handicap' treatment collected an average of 344 points (p=0.0653, t-test). Result 5 does not change for the Prolific pre-study data set.

Answer Ranks

As a third robustness measure, the given answers received ranks instead of points. This measure seemed appropriate as for the ten given questions used in the main study, the points for the best answers were very differently distributed (see Appendix 2.C). For some questions, the best answer awarded far more points than the second best. For other questions, point rewards for

[13] Participants in the main study were explicitly informed that all pre-study participants were recruited in the online participation pool of the WiSo-Lab in Hamburg. Therefore, answers might have been adjusted on the basis of this knowledge and the results based on the two different pre-studies are not fully comparable.

the first and second best answer were rather close (e.g., the best answer on question 1 awarded 85 points, but the second best answer only 7, and the best answer on question 2 awarded 27 points, while the second best answer awarded 25). Hence, the ranks 5-1 were appointed to all answers to put less weight on high point answers. Rank 5 was given to the best answer, rank 4 to the second best, rank 3 to the third best, and rank 2 to the fourth best. Rank 1 was given to all answers that awarded less points then the fourth best answer. If two answers in the top 4 awarded an equal amount of points, they received the average of their two ranks as a joint rank.[14] In the following, instead of the average sum of points, the average sum of ranks was used to evaluate the participants' performance. In addition, the average sum of ranks for WiSo-Lab pre-study answers *and* Prolific pre-study answers were applied separately.

When using the average sum of ranks and the original WiSo-Lab pre-study data set to rate the participants' performance, individuals in the 'blind handicap' and 'large handicap' treatment still perform best among all individual handicap treatments. Individuals in the 'blind handicap' treatment scored significantly higher ranks than individuals in the 'no handicap' and 'small handicap' treatment (36.75 vs. 34.98, p=0.0467; 36.75 vs. 34.99, p=0.0420, t-test). When using the Prolific pre-study data set, individuals in the 'large handicap' treatment (advantaged and disadvantaged participants collectively) scored significantly higher ranks compared to individuals from all other treatments (38.15 vs. 36.80, p=0.0792; 38.15 vs. 36.51, p=0.0476; 38.15 vs. 36.72, p=0.0614, t-test). This confirms Result 1.

When comparing the average sum of ranks and using the WiSo-Lab pre-study data set, teams still significantly outperformed individuals (38.65 vs. 35.62, p=0.0000, t-test). However, when comparing the average sum of ranks while using the Prolific pre-study data set, Result 2 becomes insignificant (37.17 vs. 37.04, p=0.3918, t-test). This effect is solely driven by the poor performance of WiSo-Lab teams when evaluated with Prolific pre-study data, though. Only using Prolific teams for the comparison with individuals renders Result 2 also highly significant for the Prolific pre-study data set (38.73 vs. 37.04, p=0.0027, t-test).

Using the WiSo-Lab pre-study data set, advantaged individuals in the 'large handicap' treatment did not outperform their disadvantaged counterparts significantly (36.28 vs. 35.27, p=0.2462, t-test). However, when using the Prolific pre-study data set, Result 3 stays significant (38.96 vs. 37.31, p=0.0788, t-test).

Similarly, disadvantaged teams in the 'large handicap' treatment did not outperform their advantaged counterparts when using the WiSo-Lab pre-study data set (37.80 vs. 39.17, p=0.1515, t-test). However, Result 4 stays significant when using the Prolific pre-study data set (35.00

[14] If the second and the third best answer would both award 10 points, they would get the shared rank of 3.5—the average of rank 4 for the second best answer and rank 3 for the third best answer. The fourth best answer would still get rank 2.

vs. 38.02, p=0.0161, t-test). This outcome is mainly driven by the bad performance of the advantaged WiSo-Lab participants in the 'large handicap' team treatment. They performed significantly worse than their disadvantaged counterparts (33.11 vs. 37.15, p=0.0031, t-test) when using Prolific pre-study data.

The bad performance of heavily advantaged WiSo-Lab teams is also the only anomaly for Result 5 as all other comparisons of the average sum of ranks between team handicap treatments yield no significant differences for both pre-study data sets.

2.6 Discussion

Surprisingly, individuals who received a large head start in their contest delivered the best performance. Participants who competed without handicaps performed worst. Contest theory, however, would predict exactly the opposite outcome. The potential explanation for these results is manifold. Firstly, participants in all treatments reported very high enjoyment in this study and specifically the given task. Therefore, working on the task may not necessarily represent costly effort. In addition, time spent on the task, the most fitting instrument to measure invested effort, did not significantly differ between treatments. Nevertheless, performance significantly differed between treatments. Hence, other factors beside time investment have to be relevant for a strong performance in the given task. Heyman and Ariely (2004) find that better paid participants invest more effort. As the winning probability for heavily advantaged participants goes towards 1, these participants had the highest expected payoff among all participants. Hence, heavily advantaged participants might have felt obligated to invest a higher level of cognitive effort to deserve their high expected payoff. This explanation can be supported by the differing fairness perceptions of the starting situation. Heavily advantaged participants perceive their starting situation as a lot less fair than participants in the treatment without handicaps. The feeling that they received their advantage unfairly could amplify the sense of obligation to perform well. Alternatively, participants who competed without handicaps might have felt pressured by the close competition and choked (e.g., DeCaro et al., 2011). However, if participants did feel a lot of pressure in this treatment, I would expect them to enjoy the study less, but this is not the case. In fact, participants in the treatment without handicaps reported the highest average enjoyment.

The second result, teams outperforming individuals across almost all treatments, is less surprising and confirms various past findings. The superiority of teams over individuals is partially credited to advantages in information processing (Kugler et al., 2012) and analytical problem-solving (Charness and Sutter, 2012), which are both highly relevant for the given task. In addition, teams spend more than twice as much time than individuals on the task. As time investment has no significant effect on performance within individual or team treatments, this primarily suggests that coordination within a team takes time. Hence, giving a task to a team instead of an individual can improve the outcome but could postpone the completion.

In general, teams seem to be less influenced by handicaps. I find close to no significant differences in performance between team treatments. The high outcomes throughout suggest that teams might maximize their performance in the given task unconditional of the treatment. One possible explanation are social image concerns. Assuming that individuals have light social image concerns about their performance in front of the experimenter, this feeling will be amplified for participants in the team treatments as they have to share their ideas via video chat with their team partner. As a result, they might exert more effort (see Mas and Moretti, 2009).

Finally, the only significant difference between team treatments is found between heavily advantaged and heavily disadvantaged teams. Interestingly, the effect is opposite to the formerly discussed effect in the individual treatments—heavily disadvantaged teams outperform their counterparts. However, as the significance is only marginal and both treatments do not significantly differ from the average across all team treatments, this result is rather a side note.

2.7 Conclusion

In conclusion, individuals show their best performance in the given, cognitively demanding task when they receive a large head start. They perform second best when the potential handicap for either competitor is unknown. Individual participants perform the worst when the contest is designed completely fairly. Teams distinctly outperform individuals but remain virtually unaffected by handicaps. In summation, my findings suggest that the behavior of individuals and teams in unfair contests cannot be perfectly predicted by the presently given theory.

As a consequence, it might not always be optimal to design a contest completely fairly and provide full information to every competitor. A favored competitor may feel obligated to redeem the positive bias toward herself. For a supervisor, hiding her personal preferences can be optimal. Incomplete information can cause competitors to hedge themselves against the worst possible state of the contest by investing the maximal amount of effort. To avoid significant influence of an unfair starting situation in a contest, assigning the task to small teams instead of individuals can mitigate the impact of handicaps and improve the overall outcome.

Appendix

2.A Additional Tables and Figures

	N	Female	Age	Time Taken Decision	Time Taken Total
Individuals					
no handicap	49	69%	40.00	3.21	6.15
small handicap (disadvantaged)	25	68%	39.64	3.32	5.77
small handicap (advantaged)	25	64%	44.36	2.80	5.99
large handicap (disadvantaged)	24	75%	37.17	3.45	7.1
large handicap (advantaged)	25	68%	43.00	3.02	7.37
blind handicap	49	51%	36.00	3.26	6.70
Teams					
no handicap	47	53%	36.28	6.89	16.65
small handicap (disadvantaged)	25	52%	32.84	7.59	15.41
small handicap (advantaged)	25	68%	35.24	7.49	16.39
large handicap (disadvantaged)	23	57%	33.17	7.06	15.98
large handicap (advantaged)	23	39%	30.13	6.95	16.10
blind handicap	46	43%	34.33	7.63	17.86

Table 2.6: **Summary of participant demographics.**

This table summarizes participant demographics across all treatments. Within the individual, respectively the teams experiment, treatments were randomly allocated. ‘Time Taken Decision’ indicates how much time participants spent on the decision page (in minutes). ‘Time Taken Total’ indicates how much time participants spent to complete the whole experiment, including the decision page (in minutes).

2.B Data Graphs for Robustness Checks

2.B.1 WiSo-Lab vs. Prolific Teams

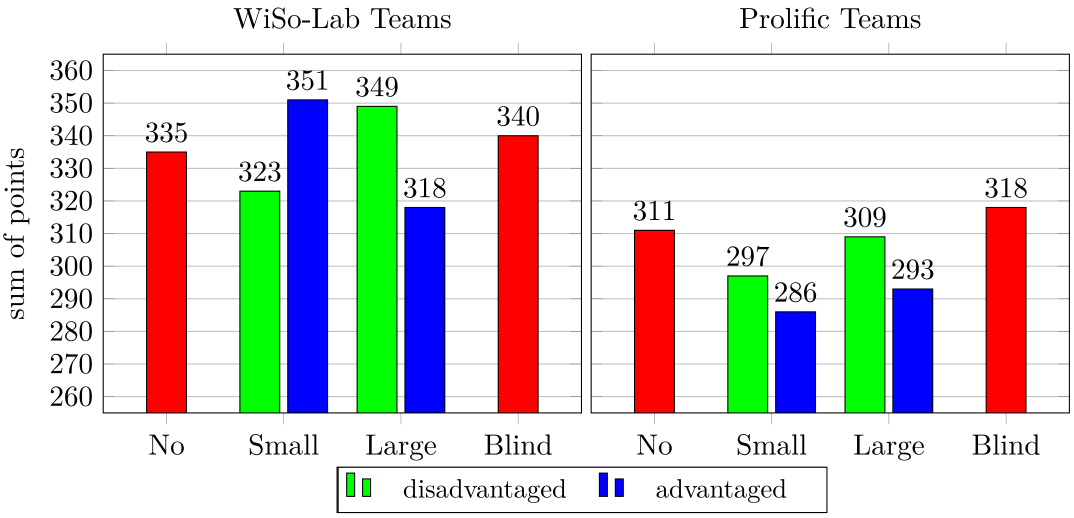


Figure 2.5: **Sum of points for all team treatments split up between WiSo-Lab and Prolific.**

The left bar graph displays the average sum of points of teams who participated via the WiSo-Lab Hamburg in all six treatments. The right bar graph displays the average sum of points of teams who participated via Prolific in all six treatments. The red bars display the average sum of points of participants in the ‘no handicap’ and ‘blind handicap’ treatments. The green bars display the average sum of points of *disadvantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments. The blue bars display the average sum of points of *advantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments.

2.B.2 Pre-Study on Prolific

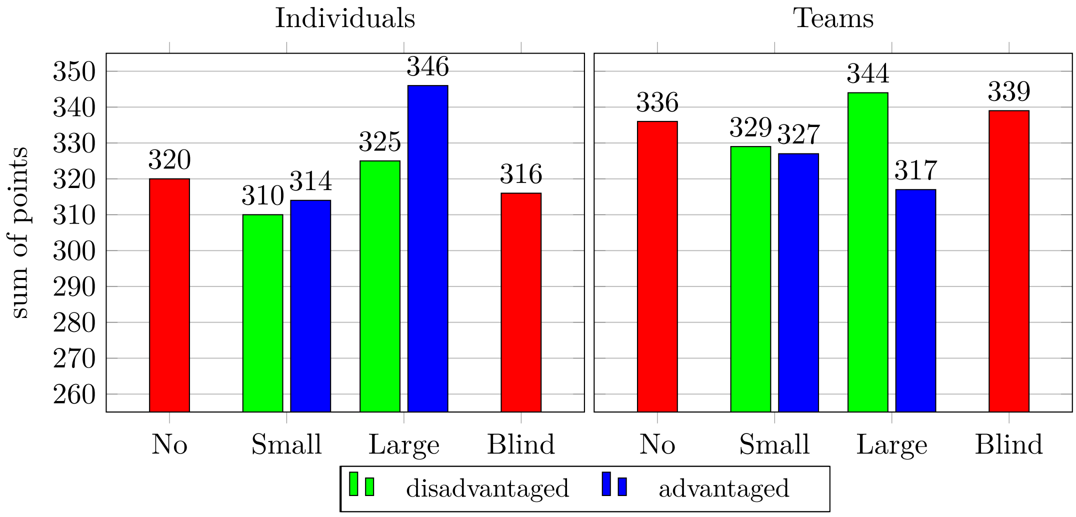


Figure 2.6: **Sum of points for all treatments with Prolific pre-study data set.**

The left bar graph displays the average sum of points of individuals in all six treatments. The right bar graph displays the average sum of points of teams in all six treatments. The red bars display the average sum of points of participants in the ‘no handicap’ and ‘blind handicap’ treatments. The green bars display the average sum of points of *disadvantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments. The blue bars display the average sum of points of *advantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments. All points are calculated using the data set of the Prolific pre-study.

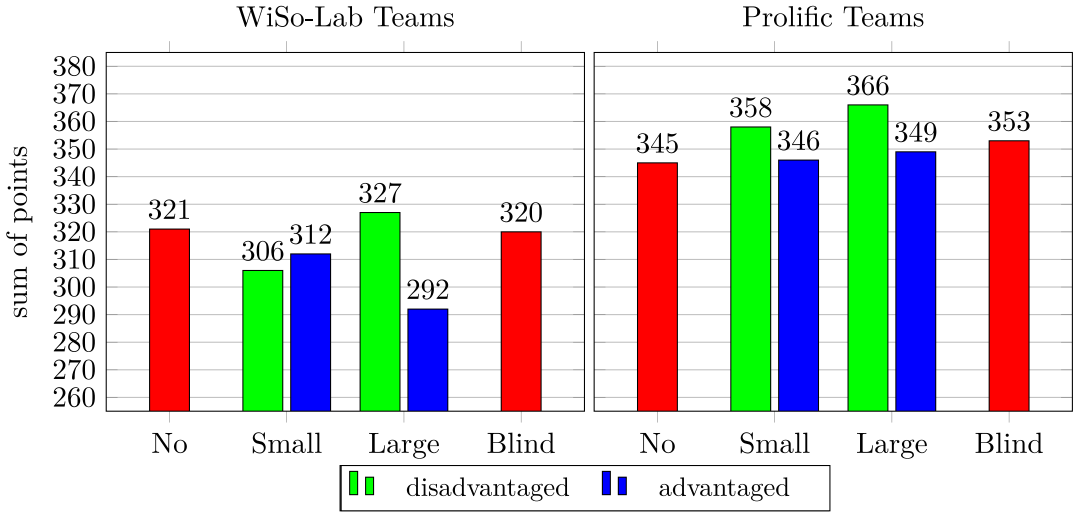


Figure 2.7: **Sum of points for all team treatments split up between WiSo-Lab and Prolific with Prolific pre-study data set.**

The left bar graph displays the average sum of points of teams who participated via the WiSo-Lab Hamburg in all six treatments. The right bar graph displays the average sum of points of teams who participated via Prolific in all six treatments. The red bars display the average sum of points of participants in the ‘no handicap’ and ‘blind handicap’ treatments. The green bars display the average sum of points of *disadvantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments. The blue bars display the average sum of points of *advantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments. All points are calculated using the data set of the Prolific pre-study.

2.B.3 Answer Ranks

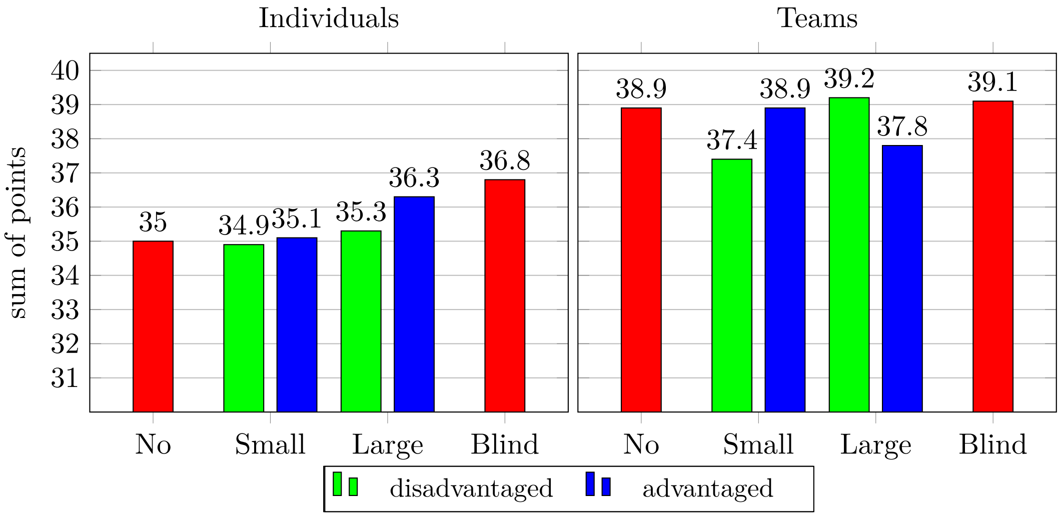


Figure 2.8: **Sum of ranks for all treatments.**

The left bar graph displays the average sum of ranks of individuals in all six treatments. The right bar graph displays the average sum of ranks of teams in all six treatments. The red bars display the average sum of ranks of participants in the ‘no handicap’ and ‘blind handicap’ treatments. The green bars display the average sum of ranks of *disadvantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments. The blue bars display the average sum of ranks of *advantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments.

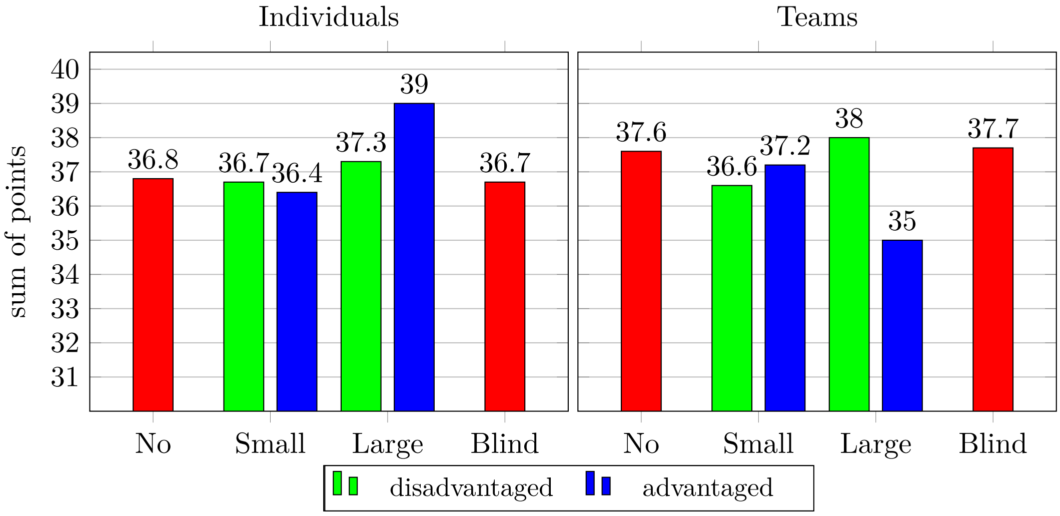


Figure 2.9: **Sum of ranks for all treatments with Prolific pre-study data set.**

The left bar graph displays the average sum of ranks of individuals in all six treatments. The right bar graph displays the average sum of ranks of teams in all six treatments. The red bars display the average sum of ranks of participants in the ‘no handicap’ and ‘blind handicap’ treatments. The green bars display the average sum of ranks of *disadvantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments. The blue bars display the average sum of ranks of *advantaged* participants in the ‘small handicap’ and ‘large handicap’ treatments. All points are calculated using the data set of the Prolific pre-study.

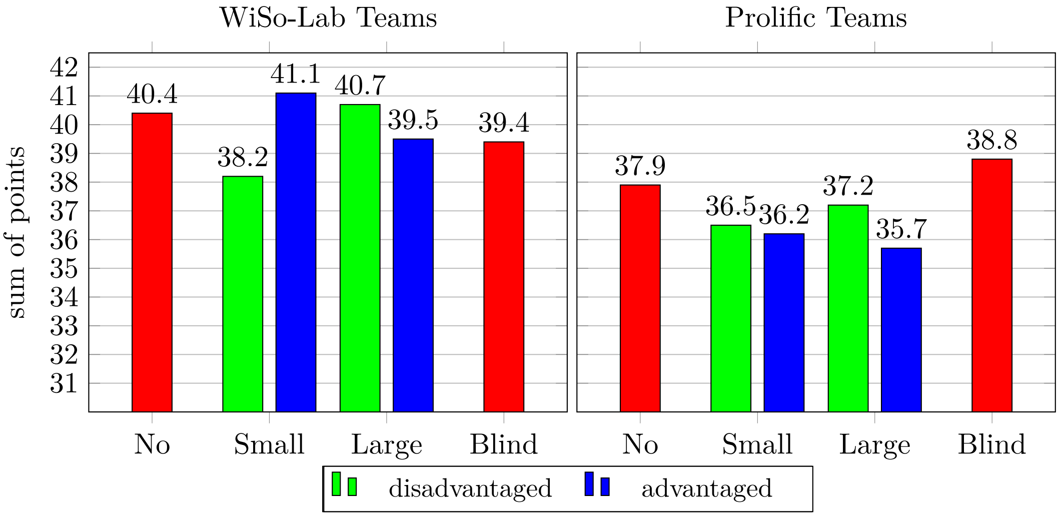


Figure 2.10: **Sum of ranks for all team treatments split up between WiSo-Lab and Prolific.**

The left bar graph displays the average sum of ranks of teams who participated via the WiSo-Lab Hamburg in all six treatments. The right bar graph displays the average sum of ranks of teams who participated via Prolific in all six treatments. The red bars display the average sum of ranks of participants in the 'no handicap' and 'blind handicap' treatments. The green bars display the average sum of ranks of *disadvantaged* participants in the 'small handicap' and 'large handicap' treatments. The blue bars display the average sum of ranks of *advantaged* participants in the 'small handicap' and 'large handicap' treatments.

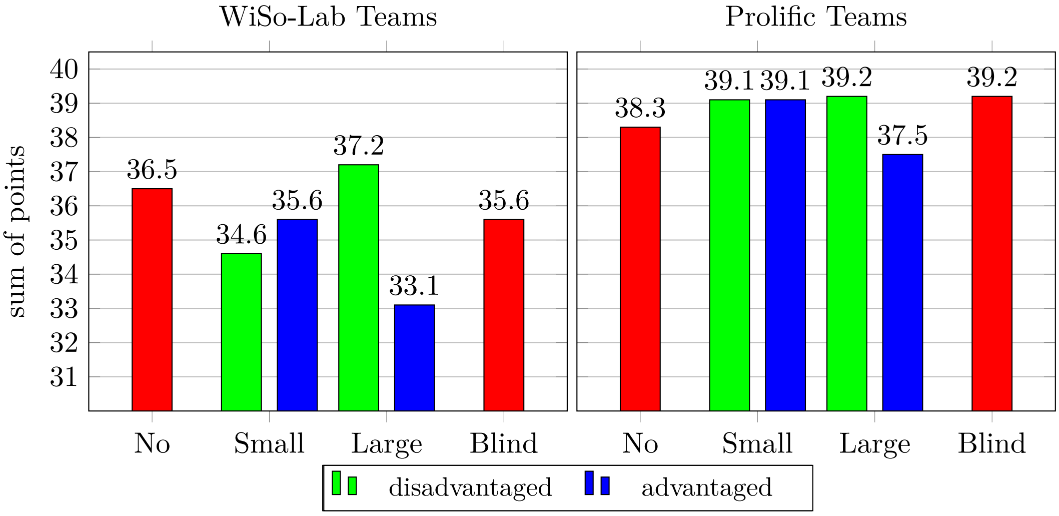


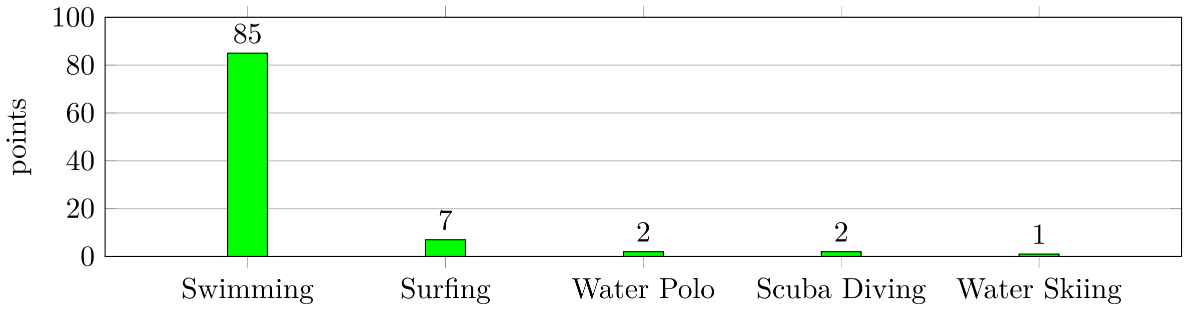
Figure 2.11: **Sum of ranks for all team treatments split up between WiSo-Lab and Prolific with Prolific pre-study data set.**

The left bar graph displays the average sum of ranks of teams who participated via the WiSo-Lab Hamburg in all six treatments. The right bar graph displays the average sum of ranks of teams who participated via Prolific in all six treatments. The red bars display the average sum of ranks of participants in the 'no handicap' and 'blind handicap' treatments. The green bars display the average sum of ranks of *disadvantaged* participants in the 'small handicap' and 'large handicap' treatments. The blue bars display the average sum of ranks of *advantaged* participants in the 'small handicap' and 'large handicap' treatments. All points are calculated using the data set of the Prolific pre-study.

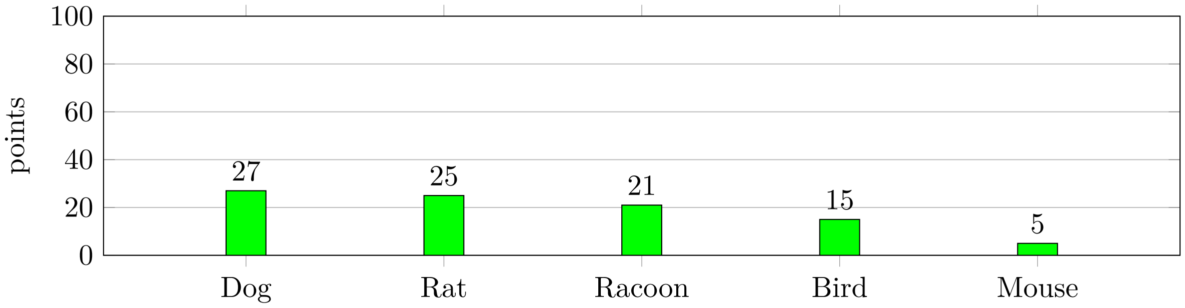
2.C *Family Feud* Game Questions and Answer Distributions

This chapter displays the distribution of answers for all ten questions asked in the main study. The points given for each answer reflect the number of times each answer was given in the original pre-study. The graphs display the five most frequent answers.

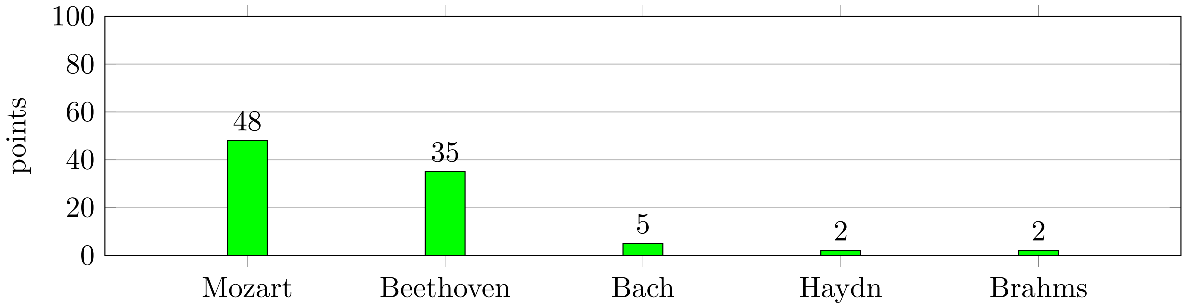
Q1: “Name a bad sport for someone who is afraid of the water.”



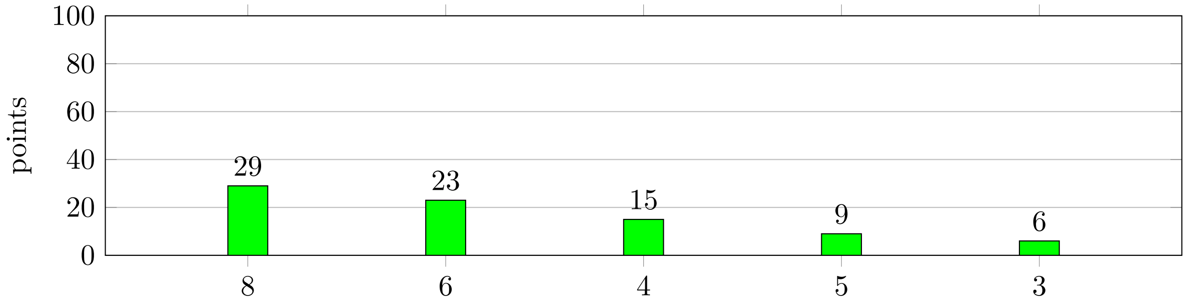
Q2: “If you put food outside for a stray cat, name another animal that might eat it.”



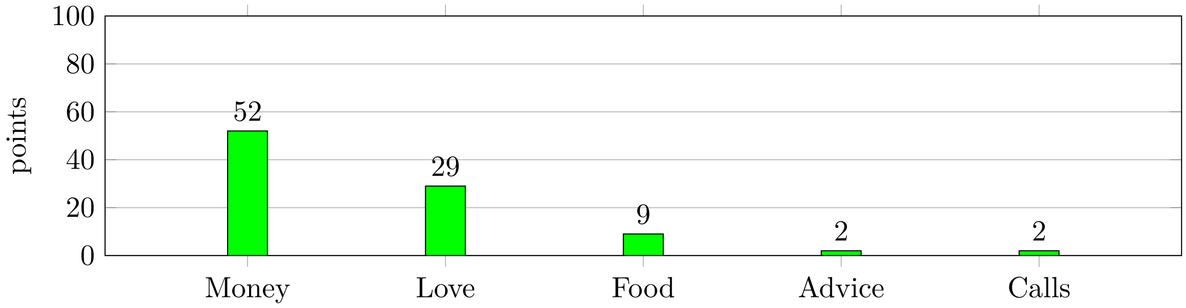
Q3: “Name a famous composer whose music is performed by a symphony orchestra.”



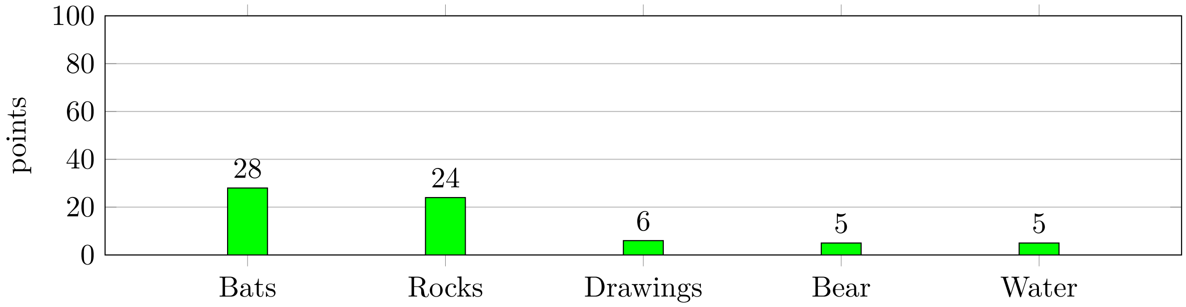
Q4: “Tell me how many glasses of water the average person drinks a day (here: 1 glass = 250ml).”



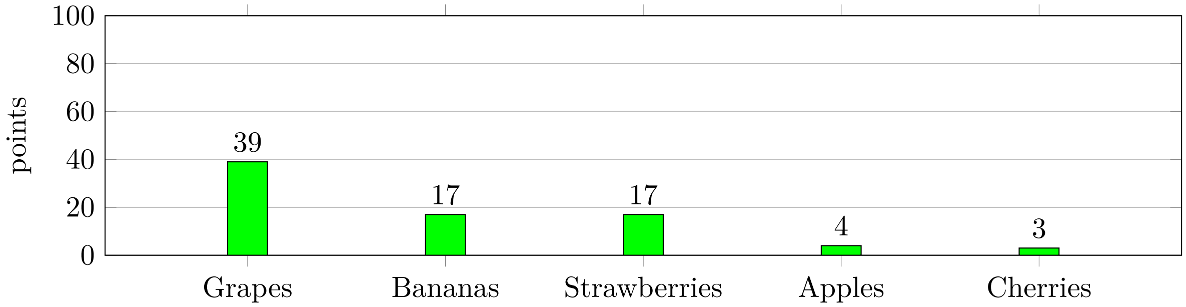
Q5: “Name something your parents still give you.”



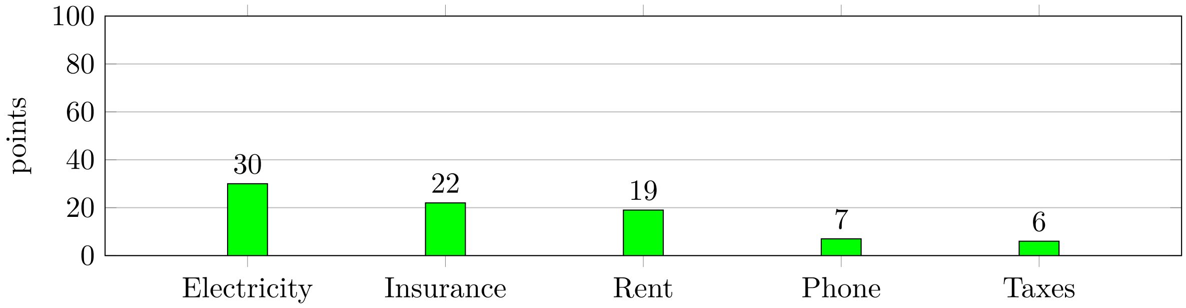
Q6: “Name something you would see in a cave.”



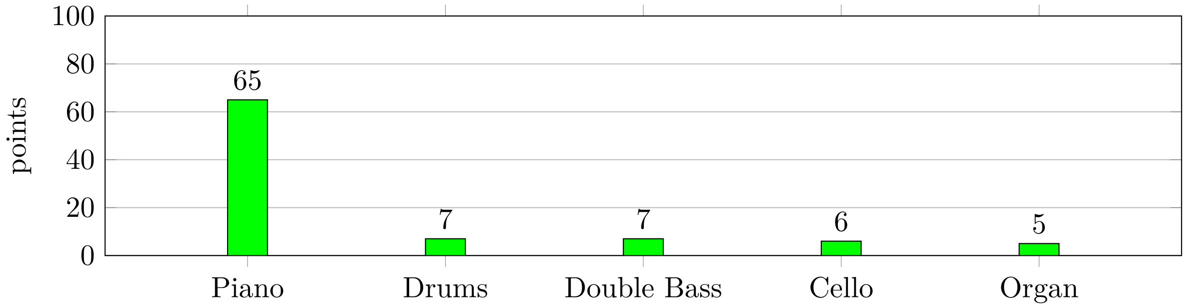
Q7: “Name a fruit you never buy just one of.”



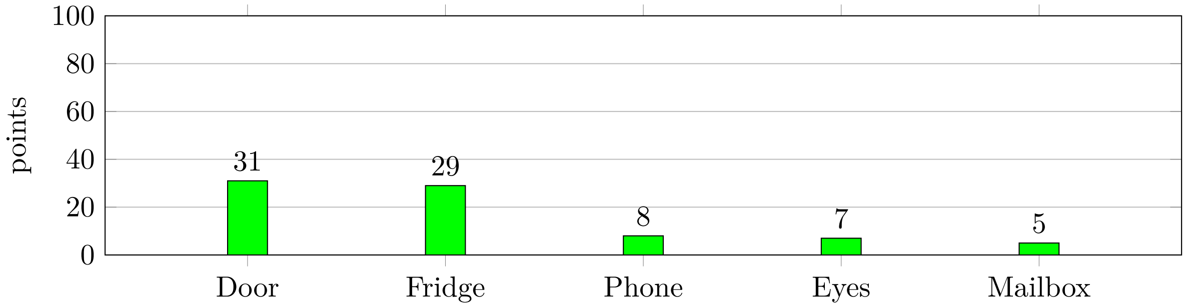
Q8: “Name a bill you’ll be paying every month for the rest of your life.”



Q9: “Name a musical instrument that is too big to carry on an airplane.”

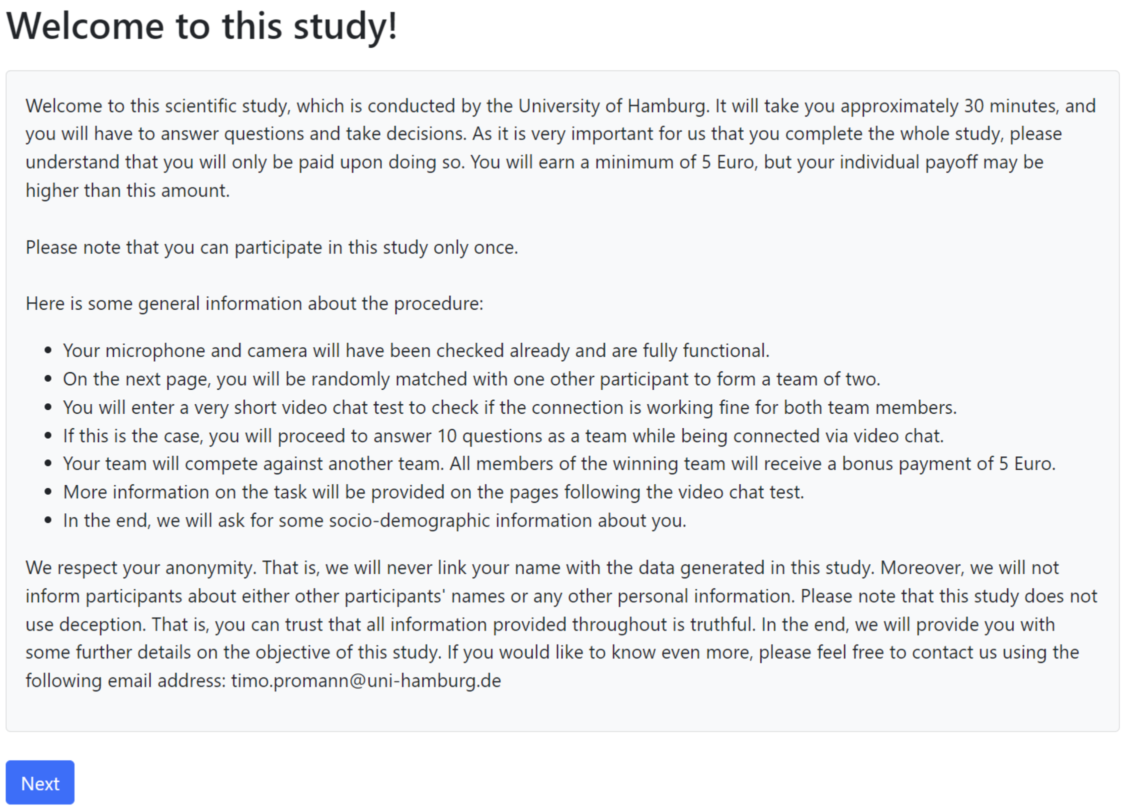


Q10: “Name something most people open at least once a day.”



2.D Experimental Instructions

This chapter features an overview of all major pages in my experimental study. Only dropout pages and waiting pages are not included. Welcome Page, Instructions, Demographics, and Debriefing vary only marginally between individual and team treatments. Hence, these pages are only displayed in the team treatment version.



Welcome to this study!

Welcome to this scientific study, which is conducted by the University of Hamburg. It will take you approximately 30 minutes, and you will have to answer questions and take decisions. As it is very important for us that you complete the whole study, please understand that you will only be paid upon doing so. You will earn a minimum of 5 Euro, but your individual payoff may be higher than this amount.

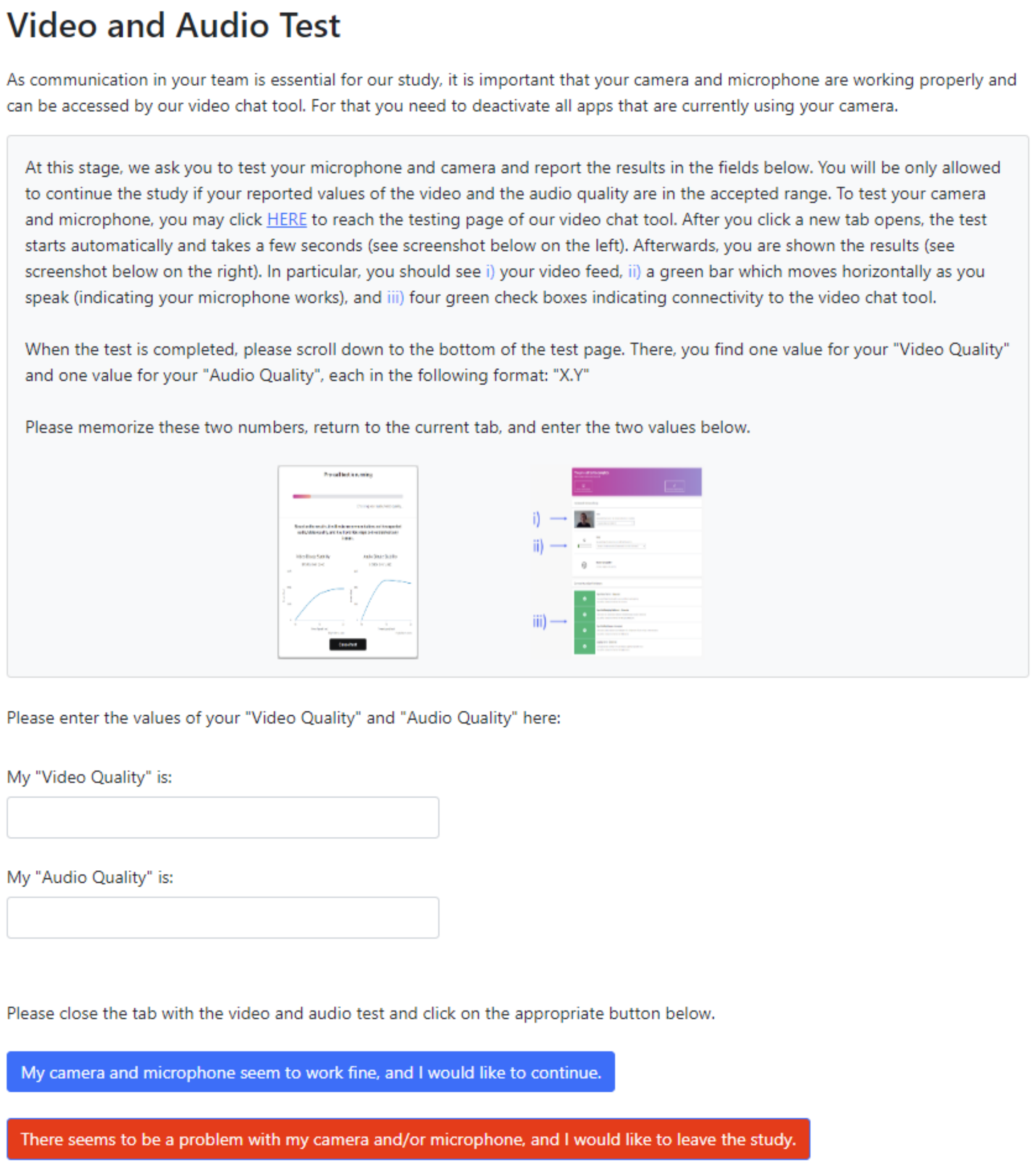
Please note that you can participate in this study only once.

Here is some general information about the procedure:

- Your microphone and camera will have been checked already and are fully functional.
- On the next page, you will be randomly matched with one other participant to form a team of two.
- You will enter a very short video chat test to check if the connection is working fine for both team members.
- If this is the case, you will proceed to answer 10 questions as a team while being connected via video chat.
- Your team will compete against another team. All members of the winning team will receive a bonus payment of 5 Euro.
- More information on the task will be provided on the pages following the video chat test.
- In the end, we will ask for some socio-demographic information about you.

We respect your anonymity. That is, we will never link your name with the data generated in this study. Moreover, we will not inform participants about either other participants' names or any other personal information. Please note that this study does not use deception. That is, you can trust that all information provided throughout is truthful. In the end, we will provide you with some further details on the objective of this study. If you would like to know even more, please feel free to contact us using the following email address: timo.promann@uni-hamburg.de

Figure 2.12: **Welcome Page.**



Video and Audio Test

As communication in your team is essential for our study, it is important that your camera and microphone are working properly and can be accessed by our video chat tool. For that you need to deactivate all apps that are currently using your camera.

At this stage, we ask you to test your microphone and camera and report the results in the fields below. You will be only allowed to continue the study if your reported values of the video and the audio quality are in the accepted range. To test your camera and microphone, you may click HERE to reach the testing page of our video chat tool. After you click a new tab opens, the test starts automatically and takes a few seconds (see screenshot below on the left). Afterwards, you are shown the results (see screenshot below on the right). In particular, you should see i) your video feed, ii) a green bar which moves horizontally as you speak (indicating your microphone works), and iii) four green check boxes indicating connectivity to the video chat tool.

When the test is completed, please scroll down to the bottom of the test page. There, you find one value for your "Video Quality" and one value for your "Audio Quality", each in the following format: "X.Y"

Please memorize these two numbers, return to the current tab, and enter the two values below.

Please enter the values of your "Video Quality" and "Audio Quality" here:

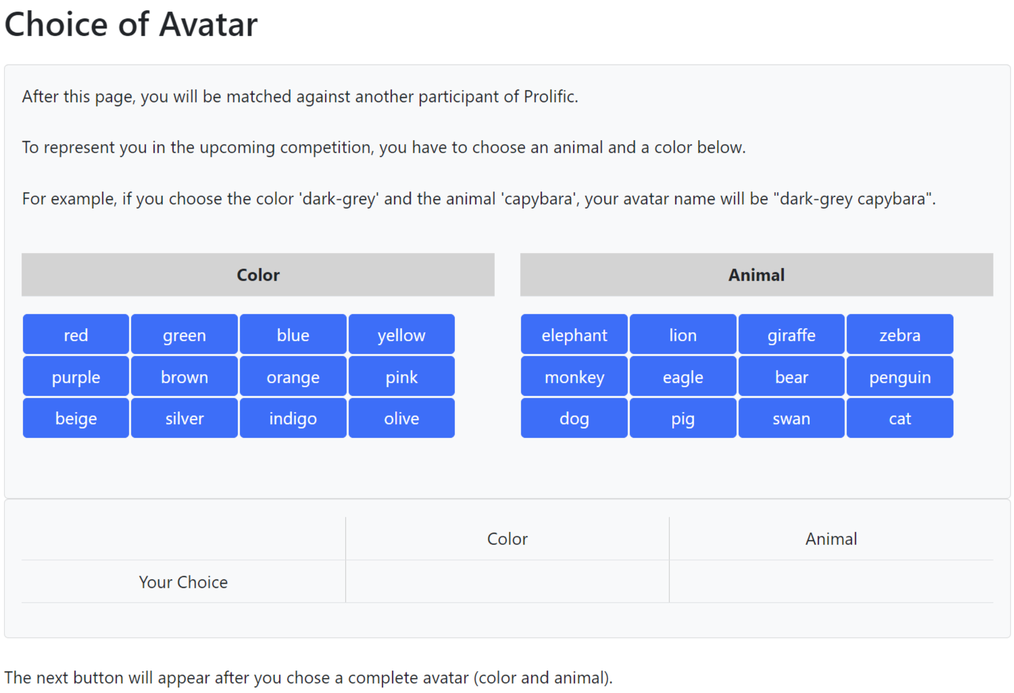
My "Video Quality" is:

My "Audio Quality" is:

Please close the tab with the video and audio test and click on the appropriate button below.

My camera and microphone seem to work fine, and I would like to continue.
There seems to be a problem with my camera and/or microphone, and I would like to leave the study.

Figure 2.13: **Video & Audio Test for Team Treatments.**



Choice of Avatar

After this page, you will be matched against another participant of Prolific.

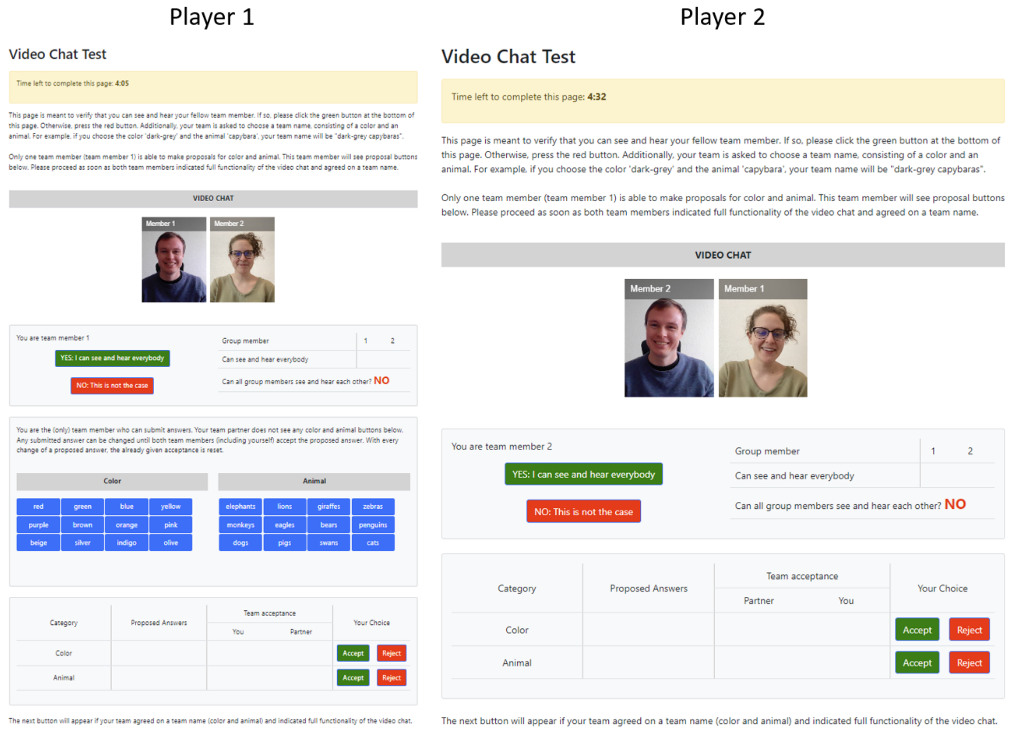
To represent you in the upcoming competition, you have to choose an animal and a color below.

For example, if you choose the color 'dark-grey' and the animal 'capybara', your avatar name will be "dark-grey capybara".

	Color	Animal
Your Choice		

The next button will appear after you chose a complete avatar (color and animal).

Figure 2.14: **Choice of Avatar - Individuals.**



Player 1

Video Chat Test

Time left to complete this page: **4:05**

This page is meant to verify that you can see and hear your fellow team member. If so, please click the green button at the bottom of this page. Otherwise, press the red button. Additionally, your team is asked to choose a team name, consisting of a color and an animal. For example, if you choose the color 'dark-grey' and the animal 'capybara', your team name will be "dark-grey capybaras".

Only one team member (team member 1) is able to make proposals for color and animal. This team member will see proposal buttons below. Please proceed as soon as both team members indicated full functionality of the video chat and agreed on a team name.

You are team member 1

Group member	1	2
Can see and hear everybody		
Can all group members see and hear each other? **NO**		

You are the (only) team member who can submit answers. Your team partner does not see any color and animal buttons below. Any submitted answer can be changed until both team members (including yourself) accept the proposed answer. With every change of a proposed answer, the already given acceptance is reset.

Category	Proposed Answers	Team acceptance		Your Choice	
		You	Partner		
Color				Accept	Reject
Animal				Accept	Reject

The next button will appear if your team agreed on a team name (color and animal) and indicated full functionality of the video chat.

Player 2

Video Chat Test

Time left to complete this page: **4:32**

This page is meant to verify that you can see and hear your fellow team member. If so, please click the green button at the bottom of this page. Otherwise, press the red button. Additionally, your team is asked to choose a team name, consisting of a color and an animal. For example, if you choose the color 'dark-grey' and the animal 'capybara', your team name will be "dark-grey capybaras".

Only one team member (team member 1) is able to make proposals for color and animal. This team member will see proposal buttons below. Please proceed as soon as both team members indicated full functionality of the video chat and agreed on a team name.

You are team member 2

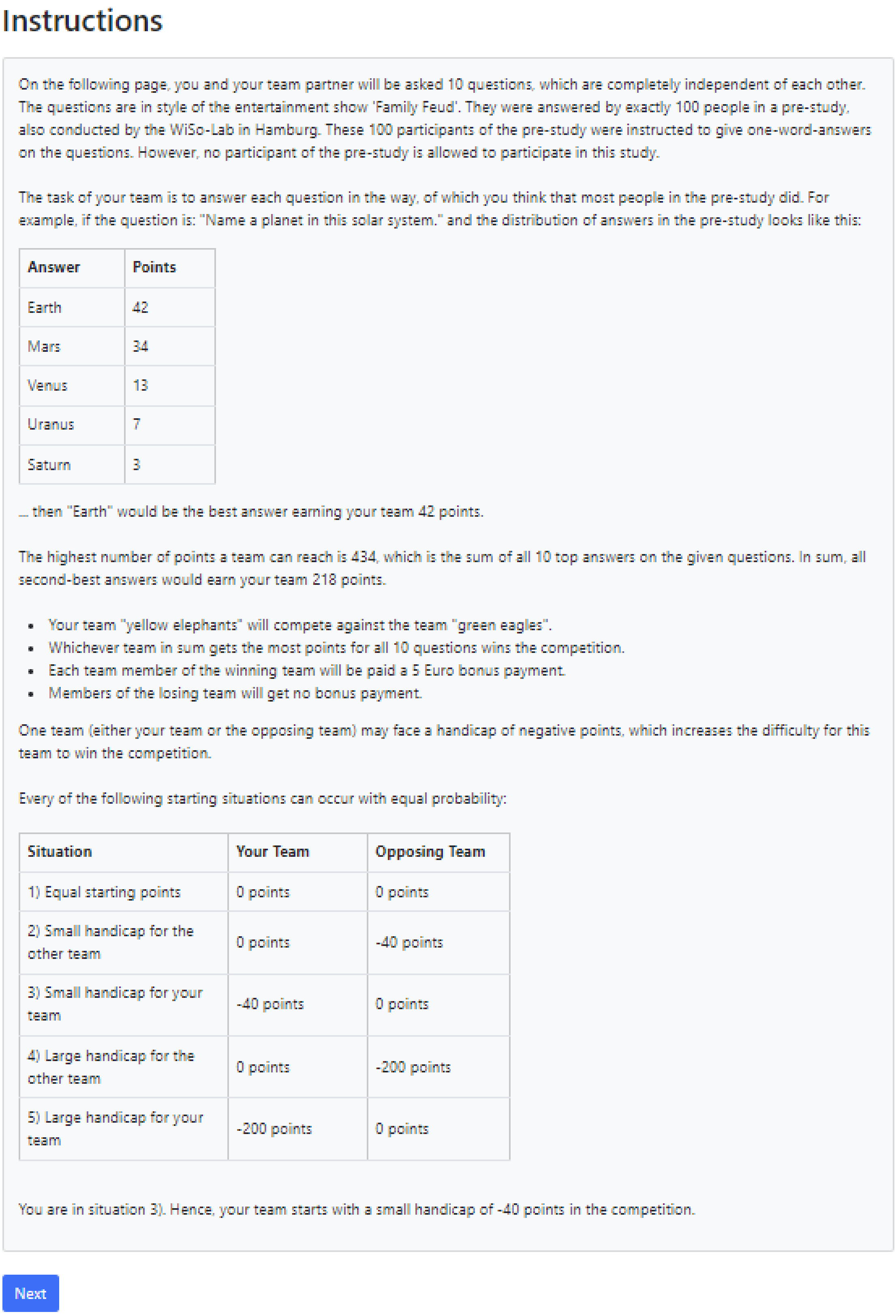
Group member	1	2
Can see and hear everybody		
Can all group members see and hear each other? **NO**		

Category	Proposed Answers	Team acceptance		Your Choice	
		Partner	You		
Color				Accept	Reject
Animal				Accept	Reject

The next button will appear if your team agreed on a team name (color and animal) and indicated full functionality of the video chat.

Figure 2.15: **Choice of Avatar - Teams.**

Participants in the team treatments also had to indicate on this page whether they could see and hear their team partner. Only player 1 was able to propose a color and an animal for the choice of their avatar, but player 2 could always reject the proposal.



Instructions

On the following page, you and your team partner will be asked 10 questions, which are completely independent of each other. The questions are in style of the entertainment show 'Family Feud'. They were answered by exactly 100 people in a pre-study, also conducted by the WiSo-Lab in Hamburg. These 100 participants of the pre-study were instructed to give one-word-answers on the questions. However, no participant of the pre-study is allowed to participate in this study.

The task of your team is to answer each question in the way, of which you think that most people in the pre-study did. For example, if the question is: "Name a planet in this solar system." and the distribution of answers in the pre-study looks like this:

Answer	Points
Earth	42
Mars	34
Venus	13
Uranus	7
Saturn	3

... then "Earth" would be the best answer earning your team 42 points.

The highest number of points a team can reach is 434, which is the sum of all 10 top answers on the given questions. In sum, all second-best answers would earn your team 218 points.

- Your team "yellow elephants" will compete against the team "green eagles".
- Whichever team in sum gets the most points for all 10 questions wins the competition.
- Each team member of the winning team will be paid a 5 Euro bonus payment.
- Members of the losing team will get no bonus payment.

One team (either your team or the opposing team) may face a handicap of negative points, which increases the difficulty for this team to win the competition.

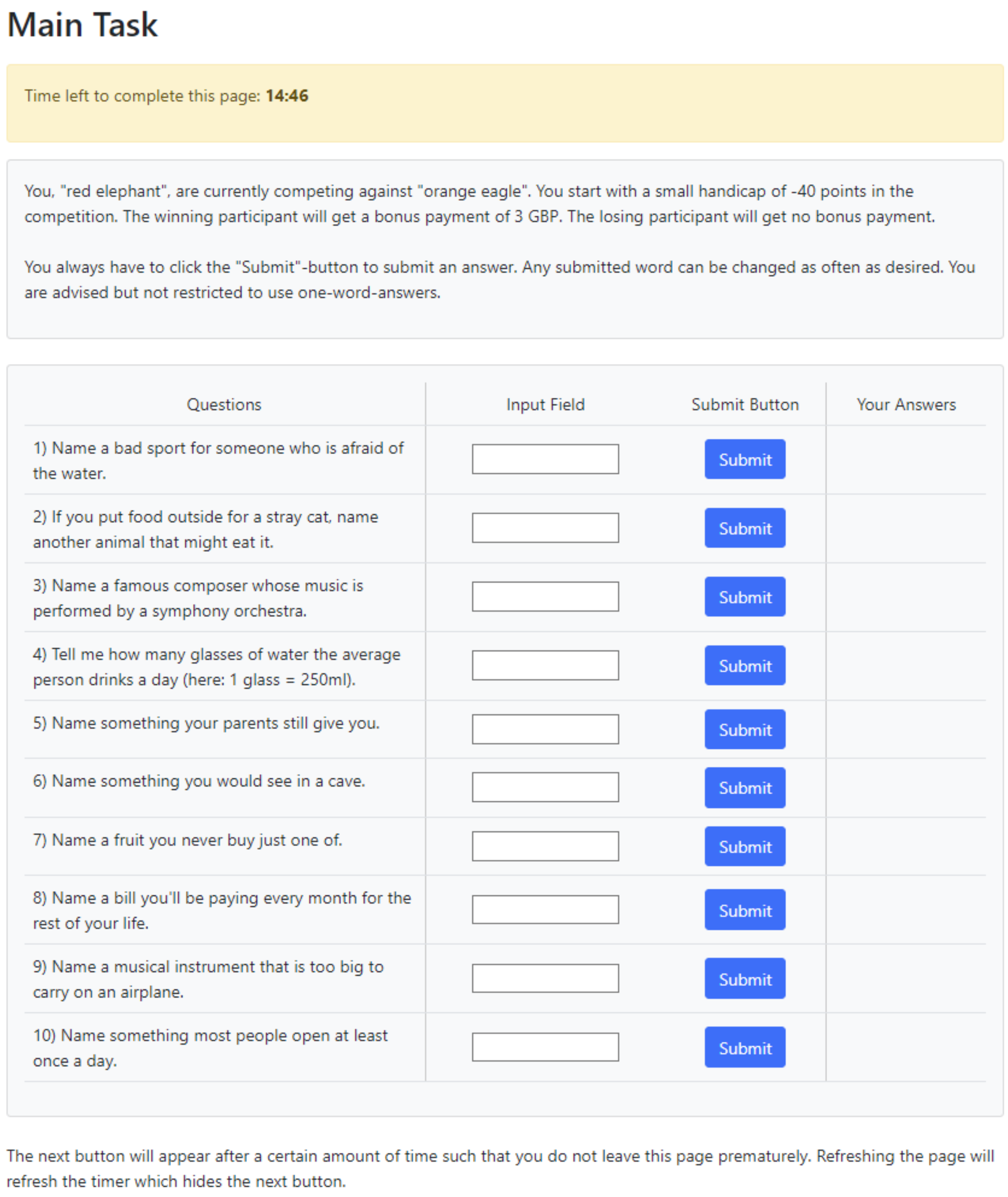
Every of the following starting situations can occur with equal probability:

Situation	Your Team	Opposing Team
1) Equal starting points	0 points	0 points
2) Small handicap for the other team	0 points	-40 points
3) Small handicap for your team	-40 points	0 points
4) Large handicap for the other team	0 points	-200 points
5) Large handicap for your team	-200 points	0 points

You are in situation 3). Hence, your team starts with a small handicap of -40 points in the competition.

Figure 2.16: **Instructions.**

The last sentence on this page varied depending on the treatment.



Main Task

Time left to complete this page: **14:46**

You, "red elephant", are currently competing against "orange eagle". You start with a small handicap of -40 points in the competition. The winning participant will get a bonus payment of 3 GBP. The losing participant will get no bonus payment.

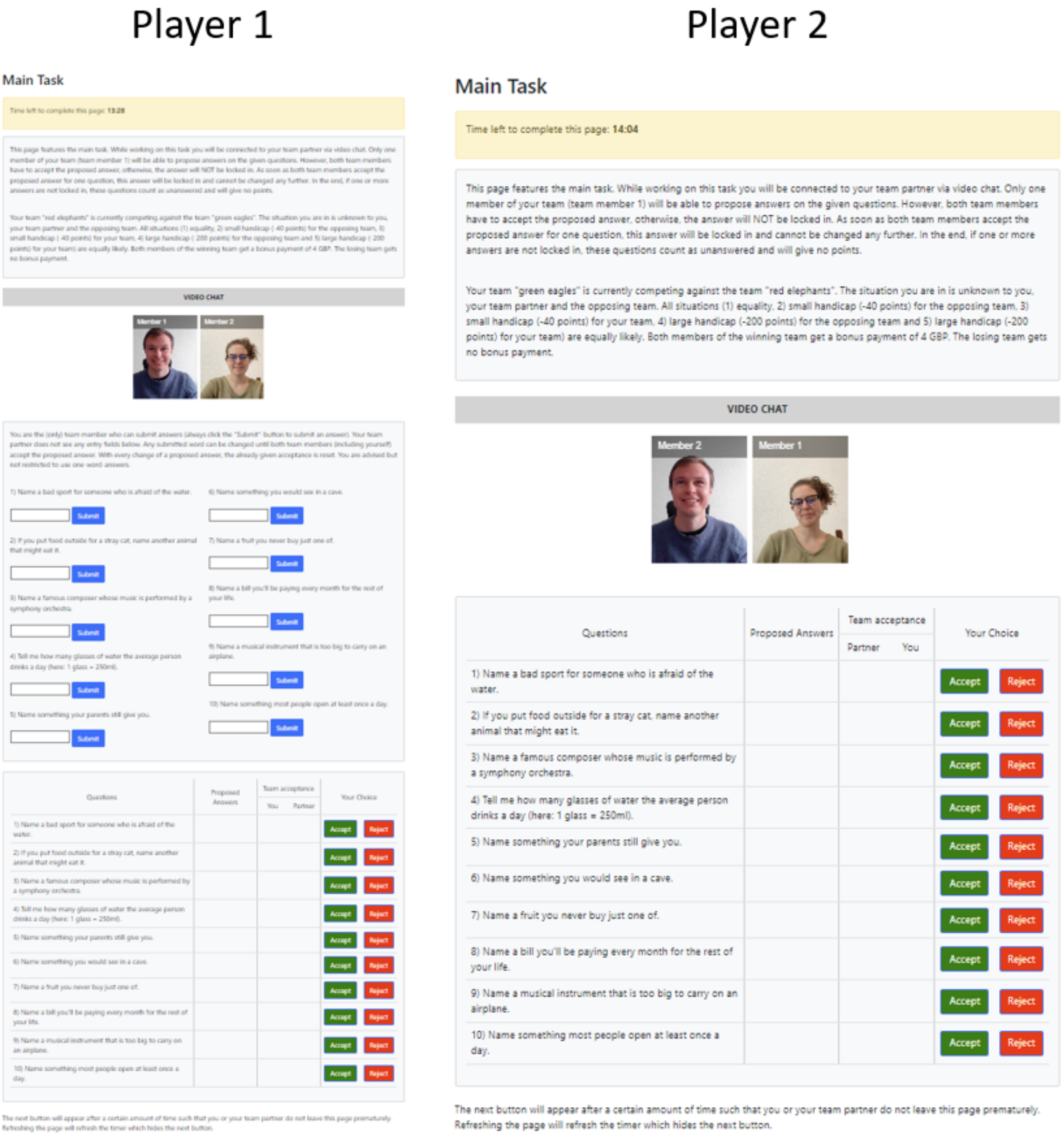
You always have to click the "Submit"-button to submit an answer. Any submitted word can be changed as often as desired. You are advised but not restricted to use one-word-answers.

Questions	Input Field	Submit Button	Your Answers
1) Name a bad sport for someone who is afraid of the water.		Submit	
2) If you put food outside for a stray cat, name another animal that might eat it.		Submit	
3) Name a famous composer whose music is performed by a symphony orchestra.		Submit	
4) Tell me how many glasses of water the average person drinks a day (here: 1 glass = 250ml).		Submit	
5) Name something your parents still give you.		Submit	
6) Name something you would see in a cave.		Submit	
7) Name a fruit you never buy just one of.		Submit	
8) Name a bill you'll be paying every month for the rest of your life.		Submit	
9) Name a musical instrument that is too big to carry on an airplane.		Submit	
10) Name something most people open at least once a day.		Submit	

The next button will appear after a certain amount of time such that you do not leave this page prematurely. Refreshing the page will refresh the timer which hides the next button.

Figure 2.17: **Main Task - Individuals.**

The information concerning the given treatment, displayed on the top of this page, varied between treatments.



Player 1

Main Task

Time left to complete this page: **13:28**

This page features the main task. While working on this task you will be connected to your team partner via video chat. Only one member of your team (team member 1) will be able to propose answers on the given questions. However, both team members have to accept the proposed answer, otherwise, the answer will NOT be locked in. As soon as both team members accept the proposed answer for one question, this answer will be locked in and cannot be changed any further. In the end, if one or more answers are not locked in, these questions count as unanswered and will give no points.

Your team "red elephants" is currently competing against the team "green eagles". The situation you are in is unknown to you, your team partner and the opposing team. All situations (1) equality, 2) small handicap (40 points) for the opposing team, 3) small handicap (40 points) for your team, 4) large handicap (200 points) for the opposing team and 5) large handicap (200 points) for your team) are equally likely. Both members of the winning team get a bonus payment of 4 GBP. The losing team gets no bonus payment.

VIDEO CHAT

You are the (only) team member who can submit answers (always click the "Submit" button to submit an answer). Your team partner does not see any entry fields below. Any submitted word can be changed until both team members (including yourself) accept the proposed answer. With every change of a proposed answer, the already given acceptance is reset. You are advised but not restricted to use one word answers.

1) Name a bad sport for someone who is afraid of the water.

Submit

2) If you put food outside for a stray cat, name another animal that might eat it.

Submit

3) Name a famous composer whose music is performed by a symphony orchestra.

Submit

4) Tell me how many glasses of water the average person drinks a day (here: 1 glass = 250ml).

Submit

5) Name something your parents still give you.

Submit

6) Name something you would see in a cave.

Submit

7) Name a fruit you never buy just one of.

Submit

8) Name a bill you'll be paying every month for the rest of your life.

Submit

9) Name a musical instrument that is too big to carry on an airplane.

Submit

10) Name something most people open at least once a day.

Submit

Questions	Proposed Answers	Team acceptance		Your Choice	
		You	Partner		
1) Name a bad sport for someone who is afraid of the water.				Accept	Reject
2) If you put food outside for a stray cat, name another animal that might eat it.				Accept	Reject
3) Name a famous composer whose music is performed by a symphony orchestra.				Accept	Reject
4) Tell me how many glasses of water the average person drinks a day (here: 1 glass = 250ml).				Accept	Reject
5) Name something your parents still give you.				Accept	Reject
6) Name something you would see in a cave.				Accept	Reject
7) Name a fruit you never buy just one of.				Accept	Reject
8) Name a bill you'll be paying every month for the rest of your life.				Accept	Reject
9) Name a musical instrument that is too big to carry on an airplane.				Accept	Reject
10) Name something most people open at least once a day.				Accept	Reject

The next button will appear after a certain amount of time such that you or your team partner do not leave this page prematurely. Refreshing the page will refresh the timer which hides the next button.

Player 2

Main Task

Time left to complete this page: **14:04**

This page features the main task. While working on this task you will be connected to your team partner via video chat. Only one member of your team (team member 1) will be able to propose answers on the given questions. However, both team members have to accept the proposed answer, otherwise, the answer will NOT be locked in. As soon as both team members accept the proposed answer for one question, this answer will be locked in and cannot be changed any further. In the end, if one or more answers are not locked in, these questions count as unanswered and will give no points.

Your team "green eagles" is currently competing against the team "red elephants". The situation you are in is unknown to you, your team partner and the opposing team. All situations (1) equality, 2) small handicap (-40 points) for the opposing team, 3) small handicap (-40 points) for your team, 4) large handicap (-200 points) for the opposing team and 5) large handicap (-200 points) for your team) are equally likely. Both members of the winning team get a bonus payment of 4 GBP. The losing team gets no bonus payment.

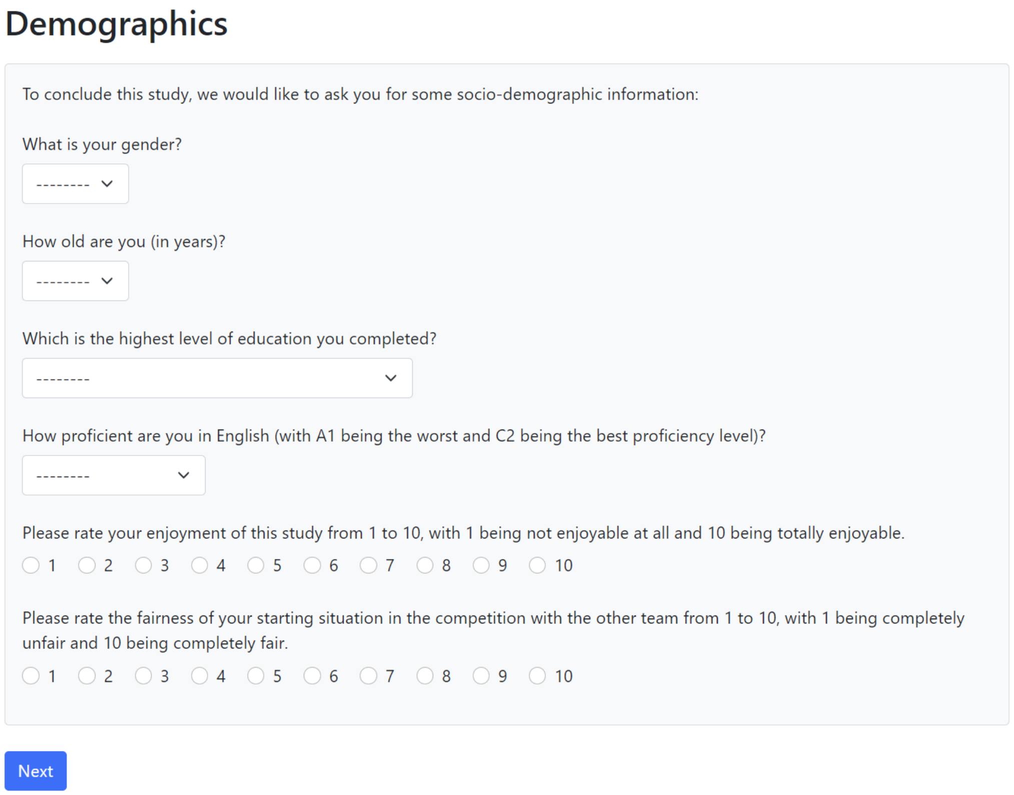
VIDEO CHAT

Questions	Proposed Answers	Team acceptance		Your Choice	
		Partner	You		
1) Name a bad sport for someone who is afraid of the water.				Accept	Reject
2) If you put food outside for a stray cat, name another animal that might eat it.				Accept	Reject
3) Name a famous composer whose music is performed by a symphony orchestra.				Accept	Reject
4) Tell me how many glasses of water the average person drinks a day (here: 1 glass = 250ml).				Accept	Reject
5) Name something your parents still give you.				Accept	Reject
6) Name something you would see in a cave.				Accept	Reject
7) Name a fruit you never buy just one of.				Accept	Reject
8) Name a bill you'll be paying every month for the rest of your life.				Accept	Reject
9) Name a musical instrument that is too big to carry on an airplane.				Accept	Reject
10) Name something most people open at least once a day.				Accept	Reject

The next button will appear after a certain amount of time such that you or your team partner do not leave this page prematurely. Refreshing the page will refresh the timer which hides the next button.

Figure 2.18: **Main Task - Teams.**

Only player 1 was able to enter answers, but player 2 could always reject the proposed answer. The information concerning the given treatment, displayed on the top of this page, varied between treatments.



Demographics

To conclude this study, we would like to ask you for some socio-demographic information:

What is your gender?

-------- ⌄

How old are you (in years)?

-------- ⌄

Which is the highest level of education you completed?

-------- ⌄

How proficient are you in English (with A1 being the worst and C2 being the best proficiency level)?

-------- ⌄

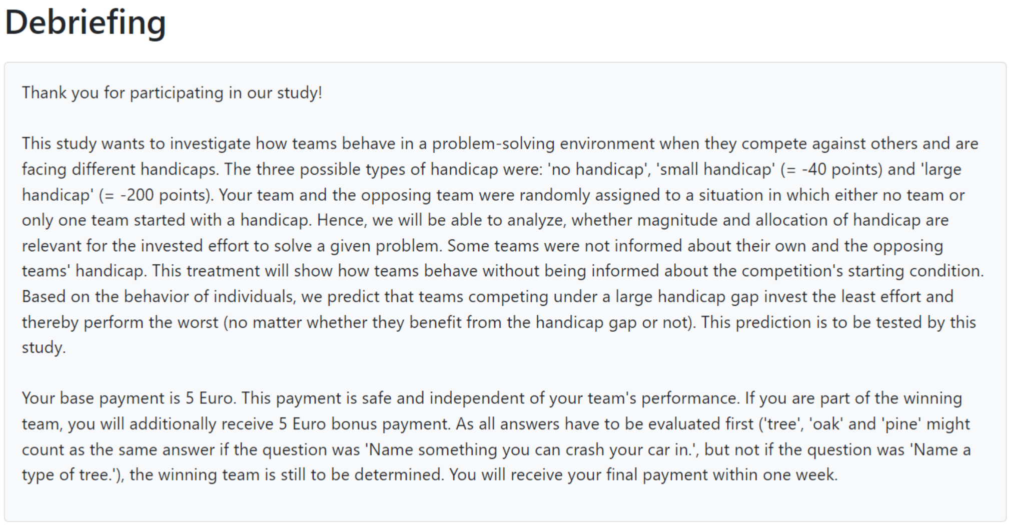
Please rate your enjoyment of this study from 1 to 10, with 1 being not enjoyable at all and 10 being totally enjoyable.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Please rate the fairness of your starting situation in the competition with the other team from 1 to 10, with 1 being completely unfair and 10 being completely fair.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Figure 2.19: **Demographics.**



Debriefing

Thank you for participating in our study!

This study wants to investigate how teams behave in a problem-solving environment when they compete against others and are facing different handicaps. The three possible types of handicap were: 'no handicap', 'small handicap' (= -40 points) and 'large handicap' (= -200 points). Your team and the opposing team were randomly assigned to a situation in which either no team or only one team started with a handicap. Hence, we will be able to analyze, whether magnitude and allocation of handicap are relevant for the invested effort to solve a given problem. Some teams were not informed about their own and the opposing teams' handicap. This treatment will show how teams behave without being informed about the competition's starting condition. Based on the behavior of individuals, we predict that teams competing under a large handicap gap invest the least effort and thereby perform the worst (no matter whether they benefit from the handicap gap or not). This prediction is to be tested by this study.

Your base payment is 5 Euro. This payment is safe and independent of your team's performance. If you are part of the winning team, you will additionally receive 5 Euro bonus payment. As all answers have to be evaluated first ('tree', 'oak' and 'pine' might count as the same answer if the question was 'Name something you can crash your car in.', but not if the question was 'Name a type of tree.'), the winning team is still to be determined. You will receive your final payment within one week.

Figure 2.20: **Debriefing.**

Honesty of groups: Effects of size and gender composition[1]

Abstract

This paper studies unethical behavior by groups and provides systematic evidence on how lying decisions are affected by group size and group gender composition. We conduct an online experiment with 1,677 participants (477 groups) where group members can communicate with each other via a novel video chat tool. Our key findings are that (i) larger groups lie more, (ii) all-male groups stand out in their proclivity to lie, (iii) already the first female in a group causes an honesty shift, and (iv) group behavior cannot be fully explained by members' individual honesty preferences.

Keywords: Group decisions, unethical behavior, lying, gender differences, online experiment, group video chat

JEL Classification: C92, J16, D70

3.1 Introduction

Motivation Many decisions are taken by groups rather than individuals. This paper considers the domain of unethical behavior (lying) and provides systematic evidence on how group decisions are affected by the size and gender composition of the group. In an online experiment with 18 treatments, we consider all group sizes up to five members and all possible male-female gender compositions. Our main findings are that (i) larger groups lie more, (ii) all-male groups stand out in their proclivity to lie, (iii) already the first female in a group causes a substantial shift towards more honest group behavior, and (iv) group behavior cannot be fully explained by members' individual honesty preferences. We also study additional issues such as decision times, talking times and the role of personal characteristics. As a methodological contribution, we have implemented a novel video chat tool that allows participants to interact face-to-face in online experiments.

There exist many settings in which non-trivial or even complex decisions are taken by groups rather than individuals. Examples abound, not only for groups of friends or family, but also in firms where boards and committees jointly take crucial managerial decisions and where (agile) project teams are constantly built anew and restructured depending on the tasks ahead. According to Lazear and Shaw (2007), 80% of large US firms rely on self-managed work teams. The prevalence of group decisions raises various interesting questions. For example, how does decision-making by groups depend on their size and composition.

[1] This chapter is co-authored by Gerd Muehlheusser, Andreas Roider and Niklas Wallmeier.

The answers to these questions might very well also depend on the type of decision, i.e. whether a given group task is *intellective* or *judgmental* in nature (see e.g. Kugler et al., 2012). Intellective decisions (such as solving complex, possibly non-routine problems) mainly rely on the group members' cognitive skills and effort, and value a performance of high quality (for a recent field experiment with groups, see e.g. Englmaier et al., 2024). By contrast, decisions that are predominantly judgmental rather reflect a group preference. One important domain of judgmental decisions is unethical (or even straightforward illegal) behavior.[2] This might range from the small (such as sugarcoating a report to a superior) to the large (as exemplified by recent high-profile corporate scandals at Volkswagen, Enron, or Worldcom, where small groups of executives and/or employees were instrumental). In fact, corporate fraud is seen by many as a topical issue and major challenge. For example, in the "Global Fraud Report" (Kroll, 2016) 75% of surveyed senior executives state that their company had become a fraud victim in the previous year. Using a natural experiment, Dyck, Morse, and Zingales (2023) find the average cost of both detected and undetected fraud in large U.S. corporations in the period 19962004 to be $360 billion per year. On a global scale, the Association of Certified Fraud Examiners (2022) estimates that the average loss of organizations due to fraud (including financial statement fraud, asset misappropriation, and corruption) amounts to 5% of annual revenues. In the light of these findings, it seems important to gain a sound understanding of the circumstances that make (groups of) decision-makers prone to unethical behavior.

A substantial body of economic literature has experimentally studied unethical behavior by individuals (for a comprehensive overview, see e.g. Abeler, Nosenzo, and Raymond, 2019). Moreover, there also exists a growing number of studies investigating unethical behavior by (small) groups. This literature (discussed in more detail below) has documented that in groups of two (*dyads*) or three (*triads*), there tends to be more unethical behavior compared to individual decision-making (for the case of triads, see e.g. Kocher, Schudy, and Spantig, 2018).

To the best of our knowledge, there is so far no systematic evidence (i.e. within the same study) whether the observed effects for dyads and triads extend to larger groups (and potentially are even amplified). Yet, larger group sizes are empirically relevant. For example, management practitioners recommend team sizes of 4 to 6 members (see e.g., Useem, 2006 and Thompson, 2017, p.32)., and empirical studies of top management teams in the U.S. find the average team size to be around 3.4 with standard deviations of 1.2-1.5 (see e.g. Haleblian and Finkelstein, 1993; Amason and Sapienza, 1997).[3] Moreover, according to estimates by the Association of Certified Fraud Examiners (2022), groups of three or more perpetrators are responsible for 38%

[2] Further examples of judgemental decisions are the evaluation of risks when deciding about an investment project, the weight placed on fairness concerns in interactions with employees or customers, or which type of candidate to hire for a particular position.

[3] See also Economist (2020), which discusses (optimal) group sizes in a variety of contexts.

of cases of corporate fraud (while the respective numbers for dyads and for individuals are 20% and 42%).

Against this background, the first aim of this paper is to provide systematic evidence on how the extent of unethical behavior varies with *group size*. We believe that there exists a major lacuna as, due to countervailing forces, the effect of group size on unethical behavior seems unclear a priori. For example, while a greater "diffusion of moral responsibility" might lead to more unethical behavior in larger groups, a potential amplification of "image concerns" could lead to the opposite effect.[4]

The second aim of this paper is to analyze the role of the *group gender composition*, a highly topical issue in both the academic and public arena.[5] For the domain of unethical decisions, various studies have experimentally investigated gender differences at the individual level (see e.g., Dreber and Johannesson, 2008; Erat and Gneezy, 2012).[6] Overall, males seem to be somewhat more prone to unethical behavior than females.[7] Much less is known about how (unethical) group decisions are affected by the group gender composition. One exception is Muehlheusser, Roider, and Wallmeier (2015) who document that all-male dyads lie more than all-female dyads, while in groups consisting of one male and one female lying is at an intermediate level (though closer to all-male groups). However, even given this finding, it seems unclear how the group gender composition affects behavior in larger groups. This is the case because in dyads the share of females in the group can take on three values only: 0, 1, and 1/2, where the latter value represents the only mixed dyad. Hence, if the behavior of mixed dyads differs from that of all-male or all-female dyads, this might be due to either a balanced group gender composition or the fact that there is *one* female or *one* male in the group. Obviously, each of these channels might lead to a different prediction for the effect of the group gender composition in larger groups.

Framework To the best of our knowledge, in the economics literature there does not yet exist a systematic analysis (i.e. within the same study) of the effects of group size and the group gender composition on unethical decision-making by groups. In this paper, we present results

[4] Evidence on the former (in the context of delegation) is provided by Bartling and Fischbacher (2012). In their survey, Abeler et al. (2019) document that image concerns (i.e., the desire to be perceived as honest) are a key driver of individual lying behavior (see also Bénabou and Tirole, 2006).

[5] For example, the role of the group gender composition is highlighted by a recent (successful) influence campaign by the "Big Three" asset managers in the US to increase the female representation on corporate boards (Gormley et al., 2023).

[6] While reality is more complex, most of the literature on gender effects focusses on potential behavioral differences between females and males. Note that, in our experiment, all 1677 subjects identified as either female or male. See also the surveys by Croson and Gneezy (2009), Bertrand (2010), Azmat and Petrongolo (2015), and Niederle (2016) who report on gender differences between males and females at the individual level with respect to e.g. time preferences, risk preferences, and social concerns.

[7] According to a survey by the Association of Certified Fraud Examiners (2022), 73% of cases of corporate fraud are committed by men and 27% by women, where these proportions are relatively stable over hierarchy levels, i.e., for employees, managers, and owners / executives.

of an online experiment that aims at providing evidence on these issues. As a methodological contribution, we have implemented a novel video chat tool that allows participants to interact face-to-face in a group setting.

In the key part of the experiment, subjects are matched into groups.[8] We adapt the die-roll paradigm of Fischbacher and Föllmi-Heusi (2013) to a group setting: All group members observe the same roll of a six-sided die, and as a group they are asked to report the outcome to the experimenter. Payoffs depend only on the group report, and the monetary incentive structure is such that, unless the number rolled is 5, a group can increase individual payoffs by reporting a different number. However, groups obtain a payoff only upon reaching an agreement on which number to report. They are given ample time to discuss their report using the video chat. If a group fails to reach an agreement at the end of the discussion period, the payoff from the group task is zero for each group member.[9]

In our experiment, we consider all group sizes from two to five group members, and for each group size we systematically vary the group gender composition between male and female subjects (i.e. for a group size of two, we consider groups with 0, 1, and 2 females, for a group size of three, we consider groups with 0, 1, 2, and 3 females, and so on). This leads to a total of 18 treatments.

Results A first set of results relates to the impact of group size on the prevalence of lying. We find that larger groups lie more (Result 1). In particular, the fraction of groups of 5 that lie is more than twice as high compared to the case of groups of 2. We also find that the group size has no impact on the intensity of lying, i.e. by how much the outcome of the die roll is misreported (Result 2). Rather, for all group sizes, conditional on lying, groups virtually always choose the report that maximizes their monetary payoffs, i.e. there is no partial lying.

A second set of results relates to the impact of the group gender composition on the prevalence of lying. For all group sizes, we find that all-male groups lie more often than all-female groups (Result 3), thereby extending the evidence for dyads provided by Muehlheusser, Roider, and Wallmeier (2015).

Compared to all-male groups, lying is also substantially lower in *almost-all-male groups* (Result 4), i.e. groups with exactly one female member. Finally, we find that when excluding all-male groups from the sample, the frequency of lying does not differ systematically across the different group gender compositions (Result 5). Together, these results suggest that all-male groups really stand out regarding their proclivity to lie. Moreover, the prevalence of lying in

[8] This group task is *Task 1*. All other parts of the experiment were played at the individual level and are discussed in more detail below.

[9] For a survey of the extensive experimental literature on voting in committees and groups, see e.g. Plott and Smith (2008, Part 6.2).

groups does not strictly decrease with the number of females in the group; it rather seems to be the first female in the group that causes an honesty shift. In particular, groups do not become more honest as the share of females further increases.

We then study the effect of group size and the group gender composition on the process of deliberation. With respect to the length of group discussions, we find that larger groups need more time to reach a decision. Interestingly, there is no difference in decision times between all-male and almost-all-male groups (Result 6). This finding suggests that the lower inclination towards lying in almost-all-male groups is not accompanied by longer group discussions.

A further set of results relates to the effect of individual honesty preferences. These were elicited through an individual task (due to Hugh-Jones, 2016) that was played after the group task, and that allows us to classify individuals as either *cheaters* or *non-cheaters*. We first document that the share of cheaters is higher among males compared to females (Result 7). We then analyze whether gender differences in honesty at the individual level can explain the observed gender effects at the group level, in particular the higher frequency of lying in all-male groups. Focussing on dyads, we show that in dyads that do not contain any cheaters, the frequency of lying is substantially higher in all-male groups than in all-female groups (Result 8). This finding suggests that group behavior cannot be fully explained by the individual group members' honesty preferences. Finally, and more generally, we study how group behavior is affected by the number of (individual) cheaters in the group. We find that not only the first cheater in the group (i.e. a *bad apple*) matters, but rather the frequency of lying increases with the number of cheaters in the group (Result 9).

We also verify the robustness of our main results using a regression analysis where we include controls such as the die roll outcome and (group averages of) various individual characteristics of group members elicited in a post-experimental survey. Last, but not least, we study those groups that fail to reach an agreement. We find that disagreement is more prevalent in almost-all-male and almost-all-female groups compared to all-male and all-female groups.

The remainder of the paper is structured as follows: Section 3.2 discusses the related literature. Section 3.3 explains the experimental design and implementation. Section 3.4 presents our results on the effects of group size and group gender composition on group lying behavior. In this section, we also present findings on decision times, the effects of individual honesty on group behavior, and on the robustness of the results in a regression analysis. Finally, we also look at groups that did not reach an agreement. Section 3.5 discusses our findings and concludes. The Appendix provides the experimental instructions, screenshots, and additional empirical results.

3.2 Related literature

Our paper is related to three strands of the literature. First, we contribute to the **literature on decision-making by groups**. While economic research has traditionally focussed on individual decision-making, by now, there is a sizeable (and mostly experimental) literature studying group decisions in settings with intellective tasks (e.g., problem solving), judgmental tasks (e.g., altruism, risk taking, honesty), or strategic tasks (e.g., prisoners' dilemma). This literature finds substantial differences compared to decision-making by individuals (for surveys, see Charness and Sutter, 2012; Kugler, Kausel, and Kocher, 2012). Virtually all group-decision studies consider either groups of two or three members (i.e. dyads or triads), and hence they do not focus on the effect of group size as the present paper does. Notable exceptions include Sutter (2005), who compares the behavior of individuals, dyads, and groups of four in a strategic task (beauty contest), Charness, Karni, and Levin (2010), who compare individuals, dyads, and triads in an intellective task (conjunction fallacy), and Engl (2022) who studies the influence of ideology on decision-making in dyads and groups of four.[10]

Second, we contribute to the experimental **literature on unethical decisions**, in particular lying. For the case of individual lying behavior, Abeler, Nosenzo, and Raymond (2019) provide a comprehensive survey and meta study. They find that, in general, many individuals do not too readily tell a lie even if doing so would yield them a benefit. They also provide evidence for two key motives underlying such behavior, namely (i) being honest and (ii) being seen as honest by others. A body of research has also investigated gender differences in lying behavior at the individual level (see e.g., Dreber and Johannesson, 2008; Erat and Gneezy, 2012; Childs, 2012; Houser, List, Piovesan, Samek, and Winter, 2016; Conrads, Irlenbusch, Rilke, and Walkowitz, 2013; Conrads, Irlenbusch, Rilke, Schielke, and Walkowitz, 2014; Muehlheusser, Roider, and Wallmeier, 2015). Overall, males tend to be less honest than females, but in some studies the observed effects are small or not statistically significant.

With respect to lying behavior in group settings, a recent body of experimental research (again mostly focussing on dyads or triads) provides evidence that groups are more prone to take unethical decisions than individuals. In doing so, many studies have adapted the die-roll paradigm of Fischbacher and Föllmi-Heusi (2013) to group settings, where group members have to make a joint decision (as does the present paper).[11] For example, Kocher, Schudy, and Spantig (2018) study lying by individuals and triads, where they focus on the role of within-

[10] Using non-incentivized experiments, the effects of group size are also studied in psychology (see e.g., Laughlin et al., 2006). For theoretical studies on the effects of group size in settings with various formal decision rules, see e.g., Mukhopadhaya (2003) and Feddersen and Pesendorfer (1998).

[11] A number of recent empirical studies have also provided evidence that behavior in die roll experiments correlates with behavior in the field (see e.g., Cohn, Maréchal, and Noll, 2015; Gächter and Schulz, 2016; Potters and Stoop, 2016; Hanna and Wang, 2017; Dai, Galeotti, and Villeval, 2018; Cohn and Maréchal, 2018).

group communication (text messages) and whether or not there exists a payoff communality among group members. They document a "dishonesty shift", i.e. a higher lying frequency in triads compared to individual decision-making. Dannenberg and Khachatryan (2020) find that the dishonesty shift between triads and individuals is reinforced in the presence of competition. Castillo, Choo, and Grimm (2022) employ the design of Kocher, Schudy, and Spantig (2018) with the difference that not the experimenter gets harmed by subjects' lying behavior, but an alternative third party (i.e. a charity). They find no behavioral difference between triads and individuals. Muehlheusser, Roider, and Wallmeier (2015) compare lying in dyads and individuals, and find no difference in behavior when not differentiating with respect to gender. When taking the group gender composition into account, they find that all-male dyads lie much more than all-female dyads, while the behavior of mixed dyads is in-between, but closer to all-male dyads.

A further set of studies employs the die-roll paradigm in group settings where, however, group members make individual decisions. For example, Conrads et al. (2013) compare the effects of individual compensation (i.e. subjects are on their own and receive payoffs according to their decision) with team compensation (i.e. subjects still decide individually, but are matched in dyads and share the joint payoff with their partner). They find that lying is more prevalent under team incentives. Likewise, Irlenbusch et al. (2020) study the role of feelings of similarity and a code of conduct in a setting where members of a dyad observe and report the (same) die roll outcome sequentially.

Apart from the die role setting of Fischbacher and Föllmi-Heusi (2013), lying behavior has also been studied employing other paradigms that introduce more complex strategic considerations, e.g. in form of cheap talk games due to Gneezy (2005).[12] In such a framework, behavioral differences between individuals and groups seem to additionally depend on whether or not a lie is expected to be believed by others (see e.g. Sutter, 2009; Cohen et al., 2009).

Finally, apart from lying, the literature has also considered other forms of unethical behavior by individuals and groups. For example, Falk, Neuber, and Szech (2020) re-consider the "mouse paradigm" of Falk and Szech (2013) and find that more individuals choose the unethical option in a group setting (where groups consist of eight members). This finding is also robust in an alternative design with donations.[13]

The third area to which our paper contributes is the **literature on how group decisions are shaped by the composition of the group**, where the group gender composition is one topical dimension of interest. In this respect, various studies look at domains such as corporate

[12] See also Abeler, Nosenzo, and Raymond (2019) for a more detailed discussion.

[13] Falk and Szech (2013), Bartling, Weber, and Yao (2015), and Bartling, Fehr, and Özdemir (2023) investigate the erosion of moral values in market settings as compared to individual decision-making.

boards (see e.g. Matsa and Miller, 2013; Gormley, Gupta, Matsa, Mortal, and Yang, 2023), judge panels (see e.g. Farhang and Wawro, 2004; Peresie, 2005; Boyd, Epstein, and Martin, 2010), hiring committees (see e.g. Bagues and Esteve-Volart, 2010; Bagues, Sylos-Labini, and Zinovyeva, 2017; Radbruch and Schiprowski, 2023), willingness to lead (see e.g. Born, Ranehill, and Sandberg, 2022), problem-solving (see e.g. Berge, Juniwaty, and Sekei, 2016), dictator games (see e.g. Dufwenberg and Muren, 2006), and confidence judgments (see e.g. Keck and Tang, 2018). However, as discussed in the Introduction, for the domain of unethical behavior, the literature on potential effects of the group gender composition is scant. To the best of our knowledge, the only study in this respect is Muehlheusser, Roider, and Wallmeier (2015), who find that small (and insignificant) differences in individual lying behavior between males and females are amplified in all-male and all-female dyads.

3.3 Experiment

In this section, we describe the experimental design (Section 3.3.1) and its implementation (Section 3.3.2). The experimental instructions are provided in Appendix 3.A.

3.3.1 Design

The experiment consists of five tasks. Task 1 is a group task, in which subjects take decisions jointly at the group level. Task 1 is the main focus of the present paper, and we consider various treatment variations, which are discussed below. By contrast, Task 2 to 5 are completed individually and are, thereby, not affected by any treatment variation. Tasks 2 to 5 serve to elicit individual preferences and characteristics; they are discussed in more detail in Sections 3.4.4 and 3.4.5 below.

In Task 1, each group needs to reach a decision whether or not to act honestly or dishonestly. In particular, we consider the die-roll paradigm of Fischbacher and Föllmi-Heusi (2013), but in a group setting. All group members observe the (same) outcome of a random die roll. They are then asked to memorize and possibly discuss the die roll outcome and to report – jointly as a group – the respective number to the experimenter.

Importantly, a group's payoff depends only on its report, but not on the actual outcome of the die roll itself. Moreover, to obtain a positive payoff, a group must reach a unanimous agreement concerning the number they jointly report.[14] If an agreement is reached, the payoff $\pi(r)$ (in £) for *each* group member is related to the reported (not necessarily true) outcome of the die roll $r \in \{1, ..., 6\}$ as follows: $\pi = r/2$ for all $r \leq 5$, and $\pi = 0$ for $r = 6$. Hence, unless

[14] In our view, requiring unanimity highlights the idea of a *joint* group decision, and this assumption is also made in Kocher, Schudy, and Spantig (2018) and Muehlheusser, Roider, and Wallmeier (2015) for the context of unethical behavior.

the true outcome of the die role is 5, a group can increase its payoff by reaching an agreement to lie, i.e. reporting a number different from the true outcome of the die roll.

If a group fails to reach an agreement within a time limit of 10 minutes (of which subjects were aware), the payoff for each group member for Task 1 is zero. To reach an agreement on the group report, group members were able to deliberate face-to-face with their fellow group members. As the experiment was played online, we implemented a video chat, which allowed group members to interact in a by-now familiar online environment (for more details, see Section 3.3.2, where we also discuss how we ensured smooth online face-to-face group discussions by implementing various functionality checks and other safeguards). The video chat gave groups the opportunity of free-form discussions (similar to many workplace environments), thereby allowing for potential gender effects to emerge naturally.[15]

As for treatment variations of Task 1, we consider group sizes $n \in \{2, 3, 4, 5\}$. Moreover, for each group size n we systematically vary the group gender composition. That is, we consider all possible combinations of female and male subjects, leading to $n + 1$ different group gender compositions. That is, each treatment is identified by the group size and the number of female subjects in the group. This leads to a total of 18 treatments (i.e. $3 + 4 + 5 + 6$).

3.3.2 Implementation

Conducting an experiment where the unit of observation is a group of subjects poses various challenges: First, it requires a relatively high number of subjects, as only a group of subjects constitutes one independent observation. Second, in our context, groups of different sizes and group gender compositions need to be able to communicate in private, which is logistically difficult to implement in a lab environment.[16] Third, it is important to avoid communication between groups. We address these issues by conducting the experiment online. It is programmed in *oTree* (Chen, Schonger, and Wickens, 2016) and was conducted in 2021 and 2022.

In order to facilitate face-to-face communication, we developed a novel platform by embedding a video chat tool in an *oTree* environment, which allows for face-to-face communication in larger groups.[17] Thereby, it is also possible to track each group member's communication patterns (e.g. the frequency, volume, and duration of contributions).[18]

[15] In their study of gender effects in group leadership decisions, Born, Ranehill, and Sandberg (2022) also emphasize the desirability of groups being able to interact and deliberate face-to-face.

[16] Moreover, because of the COVID-19 pandemic, running the experiment in-person in the lab was not feasible at that time.

[17] This tool was kindly provided by *Vonage* (see https://www.vonage.com). Previously, there existed a beta version of a video chat option for *oTree* which, however, is limited to dyads (`https://github.com/oTree-org/video-chat`). Our platform was developed independently.

[18] We refrained from recording the content of group communications. For reasons of data privacy, we would have had to alert subjects of such recordings ex ante, and we feared that this might have potentially affected not only communication, but also behavior.

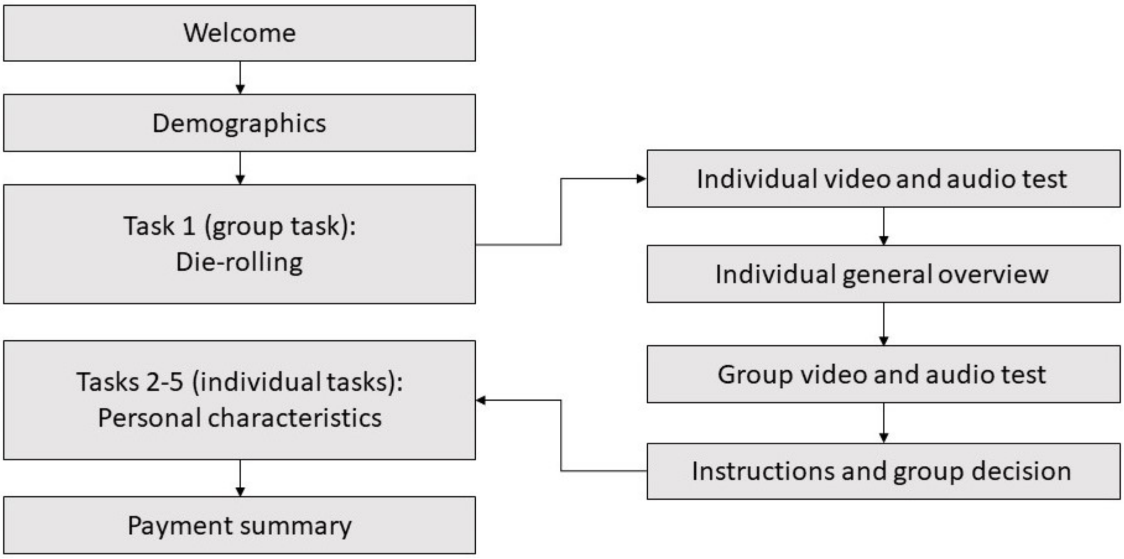

Welcome
Demographics
Task 1 (group task):
Die-rolling
Tasks 2-5 (individual tasks):
Personal characteristics
Payment summary
Individual video and audio test
Individual general overview
Group video and audio test
Instructions and group decision

Figure 3.1: **Sequence of events in the experiment.**

Task 1 is the main task. To ensure smooth communication within groups, participants had to pass a couple of technical tests, first individually and then, right before Task 1 started, as a group.

As illustrated in Figure 3.1, the experiment started with a welcome screen on which we gave subjects a general preview, in particular, about the number of tasks, and that one of them (Task 1) would be performed in a group. We informed our subjects that the details on each task (including information on the payment scheme) would be provided when the respective task is reached. Moreover, since Task 1 involves a video chat, we asked subjects to confirm that they would be willing to participate in it, and that their camera and microphone were functional. We then asked subjects for some demographic variables (age, gender, education, ethnical background, and country of residence). We elicited the demographics at the beginning of the experiment as we needed information on subjects' gender in order to implement various group gender compositions.

Subjects then proceeded to the group task, where they went through a sequence of four screens (see Figure 3.1, and for screenshots see Figures 3.6 to 3.10 in Appendix 3.A). The first three screens were meant to facilitate frictionless online face-to-face group discussions. In the first two of these screens, subjects were still on their own, i.e. they did not interact with any fellow group members yet. On the screen "Individual video and audio test", each subject individually had to perform a functionality test of their camera and microphone (see Figure 3.6 in Appendix 3.A).[19] Subjects who successfully completed the functionality test then proceeded to the screen "Individual general overview", where we gave them some basic information regarding

[19] On this screen, subjects were asked to click on a link, which opened an additional browser window and directed them to the (external) website of a provider of free video and audio tests (see `https://tokbox.com/developer/tools/precall/results`). This website automatically checks the functionality and (transmission) quality of the respective user's camera and rates them on scores ranging from 0 to 4.5 (for an example, see Figure 3.7 in Appendix 3.A). This took between 10 and 20 seconds. We asked subjects to report these scores, and they were allowed to continue if the reported score was at least 2.5 in each test.

the structure of the upcoming group interaction (see Figure 3.8 in Appendix 3.A). We did this to familiarize subjects with the setting. Afterwards, groups were formed, and group members met for the first time on the screen "Group video and audio test". On this screen, each group member had to confirm that they can see and hear all other group members before being able to proceed (see Figure 3.9 in Appendix 3.A).[20]

Finally, on the screen "Instructions and group decision", subjects first had three minutes to read the instructions for Task 1 (see Figure 3.10 in Appendix 3.A). After three minutes, the die roll was shown in the form of a short video. All group members saw the same video, and each of the six potential outcomes was equally likely to be shown. Group members were informed that the die role would be displayed for 10 seconds, and their task would be to memorize it. After a potential discussion in the video chat, groups made their report in the following way. Each group member saw a live-updating table displaying the numbers reported by each group member (including their own report). Each group member was able to change their entry as often as they wanted before an agreement was reached. A group agreement was reached (and logged in) once all members reported the same number (i.e. reports could no longer be changed after that). The group decision and the resulting payoff for each group member were then implemented.

Upon completion of Task 1, groups were dissolved, and subjects proceeded to Tasks 2 to 5, which they performed individually.[21] After Task 5, they received a summary of payoffs obtained in each of the five tasks.

For recruitment of subjects and implementation of payments we used the platform *Prolific*. As the experiment (including the group discussion in Task 1) was conducted in English, we only recruited subjects residing in the UK or the US. In total, 1677 subjects passed all technical checks (as discussed above) and were matched into 447 groups (average age: 39.6, and roughly 70% and 30% residing in the UK and the US, respectively). All subjects identified as either male or female, and the share of female subjects was 50.6%. Almost 95% of subjects have at least a high school degree, and 77% have an undergraduate degree or higher. On average, it took subjects approximately 30 minutes to complete the whole experiment (median: 29.0 minutes), and the average payoff was £5.55 (sd = 1.33). The number of group observations in each of the 18 treatments is shown in Table 3.1. In the preregistration, we specified the experimental design and that we aimed to study the effect of group size and group gender composition

[20] To ensure that group members were able to smoothly communicate with each other on the upcoming screen "Instructions and group decision", we dropped all individuals and groups that experienced or reported technical problems or were inattentive (e.g. because they reported non-admissible scores for audio and video tests) on the screens discussed so far.

[21] If a group failed to reach an agreement within a 10 minute time limit (of which subjects were aware), this was recorded as a disagreement, and subjects also proceeded to Task 2.

	Number of females in the group						
Group size	0	1	2	3	4	5	Total
2	25	29	23	-	-	-	77
3	25	25	24	25	-	-	99
4	25	27	23	30	24	-	129
5	20	25	24	22	27	24	142
	95	106	94	77	51	24	447

Table 3.1: **Number of group observations per treatment.**

through our 18 treatments.[22] Given the explorative nature of the experiment and the lack of clear theoretical predictions, we did not preregister any directed hypotheses.

3.4 Results

In this section, we present our main results. Out of a total of 447 groups, 385 groups had an incentive to lie to their advantage (i.e., they observed a die roll $r \neq 5$), and out of these groups 363 groups reached an agreement on which number to report. The analysis in the present Section focusses on these 363 groups.[23] In particular, we report on how lying behavior is affected by group size (Section 3.4.1) and by the group gender composition as measured by the number of females in the group (Section 3.4.2).[24] We also consider the impact of group size and group gender composition on decision times (i.e., the time groups needed to reach an agreement) as well as talking times within groups (Section 3.4.3).[25] In Section 3.4.4 we investigate how group members' individual honesty affects group lying behavior. In Section 3.4.5, using regression analysis, we show that our main results are robust when accounting for additional controls such as personal characteristics of group members. Finally, in Section 3.4.6 we analyze the 22 groups that had an incentive to lie, but failed to reach an agreement (resulting in individual payoffs of zero in Task 1 for all of these group members).

3.4.1 The effect of group size on lying behavior

We first explore the *frequency* of lying, i.e. the impact of group size on groups' (binary) decisions whether or not to lie about the outcome of the die roll. In addition, we then also consider the *intensity* of lying, i.e. by how much groups eventually misrepresent the outcome.

[22] We preregistered 20 group observations per treatment. When implementing the experiment, we did not know how many individual subjects and how many groups would pass all of the technical checks (as outlined above), and hence would turn into usable group observations. To take this into account, we aimed for more than 20 groups per treatment, which in the end led to the number of realized observations as outlined in Table 3.1.

[23] For an overview of how these group observations are distributed across treatments, see Table 3.4 in Appendix 3.B.

[24] Note that, out of the 363 groups under consideration, all of the groups that decided to lie lied to their own monetary advantage.

[25] While we did not record video chats, we have access to audio log data (i.e., microphone activity), which allows us to proxy the time subjects were actually talking.

Result 1. The frequency of lying increases with group size.

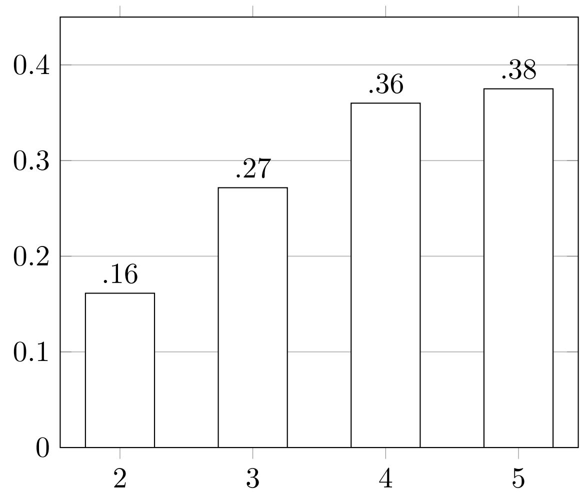


Figure 3.2: **Frequency of lying by group size.**

This figure is based on 62, 81, 100 and 120 observations for group size $n = 2, 3, 4, 5$, respectively. The frequency of lying is 16% for dyads, 27% for triads, 36% for groups of 4, and 38% for groups of 5.

The result is illustrated in Figure 3.2 and supported by a highly statistically significant Jonckheere-Terpstra test for the presence of a trend ($p = 0.002$).[26] For example, groups of four and five are more than twice as likely to lie than dyads ($p = 0.006$ and $p = 0.003$, Chi2-test). Moreover, the difference between dyads and triads as well as between triads and groups of four and five are also sizeable, but not significant (dyads versus triads: $p = 0.117$, triads versus groups of four: $p = 0.205$, triads versus groups of five: $p = 0.137$, and groups of four versus groups of five: $p = 0.818$, all Chi2-tests).

Recall that in our experiment, the group decision requires unanimity. Our finding of a positive relationship between group size and the frequency of lying is therefore in line with a recent literature (both theoretical and experimental) arguing that the individual incentive to support an unethical (or antisocial) group decision increases with the number of group members required to support it. This finding is often attributed to *guilt sharing* (or *diffusion of responsibility*), i.e. a reduction of individual moral cost based on the argument that any other group member could also prevent such a decision (see e.g., Dana, Weber, and Kuang, 2007; Bartling and Fischbacher, 2012; Irlenbusch and Saxler, 2019; Rothenhäusler, Schweizer, and Szech, 2018; Falk, Neuber, and Szech, 2020; Behnk, Hao, and Reuben, 2022; Feess, Kerzenmacher, and Muehlheusser, 2023).

Next, we consider the intensity of lying (i.e. the difference between the observed and the declared outcome of the die roll), where we have the following result:

[26] A positive effect of group size on the frequency of lying is also confirmed in a regression analysis (see Table 3.2 below).

Result 2. The group size has no effect on the intensity of lying, because partial lying does virtually not occur.

In our experiment, 97% of the groups that lie choose $r = 5$, i.e. they opt for the maximum monetary benefit. This suggests that groups perceive this as a yes/no decision, and they do not make use of the possibility to vary the intensity of lying, for example, to reduce eventual moral costs. Moreover, the percentage of lying groups reporting $r = 5$ is very high across all group sizes with 90%, 95.5%, 100%, and 98% for $n = 2, 3, 4, 5$, respectively (Jonckheere-Terpstra test for the presence of a trend, $p = 0.205$).

The issue of *partial lying* (i.e. reporting a number that is strictly larger than the die roll outcome, but strictly smaller than five) has also been studied in experiments with individual decision-making. Fischbacher and Föllmi-Heusi (2013) have attributed partial lying to image concerns (e.g. vis á vis the experimenter or future selves), inducing subjects to disguise their lying. In paper-and-pencil settings such as in Fischbacher and Föllmi-Heusi (2013) and Muehlheusser, Roider, and Wallmeier (2015), the possibility to disguise lying arises from the fact that die roll outcomes are subjects' private information, such that lies cannot be detected at the individual, but only statistically at the aggregate level. However, the possibility to disguise lying and, consequently, also the extent of partial lying, should decrease when subjects presume (or even know) that a lie can be detected. For example, in computerized die roll settings, the true outcome can be observed by the experimenter in which case the (perceived) possibility to disguise lying becomes smaller or even vanishes. Indeed, support for this hypothesis is provided by Gneezy, Kajackaite, and Sobel (2018), Abeler, Nosenzo, and Raymond (2019), and Crede and von Bieberstein (2020).[27] This reasoning can also explain the absence of partial lying in our online group setting, where the die roll is observed by the experimenter as well as by all group members.

3.4.2 The effect of the group gender composition on lying behavior

We now turn to the question how the lying behavior of groups is affected by the group gender composition. First evidence on this issue has been provided by Muehlheusser, Roider, and Wallmeier (2015), who find that all-male dyads lie significantly more than all-female dyads. The present study replicates this finding as can be seen in Figure 3.3(a), where the left-most and right-most bar correspond to all-male groups and all-female groups, respectively. In addition, the other panels of Figure 3.3 suggest that this finding is not specific to dyads, but seems to hold independent of group size:

Result 3. The frequency of lying is higher in all-male groups compared to all-female groups.

[27] See also Dufwenberg and Dufwenberg (2018) and Khalmetski and Sliwka (2019) for further theoretical contributions explaining the emergence of partial lying.

In each of the four panels of Figure 3.3, we compare the left-most bar (all-male groups) with the right-most bar (all-female groups). For each group size, lying is substantially more prevalent in all-male groups than in all-female groups, with percentage point differences of 30, 15, 35, and 25 for $n = 2, 3, 4, 5$, respectively. When pooling observations over all group sizes, a Chi2-test reveals that the difference in lying between all-male and all-female groups is highly statistically significant ($p = 0.002$). Performing such tests separately for each group size, the difference is statistically significant for dyads ($p = 0.02$) and groups of four ($p = 0.02$), and close to significance for groups of five ($p = 0.11$), but not significant for triads ($p = 0.72$). In addition, for all group sizes, lying is most prevalent in all-male groups compared to all other group gender compositions, and for $n = 2, 4, 5$ this is also true by a large margin.

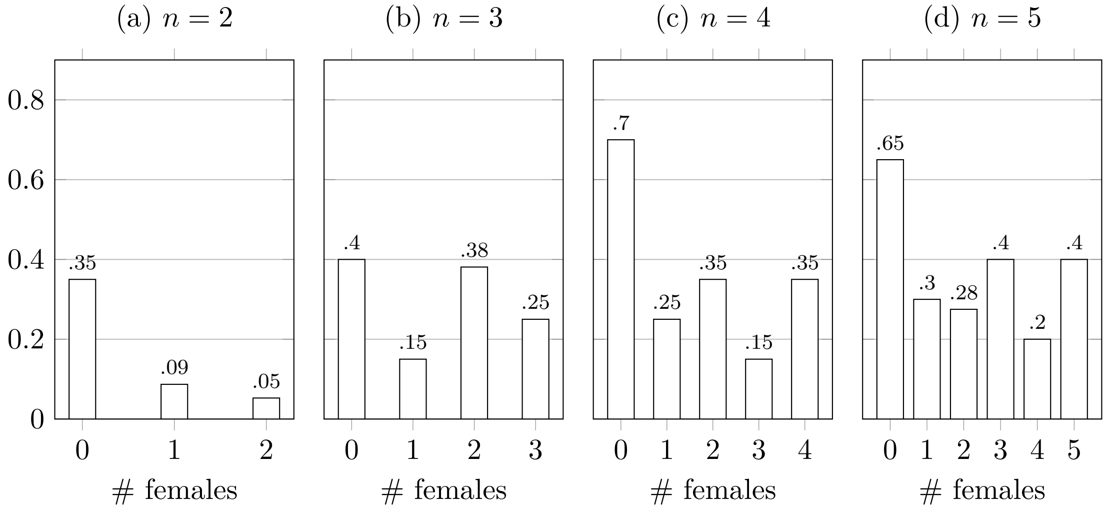


Figure 3.3: **Frequency of lying by group size and number of females in the group.**

For each group size $n \in \{2, 3, 4, 5\}$ the number of females per group is listed on the x-axis. The frequency of lying for each group gender composition is shown above the corresponding bar. For example in graph (a), depicting all dyads, all-male groups lie with a frequency of 35%, mixed groups lie with a frequency of 9% and all-female groups lie with a frequency of 5%.

Muehlheusser, Roider, and Wallmeier (2015) also find that the frequency of lying in mixed dyads (i.e. one male and one female group member) is in-between that of all-male and all-female dyads. As can be seen in Figure 3.3(a), this is also the case in our experiment. Moreover, there is additional tentative evidence that female individuals lie less than male individuals (see e.g. Dreber and Johannesson, 2008; Erat and Gneezy, 2012; Houser et al., 2016). One might thus hypothesize that, also in larger groups, the number of females in the group has a (weakly) monotone effect on group dishonesty. However, panels (b)-(d) of Figure 3.3 indicate that this is not the case. Nevertheless, a first striking observation emerges from the comparison of all-male groups with *almost-all-male* groups (i.e. groups with one female and otherwise all-male members):

Result 4. The frequency of lying is higher in all-male groups compared to almost-all-male groups.

The result is again illustrated in Figure 3.3 by comparing the two left-most bars in each panel. For each group size, lying is substantially more prevalent in all-male groups than in almost-all-male groups, with percentage point differences of 26, 25, 45, and 35 for $n = 2, 3, 4, 5$, respectively. When pooling over group sizes, the fractions of dishonest all-male and dishonest almost-all-male groups are 0.50 and 0.19, respectively, and this difference is highly statistically significant ($p = 0.000$, Chi2-test). Furthermore, performing such tests separately for each group size, the drop in dishonesty from all-male groups compared to almost-all-male groups is statistically significant for dyads ($p = 0.034$), groups of four ($p = 0.004$), and groups of five ($p = 0.027$), but not for triads ($p = 0.256$).

Interestingly, we do not find similarly consistent effects for the comparison of all-female and almost-all-female groups (i.e. groups with one male and otherwise all-female group members). While Figure 3.3 shows sizeable differences between all-female and almost-all-female groups of 20 percentage points for both group sizes $n = 4$ and $n = 5$, these difference are not statistically significant ($p = 0.144$ and $p = 0.168$, respectively). Furthermore, the fraction of dishonest all-female dyads is only 0.05, such that the scope for further reduction is limited and the difference to almost-all-female groups is insignificant ($p = 0.667$). For triads, there is a sizeable, but insignificant, increase ($p = 0.368$).[28]

Taken together, our results suggest that the frequency of lying of all-male groups stands out compared to all other group gender compositions. At the same time, replacing just one male group member by a female group member already leads to a substantial reduction in lying. In fact, our next finding suggests that it is really the *first* female in a group that matters in terms of curtailing lying:

Result 5. When excluding all-male groups, the frequency of lying is not affected by the group gender composition.

For an illustration consider the groups of five in Figure 3.3(d). Excluding all-male groups, the average fraction of dishonesty in the five remaining group gender compositions is 0.32, and for each group gender composition, the fraction of dishonest groups fluctuates around this average without showing a clear trend. A similar observation emerges for the other group sizes. This is confirmed by Jonckheere-Terpstra tests (performed separately for each group size and excluding all-male groups) which all reject the presence of a (positive or negative) relationship between group dishonesty and the number of females in the group (dyads: $p = 0.671$, triads: $p = 0.46$, groups of four: $p = 0.824$, groups of five: $p = 0.763$). In fact, for each group size also the pairwise comparisons of all group gender compositions show that group dishonesty does not differ across group gender compositions (when excluding all-male groups). In particular,

[28] When pooling over group sizes, the fraction of dishonest all-female and almost-all-female groups is 0.27 and 0.20, respectively, where the difference is not statistically significant ($p = 0.338$).

19 out of these 20 pairwise comparisons are not statistically significant (where the exception is the comparison of triads with one and two females, $p = 0.095$).

Our findings suggest that changes in group behavior are most pronounced when moving away from all-male groups. A similar finding arises in the empirical study by Matsa and Miller (2013) who exploit a legal regime change in Norway (female quotas in company boards) to study how the gender composition of the board affects crucial variables such as labor policies and profits. They find that the effects are strongest for firms led by all-male boards before the legal change.

3.4.3 Decision times and talking times

In this section we analyze how much time groups take to reach an agreement (*decision time*) and patterns of communication.[29] A first – and straightforward – hypothesis in this respect is that it takes larger groups more time to reach a decision, as the process of deliberation becomes more complex when more people are involved. Another question of interest is the effect of the group gender composition on decision times. In particular, the drop in the frequency of lying in almost-all-male groups compared to all-male groups (see Result 4) might be accompanied by longer group discussions in almost-all-male groups, during which the sole female member tries to convince the male members to refrain from lying.

In the analysis, the decision time is defined as the elapsed time (in seconds) between the time stamp when the group members are shown the instructions and the time stamp when the group decision is locked in (i.e. the time they spend on the screen “Instructions and group decision” of Figure 3.1).[30]

Result 6. (i) Decision times increase with group size. (ii) There is no difference in decision times between all-male and almost-all-male groups.

The result is illustrated in Figure 3.4. As shown in panel (a), and not surprisingly, larger groups need more time to reach a decision. While the effect is statistically significant (Jonckheere-Terpstra test, $p = 0.000$), its size is moderate with groups of four and five roughly taking 30 to 40 seconds more to reach a decision than dyads. Moreover, we find no evidence that almost-all-male groups exhibit longer group discussions. As illustrated in Figure 3.4(b), three out of the four bars are virtually identical. Moreover, also the difference between almost-all-male groups that don't lie (second bar) and all-male groups that do lie (third bar) is not statistically significant ($p = 0.47$).

The claim that discussions seemingly do not take longer in almost-all-male groups is also corroborated by the actual *talking time* of group members. While we do not record the com-

[29] The analysis of decision times and talking times was not addressed in the preregistration.

[30] Group members could already communicate with each other during the time window of 180 seconds designated to reading the instructions (i.e. before the die roll was shown). Therefore, these 180 seconds are counted as decision time. Our results remain qualitatively robust when excluding this time window.

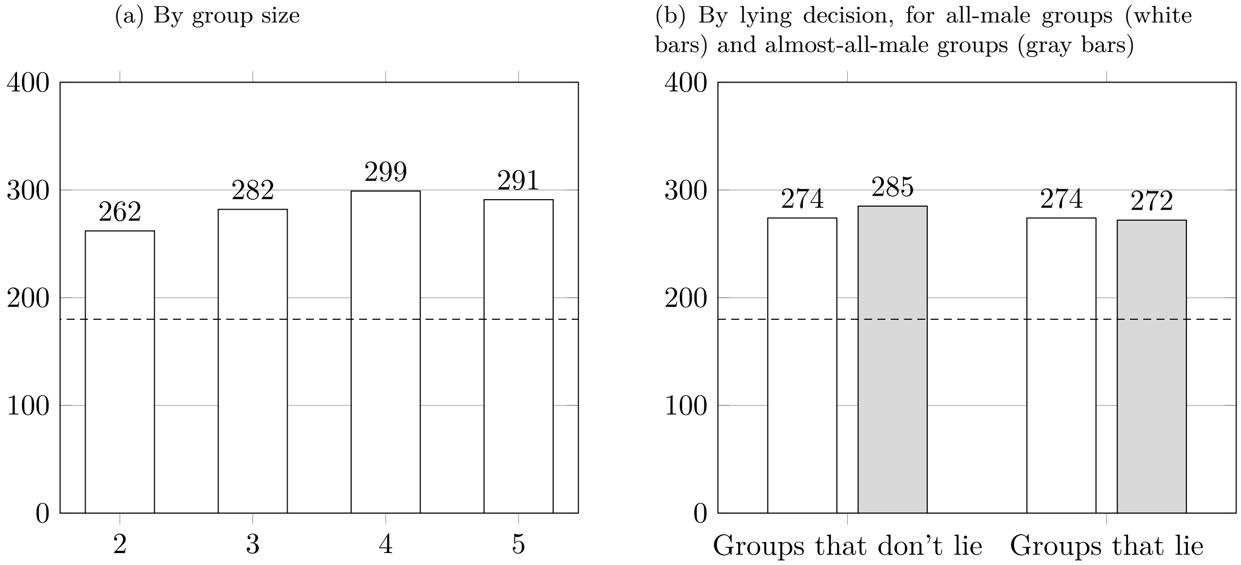


Figure 3.4: **Groups' decision times (in seconds).**

This figure shows the decision times for groups in seconds above each bar. In the left graph (a), groups are sorted by group size $n \in \{2, 3, 4, 5\}$, with groups of 4 taking the longest average time to decide (299 seconds). In the right graph (b), only all-male groups (white bars) and almost all-male groups (grey bars) are depicted. They are separated in groups that lie and groups that don't lie.

munication itself (i.e. we do not know what group members say), we can use group-member specific log data on the audio level of the microphone to measure *when and for how long* any given group member talks. This allows to construct (gender-specific) measures of the individual and also the overall talking time in the group.

For almost-all-male groups, on average, the (sole) female's share of overall talking time in the group is *lower* than $1/n$ (i.e. the share that would result from identical talking times of all n group members). In particular, we examine the behavior of almost-all-male groups in eight cases of almost-all-male groups divided by their group size ($n = 2, 3, 4, 5$) and whether or not they lied. In seven out of these eight cases, the female's share of talking time in the group is lower than $1/n$, while in the eighth case it is almost identical to $1/n$. This finding is related to Karpowitz, O'Connell, Preece, and Stoddard (2024), who study the influence of females on decision-making in mixed groups. They also find that the (single) females in almost-all-male groups participate less in group discussions.[31]

[31] In their experiment on team performance, Hardt, Mayer, and Rincke (2023) study groups of four. They find that, in gender-balanced teams, males talk significantly more than males assigned to all-male teams. Interestingly, females exhibit the opposite pattern, that is, they talk less in mixed teams compared to all-female teams. Gender differences in the participation in group discussions are also studied in the context of education, where male students are often found to be considerably more active than female students, see e.g. Lee and McCabe (2021) and the studies cited therein.

3.4.4 The effect of individual honesty on group behavior

In this section, we study the role of group members' individual honesty preferences for the group decision. On the one hand, it seems natural to presume that individual preferences (or some derived aggregate measures thereof, such as a the number of cheaters in the group) will be a key driver for group decisions. On the other hand, our previous analysis suggests that also other factors such as the size and the gender composition of the group might play a role.

In particular, we are interested in the following three questions: First, as a preliminary step, we investigate gender differences in honesty at the individual level in a separate task, for example whether males in our experiment are more prone to dishonesty than females. Second, we analyze whether gender differences at the individual level can explain the observed gender effects at the group level, in particular the higher frequency of lying in all-male groups. Third, and more generally, we study how the number of individually dishonest group members affects group behavior. This includes the question of whether there is contagion: Can one *bad apple* "spoil" an entire group?

To address these questions, we elicited a measure of honesty at the individual level after subjects had completed the group task. We did not want to employ the die-roll paradigm again, because subjects had already encountered it before. Instead we employed the task suggested by Hugh-Jones (2016). This individual task consists of six questions in the context of music. Three of the questions are arguably very challenging, but the correct answers could easily be obtained from the internet. For example, one question asks in which year the French composer Claude Debussy was born.[32] Subjects were informed that they would receive a payment of £0.5 when answering *all* six questions correctly, and 0 otherwise. Subjects were also told that they are not allowed to use the internet. Hence, in all likelihood, subjects were only able to earn the bonus by cheating (i.e. using the web to find the correct answers). Consequently, a subject is regarded as dishonest (and coded as a *cheater*) if all six questions were answered correctly. Otherwise, the subject is regarded as honest.[33] We obtain the following result on gender differences in honesty at the individual level:

Result 7. In the individual honesty task, the share of cheaters among males is larger than among females.

[32] Technically, the music quiz was Task 3 of the experiment. The instructions provided in Appendix 3.A contain all six questions of this task.

[33] Our terminology hence reflects a *consequentialist* approach in moral philosophy (see e.g., Sinnott-Armstrong, 1988), according to which unethical behavior is deemed immoral only if it actually involves a negative consequence for others (i.e. the experimenter in our setting). Alternatively, under a *non-consequentialist* (or *deontological*) approach, unethical behavior would be considered immoral per se (see e.g., Alexander and Moore, 2016). For an experimental study of the relevance of these two concepts for different domains of unethical behavior, see Feess, Kerzenmacher, and Timofeyev (2022).

In our experiment, 31 percent of male subjects are cheaters as opposed to only 25 percent of female subjects ($p = 0.014$, Chi2-test). This result is similar to earlier findings in the literature (see the discussion in Section 3.2).

This raises the question whether the observed gender effects at the group level (Results 3-5, and in particular the stark difference between all-male and all-female groups) are driven by gender differences at the individual level.

To analyze this, we consider *groups that do not contain any cheaters*. Thereby, we focus on dyads for which we have 16 (11) observations of all-male (all-female) groups without any cheaters.[34] If group behavior was mainly driven by individual honesty preferences, we should not observe any difference in lying between such all-male and all-female groups. However, we obtain the following result:

Result 8. In dyads that do not contain any cheaters, the frequency of lying is larger in all-male groups than in all-female groups.

We find that 44 percent of all-male dyads that do not contain any cheaters lie compared to only 9 percent of all-female groups without cheaters ($p = 0.053$, Chi2-test). Hence, this substantial difference suggests that group interaction plays a major role in determining lying behavior at the group level beyond any gender differences at the individual level. To further substantiate this point, the observed difference of $44 - 9 = 35$ percentage points is very similar to the 30 percentage point difference obtained for the same comparison in the main analysis where we do not exclude cheaters (see Figure 3.3(a) above).

In a next step, we study how the *number* of dishonest group members affects the lying behavior of groups. In particular, we investigate whether the group behavior is mainly affected by the presence of at least one cheater (a *bad apple*) or, more generally, by the number of cheaters in the group. To study the effect of the number of cheaters in the group, we pool over all group sizes. We obtain the following result:

Result 9. The frequency of group lying increases with the number of cheaters in the group.

The result is illustrated in Figure 3.5. We find that the first cheater leads to an increase of the frequency of lying by 7 percentage points. But not only the first bad apple matters. Figure 3.5 also shows that, when the number of cheaters in the group increases to two and three, there is an additional (and almost linear) adverse effect on group lying behavior (Jonckheere-Terpstra test for presence of a trend, $p = 0.002$). As there are virtually no groups containing four or five cheaters, these observations are not shown in Figure 3.5 (but they are included when testing). As shown in the regression analysis of Section 3.4.5, this result is robust when controlling for group size.

[34] For $n = 3, 4, 5$, the respective (low) numbers of observations are 6 (5), 3 (5), and 4 (7), respectively.

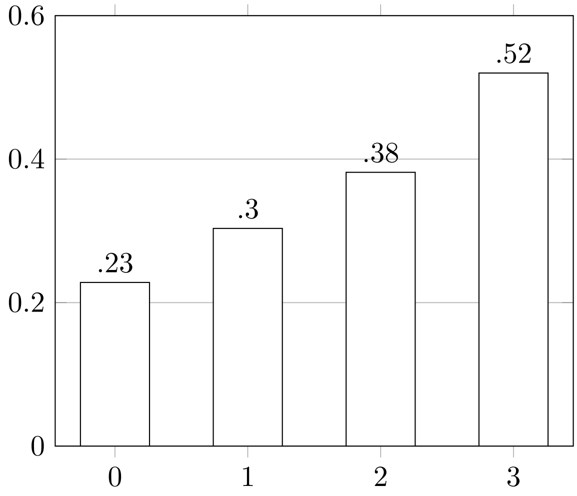


Figure 3.5: **Frequency of group lying by number of cheaters in group.**

Observations are pooled over all group sizes. This leads to 114, 145, 76 and 25 observations for groups with zero, one, two, and three cheaters, respectively. Groups with four cheaters (2 observations) and five cheaters (0 observations) are not displayed.

Our results are in line with previous findings obtained in contexts different than ours. For example, Dimmock, Gerken, and Graham (2018) empirically study work teams of financial advisors. They consider a setting of individual decision-making and study contagious effects of co-workers who previously committed misconduct (bad apples). They find that the probability that an individual commits misconduct increases with the number of bad apples in the work team. Moreover, in their experimental study of public good provision, De Oliveira, Croson, and Eckel (2015) show that group cooperation is negatively affected by the presence of highly selfish group members (bad apples). Similar to our result, they also find a gradual effect (i.e. a decline in group cooperation as the number of bad apples increases), rather than only the first bad apple being the main driver.

3.4.5 Robustness

In this section, we check the robustness of our main results on group lying behavior by conducting a regression analysis. This allows to additionally control for the observed die roll outcome and a host of personal characteristics of group members. The regression analysis confirms that group lying (i) increases with group size, (ii) is more prevalent in all-male groups, and (iii) increases with the number of cheaters in the group.

We estimate linear probability models where the unit of observation is a group. The dependent variable is the group's decision whether or not to lie about the die roll outcome (i.e. a dummy variable that is equal to one if the group lies and zero otherwise). We again confine attention to those 363 groups that had an incentive to lie to their advantage (i.e. that observed

	(1)	(2)	(3)	(4)
Group size	0.0586***	0.0718***	0.0588***	0.0534**
	(0.008)	(0.001)	(0.010)	(0.016)
All-male group		0.253***	0.246***	0.259***
		(0.000)	(0.000)	(0.000)
Number of cheaters in group			0.0505**	0.0427*
			(0.050)	(0.095)
Observations	363	363	363	363
Die roll outcome	Yes	Yes	Yes	Yes
Personal characteristics	No	No	No	Yes

Table 3.2: **Robustness of main results.**

All regressions estimate linear probability models where the dependent variable is whether or not a group lies. “All-male group” is a dummy variable indicating an all-male group. The last two rows indicate additional controls as follows: “Die roll outcome” represents dummy variables for the actual outcome of the die roll observed by the group. “Personal characteristics“ refers to group averages of members’ individual characteristics, i.e. responses to the six survey items on risk, time and social preferences, the five subscores of the IPIP Big-5 test, the Raven score, and the RMET score. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% level, respectively.

a die roll other than 5) and that did reach an agreement. In all regressions, we control for the observed die roll by including dummy variables.

The results are reported in Table 3.2, which displays the coefficients for our main variables of interest. Table 3.5 in Appendix 3.B provides the parameter estimates for all regressors included. With respect to the effect of group size (Result 1), Column (1) confirms a highly significant and positive effect of group size on the probability that a group lies (where the coefficient is stable across all specifications considered). Column (2) of Table 3.2 again documents that all-male groups stand out compared to all other group gender compositions (Results 3 and 4). In particular, the probability of lying in all-male groups is on average 25 percentage points higher, and this effect is again stable across specifications. Column (3) supports the finding that the probability that a group lies increases with the number of cheaters in the group (Result 9).

Finally, we also want to control for personal characteristics of group members.[35] We asked subjects for various individual self-assessments. This includes six validated survey items of the preference survey module of Falk, Becker, Dohmen, Huffman, and Sunde (2023) on risk, time, and social preferences. In addition, subjects provided self-assessments about personality traits in the Big-5 domain (Goldberg, 1992). To implement this, we follow Weidmann and Deming (2021) and use the (compact) 50-item IPIP scale, which yields individual measures of subjects’

[35] In the experiment, the group task was Task 1, and personal characteristics were elicited subsequently in Tasks 2, 4, and 5. Individual honesty preferences were elicited in Task 3. For details, see the instructions in Appendix 3.A.

extraversion, agreeableness, conscientiousness, emotional stability, and intellect/imagination.[36] Subjects received a flat payment of £1 for completing these self-assessments. Second, subjects were asked to solve a number of "Raven's Progressive Matrices" (Raven, 1995), where the Raven score is a widely used measure of IQ and abstract reasoning. Third, we elicited subjects' social intelligence through the well-established "Reading the Mind in the Eyes Test" (RMET, see Baron-Cohen et al., 2001). In the RMET, subjects are shown pictures of persons' eye areas and the respective subject needs to select one out of four adjectives that best describes what the displayed person is thinking or feeling (example in Appendix 3.A). In both, the Raven and the RMET task, we follow Weidmann and Deming (2021): Subjects had seven minutes to solve up to 14 Raven puzzles and were then asked to complete a 26-picture version of the RMET test. They obtained a flat payment for each task, which in our case amounts to £1.50.

For all personal characteristics elicited at the individual level, we construct measures at the group level by taking group averages. These averages are then used as controls in the regression reported in column (4) of Table 3.2. As can be seen, all earlier results are robust, solely the effect of the number of cheaters in the group becomes only marginally significant.

3.4.6 Groups with no agreement

So far, we have studied the behavior of the 363 groups (out of 385, so 94 percent, see Table 3.4 in Appendix 3.B) that observed a die roll $r \neq 5$ (i.e. had an incentive to lie) and have reached an agreement. We conclude this section by considering those 22 groups that observed a die roll $r \neq 5$, but did not reach an agreement. In all of these 22 groups, the disagreement arises because at least one group member did not make an entry in the decision window (see Section 3.3.1 above).[37] As explained in Section 3.3.1 above, we have implemented various technical checks to ensure that subjects can properly interact with their group fellows. We therefore hypothesize that the observed missing entries are intentional rather than due to technical problems. Moreover, if missing entries were due to technical problems, they should occur unsystematically.

However, this does not seem to be the case: We find that disagreement seems to cluster in almost-all-male and almost-all-female groups (six cases each, 54 percent of all disagreements). These group types account for 40.5 percent out of the 385 groups (see Table 3.4). By contrast, there are only two cases of disagreement each in all-male and all-female groups (i.e. 18 percent of all disagreements), while these two group types account for 42.3 percent of all group observations. When pooling both all-male and all-female groups and comparing them to the (also pooled) almost-all-male and almost-all-female groups, we find significantly more

[36] `https://ipip.ori.org/New_IPIP-50-item-scale.htm`

[37] In principle, a disagreement could also arise when all group members make entries that never match. However, this case did not occur in our experiment.

disagreement in the latter ($p = 0.032$, Chi2-test). No significant difference emerges for the comparison between all-male and almost-all-male groups, and all-female and almost-all-female groups, respectively. Hence, in contrast to lying behavior (see Result 4), there seems to be no gender-specific difference between these group types for the case of disagreement.

3.5 Conclusion

This paper is motivated by three related phenomena: First, decision-making by groups (rather than individuals) is ubiquitous. Second, this raises various (policy) questions related to group design such as group size and group composition. Third, a series of major corporate scandals (triggered by groups of employees of the respective firms) have attracted a lot of public and academic interest in the domain of unethical (or even illegal) decisions.

We present the results of an online experiment on unethical behavior by groups. Our primary research question is to study the impact of two crucial group characteristics, the group size and the group gender composition. To the best of our knowledge, this is the first paper to provide systematic evidence (i.e. within one study) on these issues. We adapt the widely used die-roll paradigm of Fischbacher and Föllmi-Heusi (2013) to a group setting, where each group member receives a monetary benefit when the group reaches a unanimous decision to lie about the outcome of a die roll. A total of 18 treatments captures all group sizes from two to five members, and for each group size, all possible combinations of female and male members. A second, methodological innovation of the paper is the design and implementation of a novel video chat extension for *oTree*, which is by now one of the standard programming languages to implement experiments. This video chat extension allows group members to communicate face-to-face in real time, thereby significantly enlarging the scope of communication in online experiments.

Our main findings can be summarized as follows: (i) larger groups lie more, (ii) all-male groups stand out in their proclivity to lie, (iii) already the first female group member induces a substantial honesty shift groups, and (iv) group behavior cannot be fully explained by members individual honesty preferences. In addition, we provide further results regarding the intensity of lying, as well as groups' decision and talking times. With respect to the current policy debate regarding the (gender) diversity of groups and female quotas, our findings suggest that in situations in which unethical behavior is potentially relevant, all-male groups are particularly "toxic" and should be avoided.

Our study establishes various stylized facts in a topical setting where systematic empirical evidence is still scant. In a next step, it would be interesting to explore underlying channels in more detail. Thereby, some of our findings shed some first light on possible mechanisms driving the patterns established by our experiment. For example, we find that (i) in line with earlier literature, females have somewhat stronger individual honesty preferences than males, while (ii) the honesty shift in almost-all-male groups (compared to all-male groups) is *not* accompanied by longer group discussions. Especially because of the second observation, these findings do not seem to support an explanation based on the idea that the (sole) females in almost-all-male groups finally convince their fellow male group members to refrain from lying in the course of intensive discussions. One potential explanation for the observed behavioral difference between all-male and almost-all-male groups is based on *gender-specific honesty beliefs*, i.e., males potentially believe that females have a quite strong preference for honesty. As a result, males might think that there is no point in trying to convince female group members to lie, and hence they anticipate that there would be no point in prolonged discussions. An alternative channel could be *gender-specific image concerns.* Abeler, Nosenzo, and Raymond (2019) document that for individual lying decisions, image concerns play a major role. Our results on group decisions could suggest that the strength of image concerns depends on the audience. In particular, males might be more concerned about their reputation vis-à-vis females than vis-à-vis other males in their group. This might make them less willing to lie whenever females are present.

These two channels (gender-specific honesty beliefs or image concerns) could potentially be disentangled in an additional treatment in which group decisions are taken by simple majority. For example, consider the case of triads. In case that gender-specific honesty beliefs are the key driver, there should be no difference in the frequency of lying between all-male and almost-all-male groups under simple majority (because the two males do not need the support of the allegedly honesty-minded female in order to implement a lie). However, if gender-specific image concerns are the key driver, then all-male groups would lie more than almost-all-male groups (because in the latter case, the mere presence of a female would prevent males from pushing in favor of a lie due to image concerns).

Appendix

3.A Instructions

Note: In this Appendix, we provide the instructions of our online experiment. Each heading (in bold) corresponds to a separate screen of the online experiment. At some points in the instructions, we include comments to the reader, which are marked "Note" and set in italics.

When subjects were dropped from the experiment due to the reasons discussed in Footnotes 19 and 20, they were informed accordingly, but we refrain from reporting these notification screens here.

Subjects were recruited on Prolific, where the following invitation was used: ***"Sign up for an online academic study with group and individual tasks*** *In this study, we ask you to perform a number of cognitive tasks and assessments. To participate in the study, you will need a desktop computer, laptop, or tablet (no smartphone) with a functioning camera and microphone. The study takes approximately 40 minutes."*

Welcome to this Study!

Welcome to this scientific study, which is conducted by a research team from the University of Hamburg and the University of Regensburg in Germany.

It will take you approximately 40 minutes, and you will have to answer questions and take decisions. As it is very important for us that you complete the whole study, please understand that you will only be paid upon doing so. You will earn a minimum of £4.00, but your individual payoff may be higher than this amount.

Please note that you can participate in this study only once.

Here is some general information about the procedures:

- We will first ask you for your Prolific ID and some socio-demographic information.
- Then we ask you to complete five tasks, which are independent from each other.
- In Task 1, you will interact in a group with other participants in a video chat. Your payoff may depend on the decisions by you and the other group members.
- Tasks 2 to 5 are individual tasks: There is no interaction with others, and your payoff only depends on your decisions.

- More information about the form and amount of payment will be provided at the beginning of each task.

We respect your anonymity. That is, we will never link your name with the data generated in this study. Moreover, we will not inform participants about either other participants' names or any other personal information.

For better readability, we recommend that you complete the study on a PC or Tablet (in landscape mode), not on a smartphone.

BEFORE CONTINUING, PLEASE CONFIRM THE FOLLOWINIG:

□ I AM WILLING TO INTERACT WITH OTHER PARTICIPANTS IN A VIDEO CHAT.

□ MY CAMERA AND MICROPHONE SEEM TO WORK FINE.

□ I UNDERSTAND THAT I WILL ONLY BE PAID IF I COMPLETE THE WHOLE STUDY.

Click here to proceed to the next screen.

Note: Subjects were only able to proceed to the next page when all three boxes were checked.

Demographics

We first would like to ask you for some socio-demographic information:

Question 1.1: Please enter your Prolific ID:

Question 1.2: What is your gender?

Question 1.3: How old are you (in years)?

Question 1.4: Which is the highest level of education you completed?

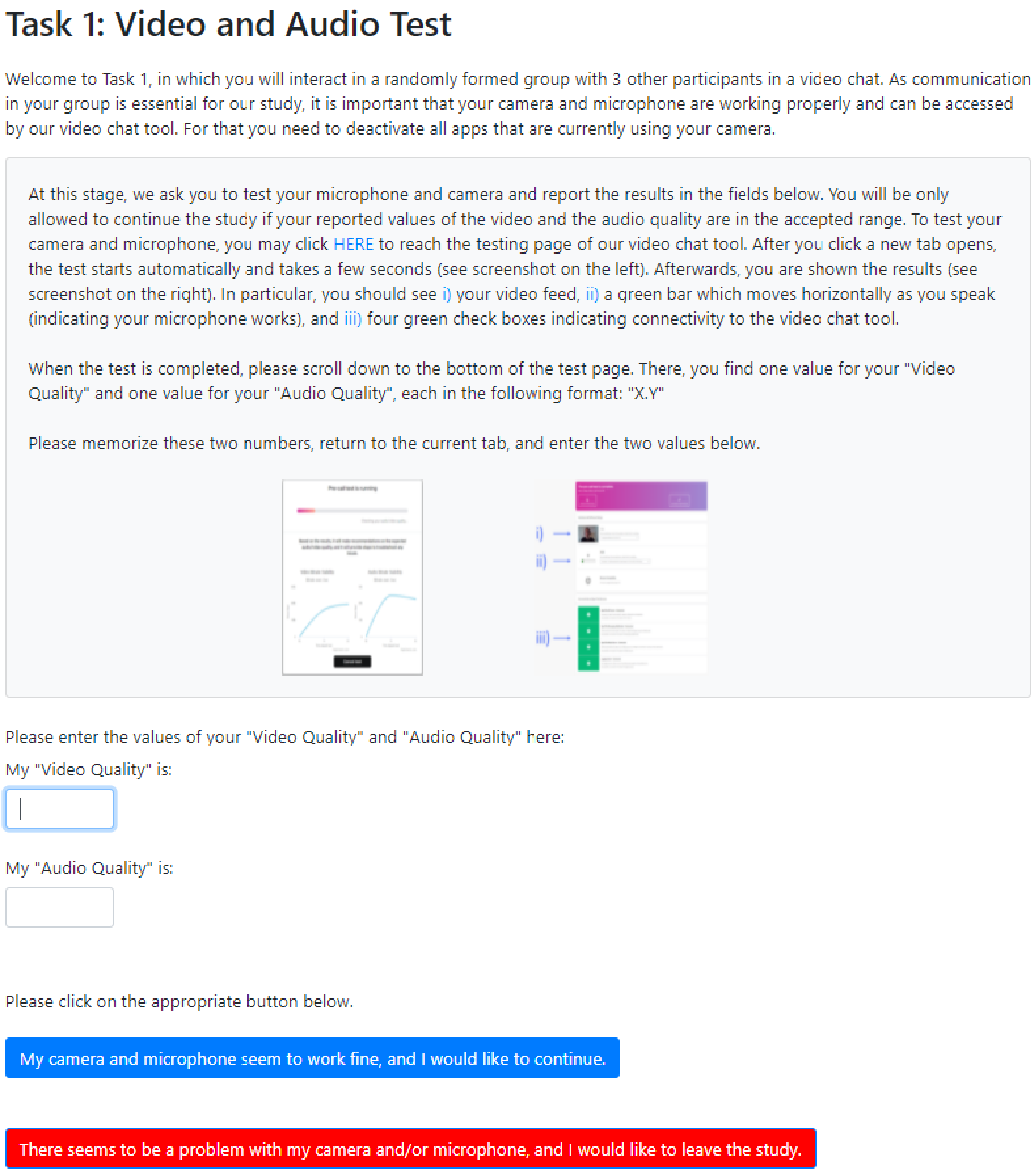
Question 1.5: What is your ethnical background?

Question 1.6: In which country do you currently reside?

Click here to proceed to the next screen.

Task 1

Note: As Task 1 is central to the present paper, on the following pages we provide screenshots of the four pages of our online experiment through which subjects proceeded in this task (plus an external test page) as displayed in Figure 3.1. On the page "Individual video and audio test", each subject individually had to perform a functionality test of their camera and microphone (see Figure 3.6), where Figure 3.7 provides a screenshot of the external test page. Subjects who successfully completed the functionality test then proceeded to the page "Individual general overview", where they received some basic information regarding the structure of the upcoming group interaction (see Figure 3.8). Afterwards, groups were formed, and group members met for the first time on the page "Group video and audio test" (see Figure 3.9). There, each group member had to confirm that they can see and hear all other group members before being able to proceed to page "Instructions and group decision" (see Figure 3.10), where subjects read the instructions for Task 1, were shown the die roll, and reports were made.



Task 1: Video and Audio Test

Welcome to Task 1, in which you will interact in a randomly formed group with 3 other participants in a video chat. As communication in your group is essential for our study, it is important that your camera and microphone are working properly and can be accessed by our video chat tool. For that you need to deactivate all apps that are currently using your camera.

At this stage, we ask you to test your microphone and camera and report the results in the fields below. You will be only allowed to continue the study if your reported values of the video and the audio quality are in the accepted range. To test your camera and microphone, you may click HERE to reach the testing page of our video chat tool. After you click a new tab opens, the test starts automatically and takes a few seconds (see screenshot on the left). Afterwards, you are shown the results (see screenshot on the right). In particular, you should see i) your video feed, ii) a green bar which moves horizontally as you speak (indicating your microphone works), and iii) four green check boxes indicating connectivity to the video chat tool.

When the test is completed, please scroll down to the bottom of the test page. There, you find one value for your "Video Quality" and one value for your "Audio Quality", each in the following format: "X.Y"

Please memorize these two numbers, return to the current tab, and enter the two values below.

Please enter the values of your "Video Quality" and "Audio Quality" here:

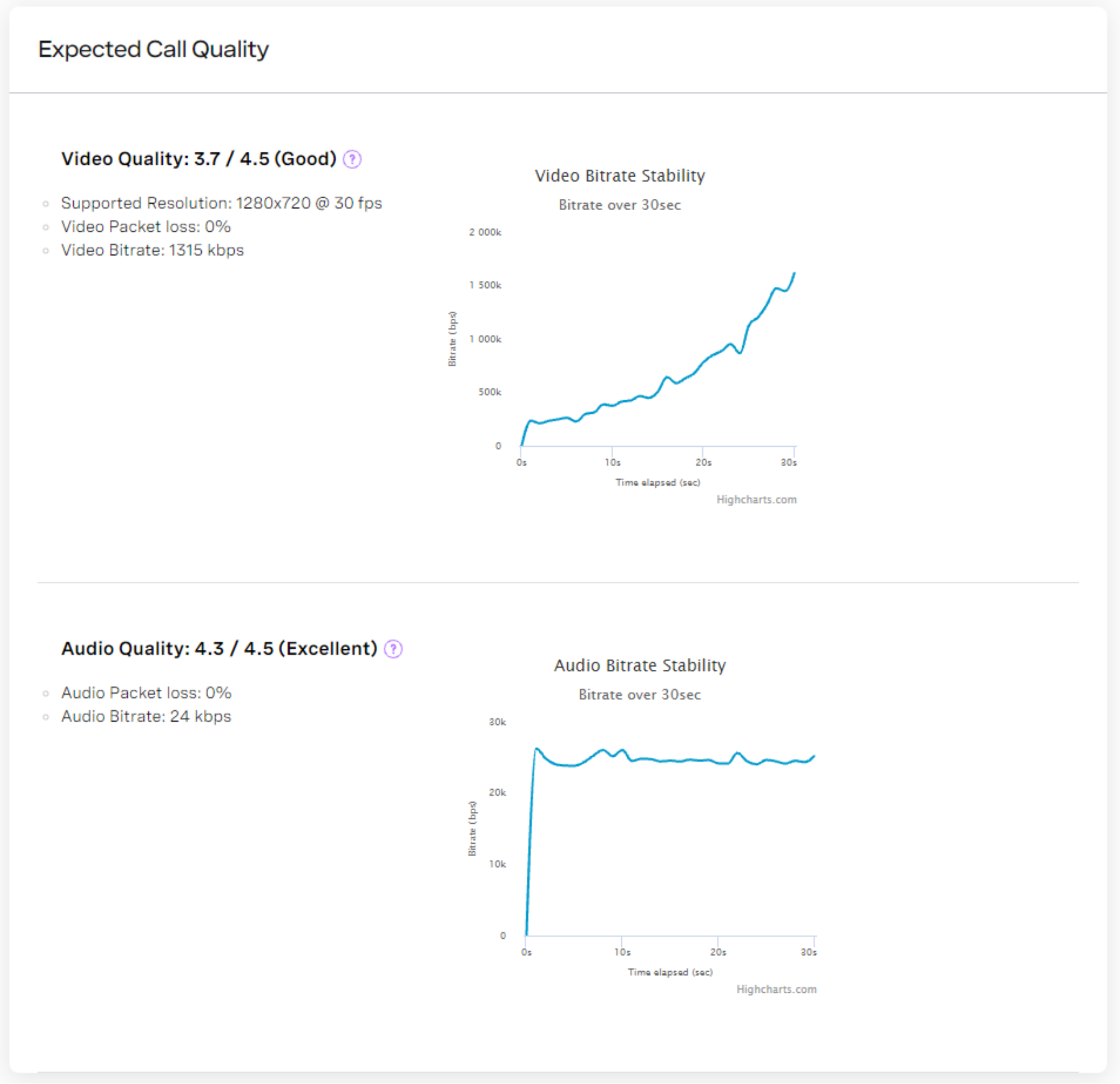
My "Video Quality" is:

My "Audio Quality" is:

Please click on the appropriate button below.

My camera and microphone seem to work fine, and I would like to continue.
There seems to be a problem with my camera and/or microphone, and I would like to leave the study.

Figure 3.6: **Screenshot of the page "Individual video and audio test".**



Expected Call Quality

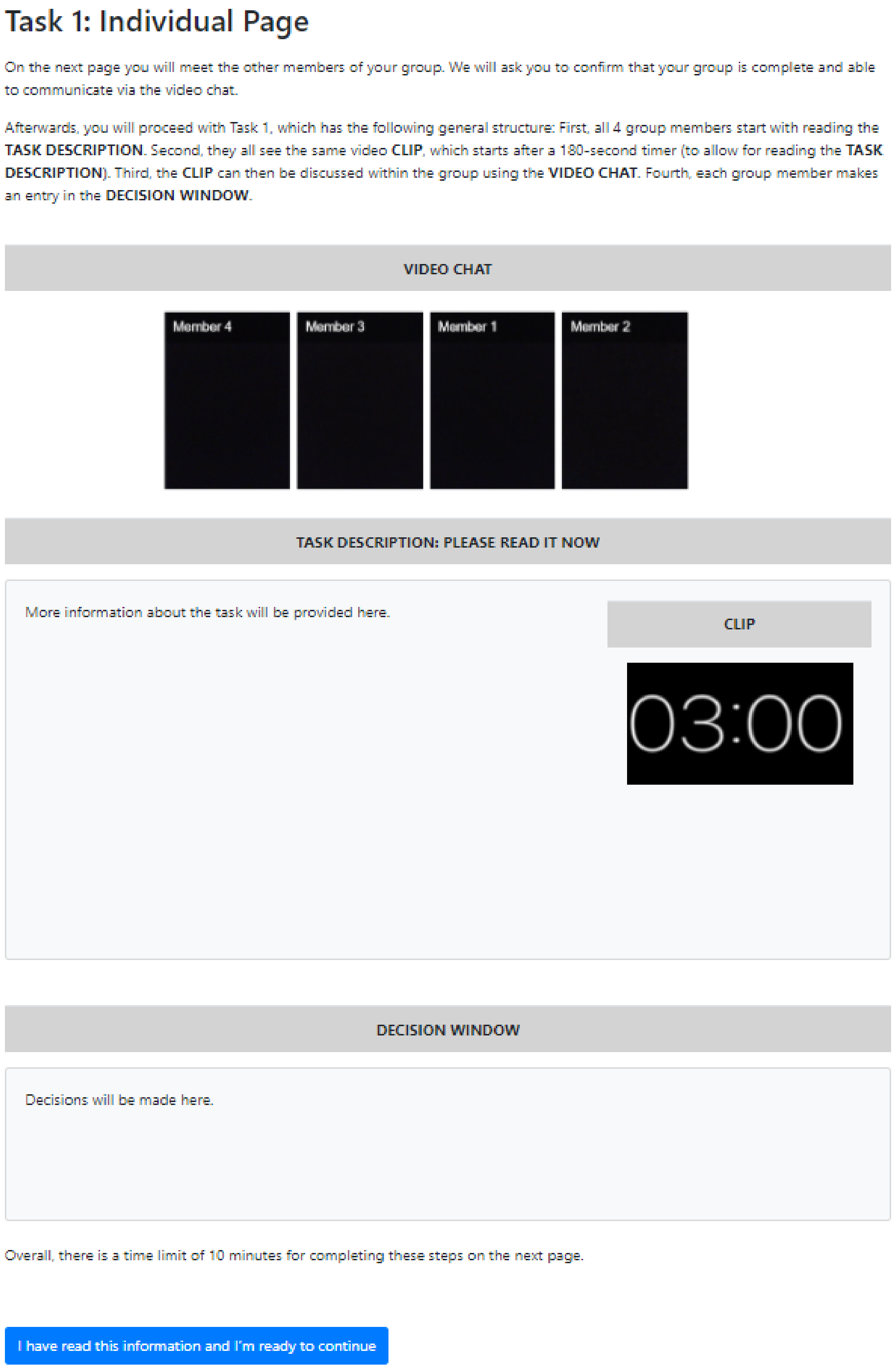
Video Quality: 3.7 / 4.5 (Good)

- Supported Resolution: 1280x720 @ 30 fps
- Video Packet loss: 0%
- Video Bitrate: 1315 kbps

Audio Quality: 4.3 / 4.5 (Excellent)

- Audio Packet loss: 0%
- Audio Bitrate: 24 kbps

Figure 3.7: **Screenshot of external test page.**



Task 1: Individual Page

On the next page you will meet the other members of your group. We will ask you to confirm that your group is complete and able to communicate via the video chat.

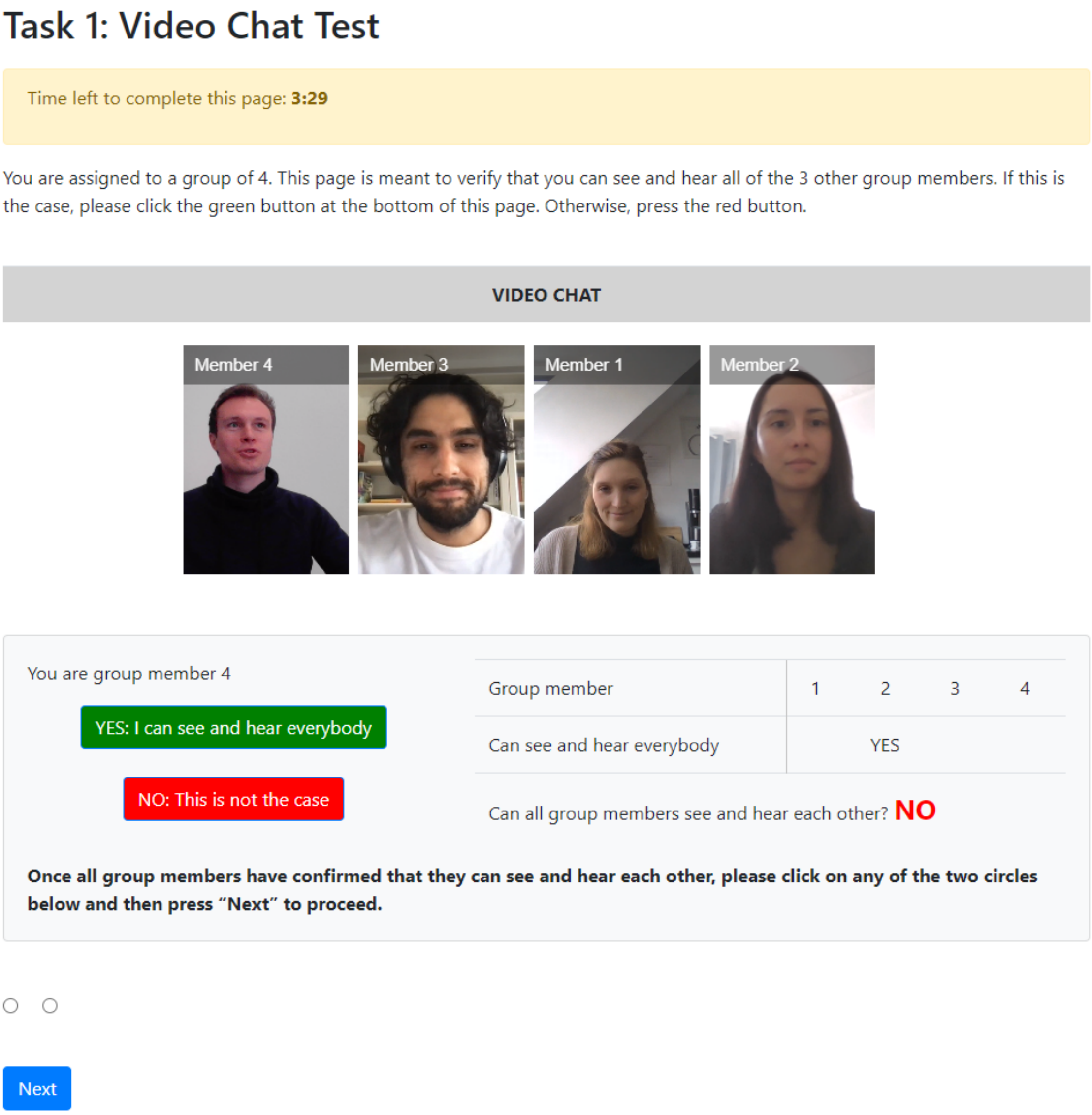
Afterwards, you will proceed with Task 1, which has the following general structure: First, all 4 group members start with reading the **TASK DESCRIPTION**. Second, they all see the same video **CLIP**, which starts after a 180-second timer (to allow for reading the **TASK DESCRIPTION**). Third, the **CLIP** can then be discussed within the group using the **VIDEO CHAT**. Fourth, each group member makes an entry in the **DECISION WINDOW**.

More information about the task will be provided here.

Decisions will be made here.

Overall, there is a time limit of 10 minutes for completing these steps on the next page.

Figure 3.8: **Screenshot of the page "Individual general overview".**



Task 1: Video Chat Test

Time left to complete this page: **3:29**

You are assigned to a group of 4. This page is meant to verify that you can see and hear all of the 3 other group members. If this is the case, please click the green button at the bottom of this page. Otherwise, press the red button.

You are group member 4

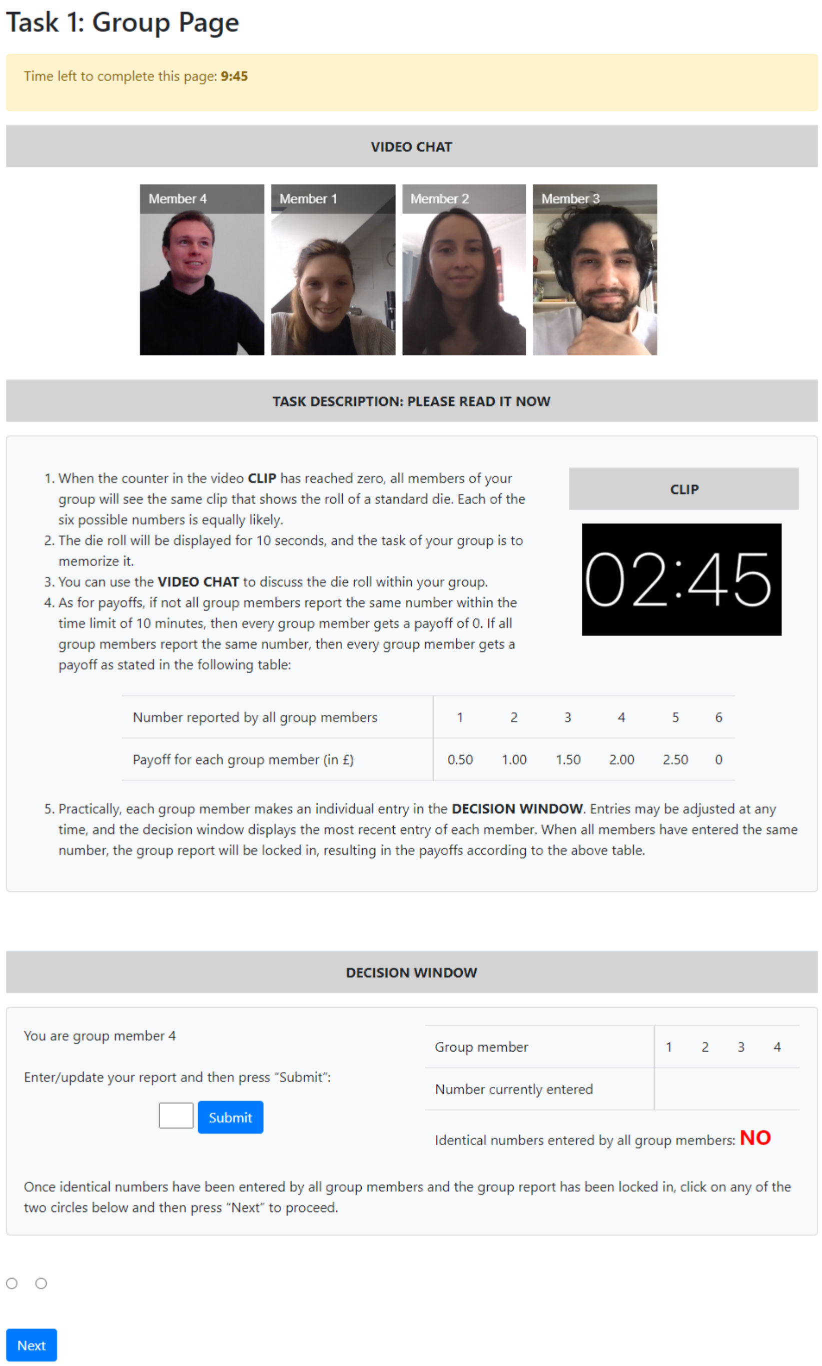
Group member	1	2	3	4
Can see and hear everybody		YES		

Can all group members see and hear each other? **NO**

Once all group members have confirmed that they can see and hear each other, please click on any of the two circles below and then press "Next" to proceed.

○ ○

Figure 3.9: **Screenshot of the page "Group video and audio test".**



Task 1: Group Page

Time left to complete this page: **9:45**

VIDEO CHAT

TASK DESCRIPTION: PLEASE READ IT NOW

1. When the counter in the video **CLIP** has reached zero, all members of your group will see the same clip that shows the roll of a standard die. Each of the six possible numbers is equally likely.
2. The die roll will be displayed for 10 seconds, and the task of your group is to memorize it.
3. You can use the **VIDEO CHAT** to discuss the die roll within your group.
4. As for payoffs, if not all group members report the same number within the time limit of 10 minutes, then every group member gets a payoff of 0. If all group members report the same number, then every group member gets a payoff as stated in the following table:

CLIP

02:45

Number reported by all group members	1	2	3	4	5	6
Payoff for each group member (in £)	0.50	1.00	1.50	2.00	2.50	0

5. Practically, each group member makes an individual entry in the **DECISION WINDOW**. Entries may be adjusted at any time, and the decision window displays the most recent entry of each member. When all members have entered the same number, the group report will be locked in, resulting in the payoffs according to the above table.

DECISION WINDOW

You are group member 4

Enter/update your report and then press "Submit":

□ Submit

Group member	1	2	3	4
Number currently entered				

Identical numbers entered by all group members: **NO**

Once identical numbers have been entered by all group members and the group report has been locked in, click on any of the two circles below and then press "Next" to proceed.

○ ○

Next

Figure 3.10: **Screenshot of the page "Instructions and group decision".**

Task 1: Completed

Thank you for completing Task 1.

The remaining four tasks are to be completed on your own.

Please click below to proceed to Task 2 (out of 5).

Task 2

Welcome to Task 2 in which we ask you for self-assessments.

You will receive a fixed payment of £1.00 for completing all questions in this task.

Question 2.1: How do you see yourself: Are you a person who is generally willing to take risks, or do you try to avoid taking risks?

Please use a scale from 0 to 10, where a 0 means you are “completely unwilling to take risks” and a 10 means you are “very willing to take risks”. You can also use the values in-between to indicate where you fall on the scale.

completely unwilling to take risks												very willing to take risks
	0	1	2	3	4	5	6	7	8	9	10	
	□	□	□	□	□	□	□	□	□	□	□	

Question 2.2: In comparison to others, are you a person who is generally willing to give up something today in order to benefit from that in the future?

Please use a scale from 0 to 10, where a 0 means you are “completely unwilling to give up something today” and a 10 means you are “very willing to give up something today”. You can also use the values in-between to indicate where you fall on the scale.

completely unwilling to give up something today												very willing to give up something today
	0	1	2	3	4	5	6	7	8	9	10	
	□	□	□	□	□	□	□	□	□	□	□	

Question 2.3: How well does the following statement describe you as a person? As long as I am not convinced otherwise, I assume that people have only the best intentions.

Please use a scale from 0 to 10, where 0 means “does not describe me at all” and a 10 means “describes me perfectly”. You can also use the values in-between to indicate where you fall on the scale.

does not describe me at all										describes me perfectly
0	1	2	3	4	5	6	7	8	9	10
□	□	□	□	□	□	□	□	□	□	□

Question 2.4: How do you assess your willingness to share with others without expecting anything in return when it comes to charity?

Please use a scale from 0 to 10, where 0 means you are “completely unwilling to share‘” and a 10 means you are “very willing to share”. You can also use the values in between to indicate where you fall on the scale.

completely unwilling to share										very willing to share
0	1	2	3	4	5	6	7	8	9	10
□	□	□	□	□	□	□	□	□	□	□

Question 2.5: Imagine the following situation: You are shopping in an unfamiliar city and realize you lost your way. You ask a stranger for directions. The stranger offers to take you with their car to your destination. The ride takes about 20 minutes and costs the stranger about 20 Euro in total. The stranger does not want money for it. You carry six bottles of wine with you. The cheapest bottle costs 5 Euro, the most expensive one 30 Euro. You decide to give one of the bottles to the stranger as a thank-you gift.

Which bottle do you give?

The bottle for: □ 5 Euro □ 10 Euro □ 15 Euro □ 20 Euro □ 25 Euro □ 30 Euro

Question 2.6: How do you see yourself: Are you a person who is generally willing to punish unfair behavior even if this is costly?

Please use a scale from 0 to 10, where 0 means you are “not willing at all to incur costs to punish unfair behavior” and a 10 means you are “very willing to incur costs to punish unfair behavior”. You can also use the values in-between to indicate where you fall on the scale.

not willing at all to incur costs to punish unfair behavior										very willing to incur costs to punish unfair behavior
0	1	2	3	4	5	6	7	8	9	10
□	□	□	□	□	□	□	□	□	□	□

Question 2.7: To which extent does each of the following 50 statements describe you? Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. Please indicate for each statement whether it is:

- 1 = very inaccurate
- 2 = moderately inaccurate
- 3 = neither accurate nor inaccurate
- 4 = moderately accurate
- 5 = very accurate

		1	2	3	4	5
1.	Am the life of the party	□	□	□	□	□
2.	Feel little concern for others	□	□	□	□	□
3.	Am always prepared	□	□	□	□	□
4.	Get stressed out easily	□	□	□	□	□
5.	Have a rich vocabulary	□	□	□	□	□
6.	Don't talk a lot	□	□	□	□	□
7.	Am interested in people	□	□	□	□	□
8.	Leave my belongings around	□	□	□	□	□
9.	Am relaxed most of the time	□	□	□	□	□
10.	Have difficulty understanding abstract ideas	□	□	□	□	□
11.	Feel comfortable around people	□	□	□	□	□
12.	Insult people	□	□	□	□	□
13.	Pay attention to details	□	□	□	□	□
14.	Worry about things	□	□	□	□	□
15.	Have a vivid imagination	□	□	□	□	□
16.	Keep in the background	□	□	□	□	□
17.	Sympathize with others' feelings	□	□	□	□	□
18.	Make a mess of things	□	□	□	□	□
19.	Seldom feel blue	□	□	□	□	□
20.	Am not interested in abstract ideas	□	□	□	□	□
21.	Start conversations	□	□	□	□	□
22.	Am not interested in other people's problems	□	□	□	□	□
23.	Get chores done right away	□	□	□	□	□
24.	Am easily disturbed	□	□	□	□	□
25.	Have excellent ideas	□	□	□	□	□
26.	Have little to say	□	□	□	□	□
27.	Have a soft heart	□	□	□	□	□
28.	Often forget to put things back in their proper place	□	□	□	□	□
29.	Get upset easily	□	□	□	□	□
30.	Do not have a good imagination	□	□	□	□	□
31.	Talk to a lot to different people at parties	□	□	□	□	□
32.	Am not really interested in others	□	□	□	□	□
33.	Like order	□	□	□	□	□

34.	Change my mood a lot	□	□	□	□	□
35.	Am quick to understand things	□	□	□	□	□
36.	Don’t like to draw attention to myself	□	□	□	□	□
37.	Take time out for others	□	□	□	□	□
38.	Shirk my duties	□	□	□	□	□
39.	Have frequent mood swings	□	□	□	□	□
40.	Use difficult words	□	□	□	□	□
41.	Don’t mind being the center of attention	□	□	□	□	□
42.	Feel others’ emotions	□	□	□	□	□
43.	Follow a schedule	□	□	□	□	□
44.	Get irritated easily	□	□	□	□	□
45.	Spend time reflecting on things	□	□	□	□	□
46.	Am quiet around strangers	□	□	□	□	□
47.	Make people feel at ease	□	□	□	□	□
48.	Am exacting in my work	□	□	□	□	□
49.	Often feel blue	□	□	□	□	□
50.	Am full of ideas	□	□	□	□	□

Click below to proceed to the next screen.

Task 2: Completed

Thank you for completing Task 2.

Click below to proceed to Task 3 (out of 5).

Task 3

Welcome to Task 3, where we ask you to complete a short music quiz.

For this task, there is no fixed payment, but you will receive a payment of £0.50 when correctly answering ALL of the following six Questions 3.1 to 3.6.

Please answer Questions 3.1 to 3.6 on your own, without looking them up elsewhere.

For each question, please select one answer from the respective pull-down list.

Question 3.1: Who wrote the composition “Für Elise”?

Question 3.2: What is Lady Gaga’s real first name?

Question 3.3: Name the lead singer of the rock group Nirvana.

Question 3.4: In what year was Claude Debussy born?

Question 3.5: How many valves are there on a standard modern trumpet?

Question 3.6: Name the town where Michael Jackson was born.

Click below to proceed to the next screen.

Task 3: Completed

Thank you for completing Task 3.

You will receive feedback with respect to your performance in Task 3 at the very end of this study.

Click below to proceed to Task 4 (out of 5).

Task 4

Welcome to Task 4 in which we ask you to solve a number of puzzles.

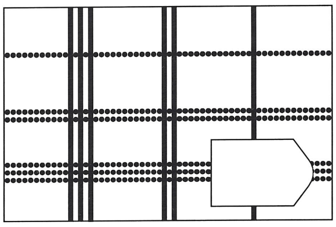
This task will take at most seven minutes, and you will receive a fixed payment of £1.50 for working on it.

Each puzzle has the same basic structure as the example below.

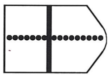
You are asked to recognize the pattern in the upper part of the puzzle by going through the different fields vertically and horizontally. Then choose the appropriate piece out of the eight possible answers provided in the lower part.

In the example below, piece 8 is the correct answer.

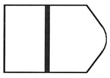
EXAMPLE:



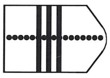
1



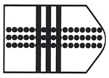
2



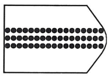
3



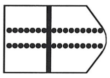
4



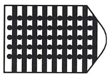
5



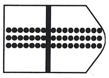
6



7



8



Your answer: □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 ■ 8

Question 4.1: There is a total of **14 puzzles** to be solved **within a time limit of seven minutes**. For each puzzle, select the correct answer.

Click below to proceed to the next screen.

Note: Each of the 14 puzzles of Task 4 was displayed on a separate page. These puzzles have been omitted from this Appendix. Once subjects reached the 7-minute time limit, they were notified of this fact and directed to Task 5.

Task 4: Completed

Thank you for completing Task 4.

You will receive feedback with respect to your performance in Task 4 at the very end of this study.

Click below to proceed to the final Task 5.

Task 5

Welcome to Task 5, for which you will receive a fixed payment of £1.50.

In this task, you will see 26 pictures, each showing a set of eyes like the one in the example below, together with four words.

EXAMPLE:

□ jealous □ **panicked** □ arrogant □ hateful

Question 5.1: For each set of eyes, choose and mark which word best describes what the person in the picture is thinking or feeling. You may feel that more than one word is applicable but please choose just one word, the word which you consider to be most suitable. Before making your choice, make sure that you have read all 4 words. You should try to do the task as quickly as possible but you will not be timed. If you really don't know what a word means you can look it up HERE.[38]

Note: Each of the 26 pictures of Task 5 was displayed on a separate page. These pictures have been omitted from this Appendix.

Click below to proceed to the next screen.

Thank you!

Note: The numbers stated on this page are meant as an example.

You have now completed all tasks.

Here is your payoff summary:

- Task 1: Payment of £2.50, because your group agreed to report a 5.
- Task 2: Fixed payment of £1.00.
- Task 3: You have correctly answered 2 of the 6 questions. As you only receive a payment for this task when you have answered all questions correctly, your payoff is: £0.00.
- Task 4: Fixed payment of £1.50. For your information only: You have correctly solved 10 of the 14 puzzles.
- Task 5: Fixed payment of £1.50.

Hence, your total payment is: £6.50. It will be transferred to your Prolific account.

Thank you again for participating in this study!

Have a nice day!

[38] By clicking on "HERE" subjects were directed to a pre-defined word list that is part of the RMET package.

3.B Additional figures and tables

(a) All groups

	No. of females per group						
Group size	0	1	2	3	4	5	Total
2	21	24	19	-	-	-	64
3	21	21	22	21	-	-	85
4	20	23	22	24	20	-	109
5	20	21	22	22	21	21	127
	82	89	85	67	41	21	385

(b) Only groups that reach an agreement

	No. of females per group						
Group size	0	1	2	3	4	5	Total
2	20	23	19	-	-	-	62
3	20	20	21	20	-	-	81
4	20	20	20	20	20	-	100
5	20	20	20	20	20	20	120
	80	83	80	60	40	20	363

Table 3.4: **Treatments with die roll outcome $\neq 5$.**

	(1)	(2)	(3)	(4)
Group size	0.0586***	0.0718***	0.0588***	0.0534**
	(0.008)	(0.001)	(0.010)	(0.016)
All-male group		0.253***	0.246***	0.259***
		(0.000)	(0.000)	(0.000)
Number of cheaters in group			0.0505**	0.0427*
			(0.050)	(0.095)
Die Roll=1	0.0649	0.0675	0.0780	0.0846
	(0.396)	(0.364)	(0.294)	(0.241)
Die Roll=2	0.101	0.0988	0.0980	0.119*
	(0.179)	(0.179)	(0.181)	(0.096)
Die Roll=3	-0.0992	-0.0846	-0.0770	-0.0315
	(0.211)	(0.273)	(0.317)	(0.676)
Die Roll=6	0.188**	0.174**	0.172**	0.203***
	(0.013)	(0.019)	(0.020)	(0.005)
Group: Average Raven score				0.0315**
				(0.015)
Group: Average RMET score				0.0165
				(0.189)
Group: Average risk attitude				0.0765***
				(0.000)
Group: Average time preference				0.0149
				(0.468)
Group: Average trust				0.00477
				(0.808)
Group: Average altruism				0.0217
				(0.380)
Group: Average positive reciprocity				-0.00414
				(0.455)
Group: Average negative reciprocity				-0.0241
				(0.176)
Group: Average BIG-5 Score Extraversion				0.00482
				(0.384)
Group: Average BIG-5 Score Agreeableness				-0.00293
				(0.709)
Group: Average BIG-5 Score Conscientiousness				-0.00688
				(0.286)
Group: Average BIG-5 Score Emotional Stability				-0.00997**
				(0.032)
Group: Average BIG-5 Score Intellect Imagination				-0.0112
				(0.123)
Observations	363	363	363	363
Die roll outcome	Yes	Yes	Yes	Yes
Personal characteristics	No	No	No	Yes

Table 3.5: **Robustness of main results (all coefficients).**

The table note for Table 3.2 applies. The reported regressions are the same, but here the coefficients of all included independent variables are shown in the table. With respect to the dummies die roll outcomes, an outcome of four serves as the reference category and is hence omitted. Recall that the analysis only considers those groups which had an incentive to lie, and hence all observations with a die roll outcome of five are excluded.

Chapter 4.

ChaTree - video chat integration in oTree[1]

Abstract

ChaTree is an open-source, online software that provides a video chat tool for online experiments using oTree. The video chat is integrated in a web-page and can be directly accessed after entering the page. With video conferences, ChaTree adds another dimension to oTree-based experiments, widening the experimental space to explore digital behavior featuring face-to-face communication. ChaTree is able to capture audio levels for each participant, offering an instrument for the investigation of dynamic group decision-making. The video chat tool has been successfully applied in an experiment with up to five participants per group chat.

Keywords: Group video chat, online experiments, experimental software

JEL Classification: C88, C92

4.1 Introduction

In experimental economics, z-Tree (Fischbacher, 1999, 2007) has been used for more than two decades to implement experiments in university-based laboratories. This software was recently accompanied by oTree (Chen et al., 2016), an open-source, online software which can be used online, in the lab, in the field, and in classrooms. oTree feels like an update to the pioneer z-Tree, providing nearly limitless opportunities for creating experimental games to study economic behavior, much like z-Tree but without spacial constraints. Adding a digital environment as an experimental laboratory simplifies participation and thereby increases the potential pool of participants nearly infinitely. Next to amplifying the number of participants, online participation also increases the representativeness of the participating group, as university laboratories are usually limited to students. After z-Tree, oTree is the next step to recreate and investigate all sorts of interactive environments. As an add-on to oTree, the new software package *ChaTree*, which is presented in this paper, offers the opportunity to study face-to-face communication not only in a lab environment but also in online experiments via video chat.

So far, the only experimental studies in economics using remote face-to-face communication have been executed in a lab employing a complex combination of software and equipment. However, they all find that communication in video conferences does not produce significantly different outcomes compared to in-person face-to-face communication. Bos et al. (2001) examine behavior in a trust game, Brosig and Weimann (2003) study cooperation and Grözinger et al.

[1] This chapter is co-authored by Jan-Patrick Mayer, Gerd Muehlheusser, Andreas Roider, Eugen Tereschenko and Niklas Wallmeier.

(2020) investigate creative collaboration, with each of them comparing a similar set of communication channels. Consistently over all three studies, participants performed similarly well using remote and non-remote face-to-face communication while significantly outperforming the other communication channels like text chat or audio conferences. These findings suggest that studies involving face-to-face communication can also, in addition to the lab, be implemented online.

Next to the finding that in-person face-to-face communication yields similar results as communication through video conferences, studying behavior in a video chat is of important value in itself. Even before the COVID-19 pandemic, working from home and meeting online were practices slowly increasing in usage. Since the pandemic, this increase has been highly amplified. Von Gaudecker et al. (2020) analyze labour supply data of the Netherlands, a model state of western Europe, to find that working hours at the workplace and at home split almost equally in late March 2020. Bick et al. (2023) report that, in the U.S., the share of hours worked from home increased from 14.4% in February 2020 to 39.6% in May 2020. While these numbers are strongly affected by drastic measures to contain the pandemic in 2020, working from home became a valid alternative for various occupations in the years after. With these transforming work environments, video conferences very quickly became a common communication tool. Therefore, investigating behavior in and with video conferences will provide valuable insights.

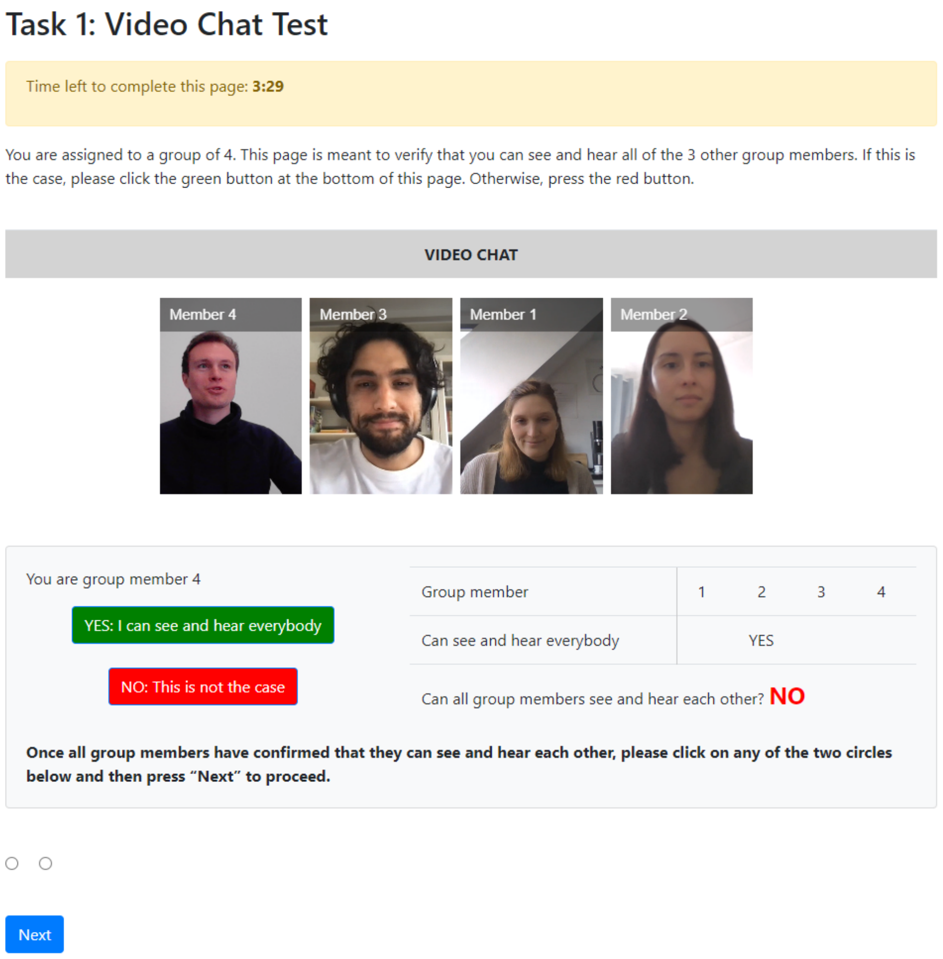
4.2 Features

The elegance of ChaTree consists of the ease of usage, the small margin of participant application error, and the protection of data privacy. The video chat is directly embedded in one experimental screen and users automatically participate once camera and microphone access are allowed. Usernames are predetermined by the experimenter and the user interface of the video chat is very plain, allowing each participant to fully focus on the other group members. The only active option for participants is to mute/unmute each user in their video conference (including themselves) to mitigate disturbances from noisy backgrounds.

The video chat tool was successfully applied in Promann (2024) for teams of two and in Muehlheusser et al. (2024) for group sizes of two, three, four, and five. The number of video conference participants is not limited by the software itself, but every incoming and outgoing video stream is hosted by each participant's device. Hence, the video chat's functionality will be impaired by increasing the number of participants and the quality of the worst participating device provides a natural limit. Poor quality video conferences can be mitigated by adding a screen for testing the video and audio quality of each participant's device[2]. To further ensure

[2] E.g., `https://tokbox.com/developer/tools/precall/results?scalableVideo=true` provides such a test page.

a fully functional video chat conference for each experimental group, a video chat test page, as shown in Figure 4.1, is recommended.



Task 1: Video Chat Test

Time left to complete this page: **3:29**

You are assigned to a group of 4. This page is meant to verify that you can see and hear all of the 3 other group members. If this is the case, please click the green button at the bottom of this page. Otherwise, press the red button.

VIDEO CHAT

You are group member 4

YES: I can see and hear everybody

NO: This is not the case

Group member	1	2	3	4
Can see and hear everybody		YES		

Can all group members see and hear each other? **NO**

Once all group members have confirmed that they can see and hear each other, please click on any of the two circles below and then press "Next" to proceed.

○ ○

Next

Figure 4.1: **Video chat test.**

On this page, participants test whether the video chat is working flawlessly for every group member. They automatically join in a video chat with all group members and have to confirm that audio and video communication is working.

In addition, the oTree add-on chaTree provides the opportunity to capture audio levels of talking individuals. This means that every time a video chat participant speaks, the increased audio level will be registered for this participant. Hence, each change of speaker as well as the total amount of talking time for each participant will be captured. This provides another instrument for analyzing the dynamics of group decision-making and the impact of dominant speakers on the outcome of a group discussion.

4.3 Requirements and Implementation

The video chat embedment is provided for oTree5 on GitHub[3]. Application in an oTree3 environment is possible as well with some minor alterations.

[3] `https://github.com/TimoPromann/ChaTree`

The current version of the video chat's implementation code requires a Video API account of *Vonage*, the video chat provider, and a *Mongo DB* account to acquire a database for temporary data storage. The primary use of the database is to store identification keys of video chat participants. This way, a group that is matched within oTree will also be matched within one video chat room while the next group is sent to a different room. In addition to the identification keys, only the participants' voice levels are stored on the database. The identification keys consist of various letters and integers in random order while the voice levels are integers between 0 and 1. Therefore, no private information of any participant is externally stored. However, the database can easily be substituted if required.

The GitHub repository provides detailed instructions on how to integrate the video chat in an oTree project. Next to the code for the video chat, the repository supplies code for the voice level capture. A README file will provide guidance on how to set up a Vonage API and a Mongo DB account. The README will further point out necessary alterations of settings. Using Vonage API will entail costs of around 10$ per month. Mongo DB offers one database model with very limited storage capacities, which is free of charge. As the data stored on the database is marginal, the free model will suffice. The server used for hosting the experiment will not bear the main load of computing capacity as every incoming and outgoing video stream is hosted by each participant's device. Nevertheless, the data traffic will reach higher magnitudes than with a simple survey, so that using a reasonably small session size is recommended. Sessions with around 50 participants simultaneously interacting in multiple video chats were conducted flawlessly in Muehlheusser et al. (2024). It is recommended to test the final setup with as many devices, browsers, and operation systems as possible and adjust as needed in case of unforeseen issues.

On Social Norms and Observability in (Dis)honest Behavior[1]

Abstract

Transparency and observability have been shown to foster ethical decision-making as people tend to comply with an underlying norm for honesty. However, in situations implying a social norm for *dis*honesty, this might be different. In a die-rolling experiment, we investigate whether observability can also have detrimental effects. We thus introduce a norm nudge toward honesty or dishonesty and make participants' decisions observable and open to the judgement of other participants in order to manipulate the observability of people's decisions as well as the underlying social norm. We find that a nudge toward honesty indeed increases the level of honesty, suggesting that such a norm nudge can successfully induce behavioral change. Our introduction of social image concerns via observability, however, does not affect honesty and does not interact with our norm nudge.

Keywords: Lying, cheating, social norms, image concerns, nudging, behavioral change

JEL Classification: C91, C92, D01, D91, M14

5.1 Introduction

People frequently engage in dishonest behavior, which can - depending on the context - entail large costs to individuals, organizations, and society more broadly. For policy design that effectively fosters honesty, it is thus pivotal to explore and identify relevant behavioral mechanisms. In the present study, in particular, we focus on social norms, people's preferences for being seen as honest, and the way these concepts interact. For that purpose, we set up an experiment in which we vary whether information about others' reporting represents team members as being honest or dishonest, and whether own reporting behavior is observable to team members.

A recent meta-analysis of experimental studies on (dis)honesty concluded that people tend to have a preference for being honest and for *being seen as honest*, i.e., people have image concerns (Abeler et al., 2019). Previous research shows that *norms* also play an important role in people's prosocial behavior (e.g., Bicchieri, 2016; Bicchieri et al., 2022; Cialdini et al., 1990; Krupka and Weber, 2013) and that "whenever individuals believe they are expected by their group ... to behave according to a given standard, and also expect the norm to be generally followed, they usually comply." (Bicchieri and Muldoon, 2011).

In many decision situations, a norm for honesty prevails. If this is the case, then the previous literature suggests that observability - and thus image concerns - encourages honesty. However,

[1] This chapter is co-authored by Christoph Huber, Christos Litsios and Annika Nieper and has been published as Huber et al. (2023) in the *Journal of Economic Behavior & Organization.*

it is easy to conceive of decision situations in which there appears to be a norm for *dis*honesty - that is, in which one is expected to comply with others' *dis*honest behavior. In those cases, we propose that it is reasonable to expect observability to encourage *dis*honesty instead. Consider the example of fraudulent practices in organizations, for instance, where there prevails a social norm - as well as an expectation - for unethical behavior internally, which might differ from external norms. Accordingly, this mechanism may result in corruption, fraud, and related malfeasance (e.g., Slemrod, 2007; Dyck et al., 2023). Similar applications include people observing the behavior of any reference group, their peer group, or even their leaders, and thus learn about the prevailing norm, which might just as well not be ethical or honest in nature (e.g., Acemoglu and Jackson, 2015; Ajzenman, 2021).[2] Examining the role of observability under different social norms is thus key to advancing our understanding of (dis)honest behavior more generally, and it directly relates to topics that span important domains in economics and the social sciences.

Previous research has shown that targeting people's concern for being seen as honest can increase honesty (Bo et al., 2015; Bodenschatz and Irlenbusch, 2019). People cheat less in an experimental die-rolling task if they are observed by others (e.g., Gneezy et al., 2018; Fries et al., 2021; Bašić and Quercia, 2022). Nevertheless, increasing people's concern for how others view them can also have no effect on honesty; it can even increase dishonesty. This can be seen, for instance, when one person, who benefits from mutual dishonesty, can observe another's decision (Weisel and Shalvi, 2015; also see Kocher et al., 2018). These mixed findings might result from differing expectations of others to behave (dis)honestly based on the situation, as different situations can evoke different social norms (Akerlof and Kranton, 2005). Observing others' actual behavior, in particular, may alter the empirical expectations on what constitutes the social norm in a given context (Bicchieri, 2016). A number of previous studies have used such information provision as a social norm nudge - that is, to evoke perceptions of "good" and "bad" norms (Bicchieri and Dimant, 2019; see Bicchieri and Xiao, 2009, for example, in Dictator games; and Isler and Gächter, 2022, in honesty and cooperation).

The present study stands out in this body of research because it applies at least two novel design elements: first, we introduce the interaction between information provision (observing other participants' decisions) and observability (one's decision is observed by other participants); second, both information provision and observability relate to the same group of peers - i.e., the participants one observes in a first stage are precisely those who then observe one's own decision in a second stage. With this novel setup and the introduction of an additional

[2] If there is widespread dishonesty in a specific context and if that is widely known or perceived, then one might infer that there is a social norm for dishonesty, which may affect others' behavior. Important applications include tax evasion (Slemrod, 2007), insider trading (Acharya and Johnson, 2010), academic cheating (Jensen et al., 2002), or misrepresenting information in politics (Swire-Thompson et al., 2020), among other examples.

stage in which the same peers provide feedback on one's decision, we also introduce an element of accountability that is not present in previous work.

In this study, we experimentally manipulate observability and examine the role of social image concerns in (dis)honesty decisions under different social norms. If norms shape behavior and observability affects adherence to norms, then the introduction of observability should increase honestand dishonest behavior alike, in line with the respective norms. We thus propose that the ability of image concerns to increase honesty relies on a social norm for honesty. If there prevails a social norm for dishonesty instead, image concerns might even have a backfiring effect. We test this proposition in an experimental die-rolling task with a factorial treatment design, in which we manipulate people's perception of the prevalent social norm, as well as their social image concerns. Following Bicchieri and Dimant (2019), we induce a social norm nudge by manipulating subjects' empirical expectations on how other people would act in a comparable situation. For that, each participant is grouped with three other participants and then presented with (social) information about those participants' decisions. To vary subjects' social image concerns, we manipulate the observability and accountability of their decisions: in one set of conditions, a participant's decision remains unobserved by others, while in another set of conditions their decision is shown precisely to the reference group of participants used to induce the norm nudge. Note, however, that in contrast to previous literature, neither the framing (e.g., Dimant et al., 2020) nor the rules of the game (e.g., Serdarevic, 2021) differ between treatments.

Our results demonstrate that social norm nudges can be effective in mitigating dishonest behavior. We find significantly less dishonesty when participants are presented with others' honest behavior, compared to when they are presented with others' dishonest behavior. In our anonymous online setting, we find no effect of observability on people's (dis)honesty decisions. Importantly, our design also allows us to elicit empirical and normative expectations on the inherent social norm in each individual treatment condition. For that purpose, we implement a belief elicitation inspired by the Krupka and Weber (2013) paradigm: Participants are asked to assess the social appropriateness of lying in the die-roll task, as well as the prevalence of actual lying, and are incentivized in a coordination game setup; specifically, they receive a bonus payment if they anticipate the most frequently given assessment (for a more general discussion on the use of coordination games as an elicitation method, see Schmidt et al., 2022). Although we do observe a significant uptick in lying when norms shift toward dishonesty, we note that lying is, on average, not seen as socially appropriate in any of our treatments.

This study contributes to the literature on dishonest behavior in a number of important ways. First, we add to our understanding of how social norms, as well as transparency and observability, can shape (dis)honest behavior. To the best of our knowledge, the present study is the first to explicitly look into a potential interaction effect between these well-established drivers of (dis)honesty and thereby shed light on a potential damaging effect of increasing observability. We apply peer observation as norm-nudge stimuli (e.g., Bicchieri and Dimant, 2019) and thereby follow recent calls for examining "peer-nudging" as a means of changing social-norm perceptions (Isler and Gächter, 2022), and, eventually, actual behavior.[3] This allows us to identify both empirical and normative expectations as important factors in determining honesty. And, indeed, our results show that observing a reference group of peers behaving in a certain way does shift participants' expectations - even in an anonymous online setting. This provides evidence of a norm nudge that can successfully induce behavioral change in ethical decision-making. In a similar vein, with our novel experimental setup in an anonymous online framework, we also contribute to recent discussions on dishonesty in the digital age (Cohn et al., 2022).

The remainder of this paper is organized as follows. In Section 5.2, we review relevant literature related to our research questions and outline the hypotheses to be tested. Section 5.3 describes our experimental design and procedure. In Section 5.4, we present our main results. Section 5.5 concludes.

5.2 Related Literature and Hypotheses

Ever since Becker (1968) published his rational crime model, economists and other social scientists have studied in what situations and with what motives people behave honestly. Of greatest interest are situations in which honest behaviour departs from the standard economic model's prediction - that is, when lying would maximize payoffs. In their seminal work, Fischbacher and Föllmi-Heusi (2013, FFH) reported 39% of experimental participants to be "fully honest." Much of the following economics literature on lying and dishonesty uses, adapts, and extends their clever but simple incentivized experiment: Participants privately roll a six-sided die, are asked to memorize the number that came up, and are subsequently asked to report this number, where different numbers are associated with different monetary payoffs. Lying in this task is defined as a participant reporting a number different than the one they rolled and is measured by the overall distribution of reports in lieu of observing individuals' die rolls to maintain anonymity. This and a number of related experiments have identified lying costs, i.e., intrinsic costs derived from deviating from truth-telling, which foster lying aversion (e.g., Ellingsen and Johannesson, 2004; Kartik, 2009; Gneezy et al., 2013, 2018; Abeler et al., 2019).

[3] A related yet distinct concept is "meta-nudging" (Dimant and Gesche, 2023; Dimant and Shalvi, 2022), which aims to achieve behavioral change by nudging people indirectly via "social influencers"; that is, those with the ability to enforce other's behavior and norm adherence. "Peer-nudging," by contrast, refers to nudging people through social information about their peers.

In a meta study of 90 different experiments using variants of the FFH design, Abeler et al. (2019) concluded that overall, people “lie surprisingly little” and named “a preference for being seen as honest and a preference for being honest” as the primary motivations (Abeler et al., 2019, p. 1115).

Several researchers have taken up the notion that people wish to be perceived as honest and have extended a simple model of lying costs to incorporate what Abeler et al. (2019) termed a “reputation for honesty” - i.e., social image (or social identity) concerns (e.g., Dufwenberg and Dufwenberg, 2018; Khalmetski and Sliwka, 2019). More generally, one can distinguish self-image concerns, when behavior remains unobserved by others; social image concerns, when behavior is observed by others but payouts remain independent; and reputational concerns, when behavior is observed by others and entails interrelated payouts (see Bolton et al., 2021). The main conceptional difference between image concerns and reputational concerns is thus rooted in the payoff independence between decision maker and observer(s). Gneezy et al. (2018) compared lying behavior in an FFH-type experiment between an observed and a non-observed condition. In the observed condition, decisions were made on a computer and could be observed by the experimenter; in the non-observed condition, however, sealed envelopes were used to ensure anonymity. The researchers found more prevalent lying in the non-observed condition than in the observed one, suggesting that social image concerns are an important determinant in lying behavior. In another die-rolling task, Fries et al. (2021) implemented different levels of observability of participants’ die rolls and reports. They found that an increase in the die roll’s observability, in particular, could facilitate honesty, catering to social signaling motivations, as perfectly identifying liars becomes possible. While most of the previous work in this direction has focused on people giving up monetary rewards in order to appear honest, Barron et al. (2021) observed a willingness to lie for a desirable (social) image. Bolton et al.’s (2021) results from a series of dictator game experiments also provide valuable insights for the research questions under investigation in the present study: They reported that observability - and thereby social image concerns - could only have little or even negative effects in certain situations.

Social norms are closely related to social image concerns and present another prominent for (dis)honest behavior. The underlying idea in this line of research is that utility also depends on the perceived distribution of lying costs, in the society or among a particular reference group (e.g., Weibull and Villa, 2005; Gibson et al., 2013). Social norms, more generally, are rules that prescribe appropriate behavior and build upon an individual’s expectations about how others behave (empirical expectations) and about how others believe one ought to behave (normative expectations; Bicchieri, 2016).[4] In this regard, a few studies have experimentally induced norm-nudge stimuli, that is, interventions that aim to change aforementioned expectations (Bicchieri and Dimant, 2019). In a recent study, Dimant et al. (2020), for example, sought to achieve behavioral change by norm-nudging (dis)honesty. They made use of framing effects in presenting participants with information about a majority of participants having been honest or a minority of participants having been dishonest (or the analogue normative information, depending on the treatment) before asking participants to make a decision themselves. Nevertheless, they reported null effects, as the intervention did not achieve a shift in participants' perception of the prevailing social norm. In a closely related study, Serdarevic (2021) introduced a different nudge: Societal expectations were varied by adding information on the expected truthfulness in an FFH-type decision task (participants were informed that they either "have to report truthfully," "do not have to report truthfully," or no additional information was given in a control condition). Serdarevic reported significantly more dishonesty among participants encouraged to misreport, and indeed, she identified a shift in what is seen as socially appropriate to be driving this behavioral change.[5] Likewise, empirical data from Ajzenman (2021) highlights the importance of social norms in (dis)honest behavior: He found an association between an uptake in students' cheating and revelations of corruption among local officials. Several recent studies have also investigated the role of social proximity with respect to social norm compliance: Dimant (2019), for example, found that social proximity amplifies the contagion of anti-social behavior, in particular; and Bicchieri et al. (2022) also identified social proximity as a key ingredient for norm compliance among peers.

While the literature heretofore discussed focuses on how people's individual willingness to lie is shaped by intrinsic lying costs, social image concerns, and social norms, several studies have examined lying behavior in a social context more explicitly by investigating group behavior. In the context of corruption, Weisel and Shalvi (2015) introduced a dyadic game in which two players sequentially play a die-rolling task. By varying the extent to which the two players' payoffs were aligned, they found vast dishonesty with perfectly aligned incentives at almost 50% higher levels than in individual decisions. Moreover, Kocher et al. (2018) observed significantly more lying when decisions are taken as a group of three than as individuals, even without

[4] Some of the literature we are citing uses the expressions empirical and normative expectations when referring to what others call descriptive (how others behave) and injunctive (how one believes one ought to behave) norms. For consistency, in this paper we only refer to empirical expectations and normative expectations when referring to these concepts (also see Bicchieri and Dimant, 2019, for example).

[5] In an experiment on corruption behavior, Köbis et al. (2015) manipulated empirical expectations in a positive or negative way by priming participants with information suggesting that either "almost nobody" or "almost everybody" made a corrupt decision. Similarly, participants in Lois and Wessa (2020) received false feedback about the average level of (dis)honest behavior to induce empirical and normative expectations of the social norm in each of three different treatments. While both of these studies found less ethical behavior in a pro-corruption norm condition and a condition suggesting high cheating among peers, respectively, we regard these methods as (at least borderline) deceptive, as the social information presented to participants is fictitious.

payoff commonality among group members. Importantly, they also identified a an upward shift in group members' expectations about others' lying behavior through communication, which may have additionally facilitated their own dishonesty. These studies highlight that, depending on the particular situational context, people might be even more inclined to lie in the presence of others.

Taken together, previous literature has shown that both social image concerns and social norms can have a positive effect on honesty, whereas group settings, in which people's payoffs depend on one another, might increase cheating. Nevertheless, cheating can be prevalent in group decisions, even when payoffs are not interrelated, such as when common expectations about others deteriorate toward dishonesty. We hypothesize that social image concerns do increase honesty when the norm is to be honest, but that such concerns can have a backfiring effect when there is a norm for dishonesty - social image concerns might lead to stronger conformity to the underlying social norm and thereby encourage honesty *or* dishonesty depending on the situational context. While, in principle, these propositions do not rely on social proximity, we expect them to be stronger with a shorter social distance to the reference group. Hence, we postulate the following three main hypotheses to be tested:

H1 People behave more honestly when there is a social norm for honesty than when there is a social norm for dishonesty.

H2 People's decisions to behave honestly or dishonestly are affected by social image concerns.

H3 Social image concerns and social norms have an interactive effect on honesty. Specifically, social image concerns increase honesty under an honest norm and decrease honesty under a dishonest norm.

5.3 Experimental Design and Implementation

We examine (dis)honest behavior and test our hypotheses using a die-rolling game (Fischbacher and Föllmi-Heusi, 2013), in which behavior has been shown to correlate with various instances of unethical behavior in naturally occurring settings (e.g., Dai et al., 2018, Hanna and Wang, 2017). In this game, participants roll a computerized fair six-sided die, which can be any integer between 1 and 6: $d \in \{$ "⚀", "⚁", "⚂", "⚃", "⚄", "⚅"$\}$. The computerized die roll is implemented by presenting a randomly chosen video recording of an actual die roll on participants' screens. Each participant is then asked to enter their rolled number, making each individual decision fully observable by the experimenter in all treatments (Kocher et al., 2018). Only one of the six possible outcomes yields a bonus payment for participants, such that they have to make a binary decision between lying or not lying. Thus, if participants report having rolled "⚃", they earn a bonus of £1.50 on top of the equally sized reward for participation. If they report having rolled any other number (1, 2, 3, 5, or 6), they earn no bonus. This design choice rules out partial lying (Kajackaite and Gneezy, 2017), which might be affected by observability (Gneezy et al., 2018; Abeler et al., 2019).[6] In comparison to a binary coin toss, this design also increases the proportion of participants with an incentive to lie, as the ex ante probability of missing out on the payoff-maximizing outcome ("⚃") is $\frac{5}{6} = 83.3\%$. At the same time, participants are generally familiar with six-sided dice such that the randomization device and corresponding probabilities are intuitive and easy to understand.

5.3.1 Experimental Treatments and Procedure

We employ a 2 × 2 between-subjects design, in which we vary the observability of a subject's decision and introduce norm nudges to foster an honest or dishonest social norm; see Table 5.1. Each participant is randomly assigned to one of the four treatments.

		Observability	
		Public	Private
Norm nudge	Honest	*honest-public* (T1)	*honest-private* (T2)
	Dishonest	*dishonest-public* (T3)	*dishonest-private* (T4)

Table 5.1: **Treatment overview.**

To operationalize our treatment variations, we apply the following experimental procedure. Before an experimental session begins, each participant is randomly assigned either role A or role B. The experiment then consists of two stages. In the first stage, only A players play the die-rolling game, as described above. Each A player can then be categorized as honest or dishonest: If an A player reports having rolled "⚃" while their die roll yielded a different number, they are categorized as *dishonest*; if they correctly report the number of pips on their die roll, they are categorized as *honest*. If an A player's report is different from "⚃" but also different from the outcome of their die roll, they are excluded from further stages of the study, as their behavior could be seen as dishonest but they did not lie to increase their monetary payoff. Once a sufficient number of A players have made their decisions, we form groups of three consisting of either three honest or three dishonest A players.

[6] Gneezy et al. (2018) and Abeler et al. (2019) found that when participants can lie partially (i.e., report a payoff-enhancing but not the payoff-maximizing outcome), they do so less in observable compared to non-observable conditions; full lying (i.e., reporting the payoff-maximizing outcome), in contrast, is hardly affected by choices being observable to the experimenter. Crede and von Bieberstein (2020) nevertheless found considerably less lying when participants were explicitly made aware of their choices being observable by the experimenter. As our experimental instructions do not explicitly mention that the experimenter is able to track outcomes (in line with Gneezy et al. (2018) and Abeler et al. (2019)), and as partial lying is not possible in our experiment setup, we expect potential effects of experimenter observability to be negligible and constant across treatments.

In the second stage, B players are asked to complete the same die-rolling game. They are informed that they have been matched with three A players to form a team, but the exact matching mechanism is not revealed.[7] B players in all treatments then see their team's die rolls *and* their reports, such that there is no uncertainty about the extent of (dis)honesty in the reference group.[8] This procedure serves as our social norm manipulation, nudging participants' perceptions of the social norm toward honesty or dishonesty. As the experiment is conducted online, each participant naturally has a large spacial and social distance to other participants. To reduce the perceived social distance, they are informed about forming a "team" with other participants, which are each represented by a self-selected avatar (from a set of possible gender-neutral avatar choices; see Abraham et al., 2023, for example).

In the set of *public* treatments, we then seek to induce observability. The outcomes of B players' die rolls and their reported numbers are thus explicitly communicated to their team. In light of the high level of subject anonymity in an online experiment, we introduce an additional feature to strengthen this condition: After B players have made their decisions, the same group of A players is asked to provide each matched B player with feedback about their behavior in the game (without any monetary consequences).[9] Players are informed about this procedure in the experimental instructions. In the set of *private* treatments, in contrast, A players are not informed about B players' behavior and do not provide any feedback, instead moving directly to a concluding survey. Figure 5.1 provides an overview of the experimental procedure.

[7] Presenting a non-representative sample without mentioning the precise sample selection follows the goal of inducing the perception of a social norm for honesty or dishonesty. On a general level, the omission of information is not necessarily regarded as deception in experimental economics (e.g. Hey, 1998; Hertwig and Ortmann, 2008; Wilson, 2016) and a recent survey among student participants and experimental economics researchers showed that presenting a non-representative sample was only regarded as rather deceptive by around 20% of researchers and by even fewer students (mean rating 3.76 on a seven-point Likert scale); the majority of researchers regarded this method as appropriate (mean rating 4.76 on a seven-point Likert scale; see Charness et al., 2022). In fact, our use of a subtle and neutral phrasing (e.g, "You are now matched with three A players.") is in line with the proposed alternatives to presenting a non-representative sample as reported in Charness et al. (2022).

[8] Every group of three A players is separately matched with six B players to increase monetary efficiency because of budgetary constraints.

[9] In particular, the feedback consists of two parts: a numerical rating on a scale from "very negative" (-3) to "very positive" (3) and a verbal response in the form of an adjective describing how participants perceived the B player's behavior. In addition, each A player has the opportunity to revise both their numerical rating and their verbal response after being informed about the feedback from the other two A players in their group. We thereby seek to mimic a simplistic group discussion involving potential gossip concerning B players' behavior to further increase social image concerns (see, for example, Bénabou et al., 2020).

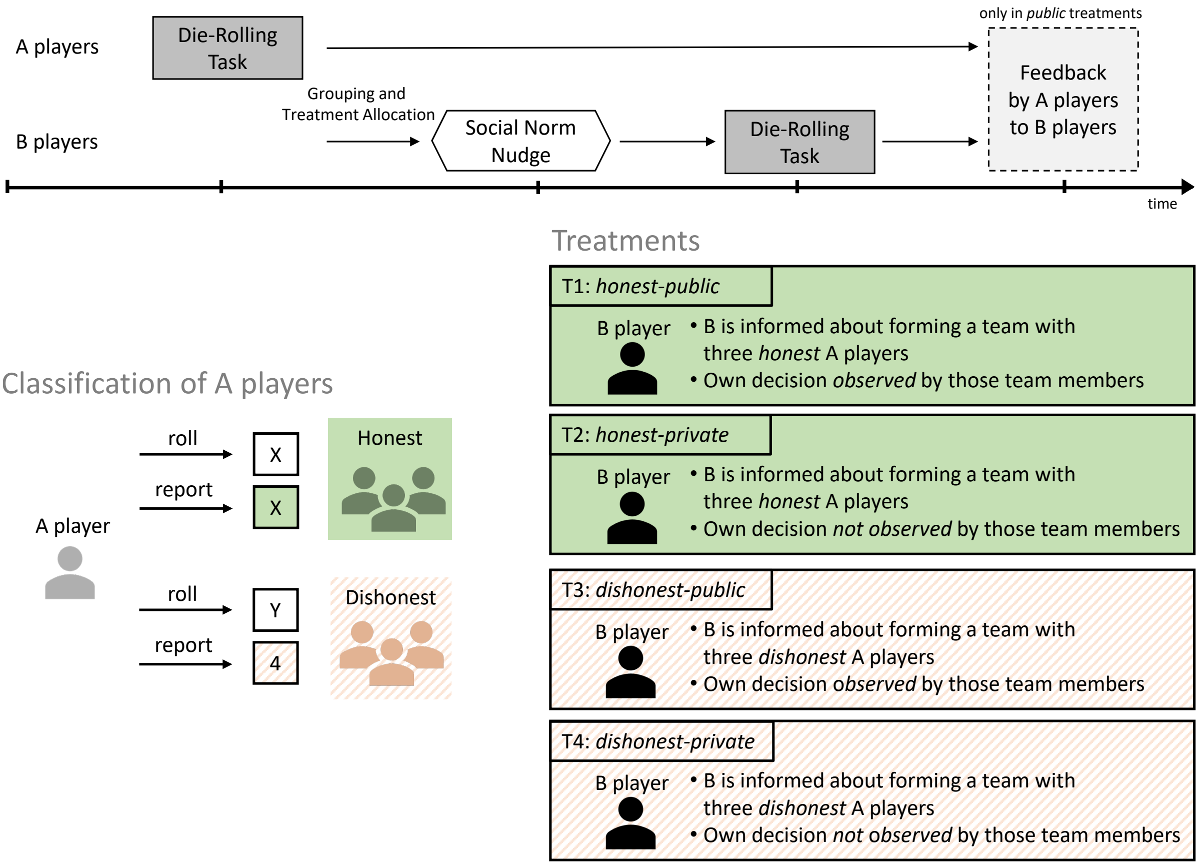

only in *public* treatments
A players
Die-Rolling Task
Feedback by A players to B players
Grouping and Treatment Allocation
B players
Social Norm Nudge
Die-Rolling Task
time
Treatments
Classification of A players
roll
X
report
X
Honest
A player
roll
Y
report
4
Dishonest
T1: *honest-public*
B player
• B is informed about forming a team with three *honest* A players
• Own decision *observed* by those team members
T2: *honest-private*
B player
• B is informed about forming a team with three *honest* A players
• Own decision *not observed* by those team members
T3: *dishonest-public*
B player
• B is informed about forming a team with three *dishonest* A players
• Own decision *observed* by those team members
T4: *dishonest-private*
B player
• B is informed about forming a team with three *dishonest* A players
• Own decision *not observed* by those team members

Figure 5.1: **Overview of the experimental procedure and design.**

This figure schematically outlines the experimental procedure, classification of A players into honest and dishonest players, and the treatment allocation for B players. In the die-rolling task, participants first see a video of a six-sided die roll and are then asked to report the rolled number. Regardless of the actual die roll, they receive a bonus payment if they report the number 4. A players are then classified depending on their report: if they report the number of their die roll, they are classified as honest (with $X \in \{1, 2, 3, 4, 5, 6\}$); if they report 4 instead of the number of their die roll, they are classified as dishonest (with $Y \in X \setminus \{4\} = \{1, 2, 3, 5, 6\}$). In the social-norm-nudge stage, B players are then presented with the die rolls and reports of either three honest A players (T1 and T2) or three dishonest A players (T3 and T4) before going through the die-rolling task themselves. Note that the 'Feedback by A to B players' stage only appears in the two *public* treatments (T1 and T3).

Following B players' decisions in the die-rolling game but before receiving A players' feedback, we elicit their personal norms (Bašić and Verrina, 2023; Bicchieri and Chavez, 2010; Bicchieri and Xiao, 2009) and their normative and empirical expectations (Krupka and Weber, 2013; in that order). Participants are asked how socially appropriate they believe reporting "⚃" to be when a different number was rolled (personal norm) and how socially appropriate they believe most others find this behavior (normative expectation). Both variables are measured on a six-point scale from "very socially inappropriate" to "very socially appropriate." We then elicit empirical expectations by asking, "What percentage of participants do you think reports '⚃' when a different number ... was rolled ... ?" Correct responses concerning normative and empirical expectations, respectively, are rewarded with £0.25. Finally, we measure participants' social and honest image concerns with seven additional survey items.[10]

5.3.2 Experimental Implementation

The experiment was programmed and conducted using oTree (Chen et al., 2016). Participants were recruited via *Prolific* (see, for example, Palan and Schitter, 2018). Participants were paid a participation fee of £1.50 and were able to earn a bonus payment of up to £2.00 depending on their decisions in the experiment, as outlined in the previous section. Participants needed to be at least 18 years old, speak English fluently, have an approval rate on Prolific of at least 90%, and have previously participated in at least five studies.

We recruited a total of 1,629 participants. As we are mainly interested in B players' behavior but required a sufficiently number of A players acting as a reference group and providing feedback to B players in the public treatments, we randomly assigned role A with a 1/3 probability, and role B with a 2/3 probability. This procedure resulted in a sample of 543 A players and 1,086 B players. The 543 A players were then allocated into 181 groups of three players each, either all *honest* or all *dishonest*. Each of these groups was matched with six B players; that is, each group acted as the reference group for six different B players.[11] Participants who did not report any number in the die-rolling game were excluded. Additionally, we applied the following preregistered exclusion criteria: participants who misreported the number that they saw but did not report "⚃", participants who were timed out prior to reporting the die roll,[12] and participants who saw the computerized die roll of "⚃" were excluded from the analysis (since those participants had no incentive to lie).[13] This led to a final sample of 1,192 participants, 409 in the role of A players and 783 in the role of B players.

[10] The survey items on participants' social image concerns are "I was concerned about what others think about me," "it was important to me that my team members would perceive me in a positive way," "it was important to me that my team members would accept me," and "I thought about what information my team members might share about me to another person." The survey items on their honest image are "I wanted others to think I am a person who tells the truth," "I wanted others to think I am a person who does not misrepresent facts," and "I wanted others to think I am a person who does not lie." All items are measured on a seven-point Likert scale from "totally disagree" to "totally agree" (see Wu et al., 2015).

[11] Note that B players, on the other hand, were informed that they were matched with three A players to form a team. With this setup, each group of three A players was separately shown to six B players and acted as their reference group, and each group of three A players in the *public* treatment gave feedback to six B players.

[12] Because the experiment is interactive, all participants in our experiments are asked to provide timely responses such that their respective partners in the experiment do not have excessive wait times. We used pilot data to establish an average participation time. Participants automatically proceeded to the next page when they took too long to respond but thereby forewent any bonus payment.

[13] In addition, we excluded 99 observations from participants who took part in two sessions due to a technical error. In those cases, only data from one's second participation was excluded.

5.4 Results

Overall, 38% of participants were female, the average age was 29 years, they earned a median income of about US$15,000 per year, about half of all participants were students, and the majority was European (22.6% British, 18.2% Portuguese, 11.2% Polish, and 7.5% Italian). The experiment had a median duration of 22 minutes, and participants earned an average of £2.24 (Std. dev: £0.75). Table 5.5 in the Appendix summarizes the distribution of participant demographics across treatments.

In this section, we will first investigate whether our novel treatment variations yield their intended effect. We will thus examine to what extent the induced social norm nudge affects normative and empirical expectations, as well as personal norms (section 5.4.1), and further look into differences in social image concerns between the two observability treatments, public and private (section 5.4.2). After having established the mechanics of our treatment variations, we will go on to analyze the extent to which they affect lying behavior (section 5.4.3). We differentiate between *honest* and *dishonest* on the social-norm-nudge dimension and *private* and *public* on the observability dimension. As outlined in previous sections, combining the two dimensions results in four treatments: *honest-public* (T1), *honest-private* (T2), *dishonest-public* (T3), and *dishonest-private* (T4).

5.4.1 Social Norm Nudge

For the social-norm-nudge dimension, B players are matched with three honest or three dishonest A players in the *honest* and *dishonest* condition, respectively. After observing those three A players' decisions in the die-rolling game, B players conducted this task themselves. The aim of these treatment variations is to change participants' perceptions of social norms - that is, their understanding of what one ought to do and their expectations of what others do. We thus elicited participants' normative and empirical expectations of the social norm using an incentivized coordination task (Krupka and Weber, 2013). Assuming our treatment variations are successful in manipulating social norms, we expect a shift toward dishonesty in the *dishonesty* nudge treatment in comparison to the *honesty* nudge treatment because participants would perceive lying as more socially appropriate and expect others to lie. We also elicited participants' personal norms (Bašić and Verrina, 2023) to enable us to distinguish a shift in societal expectations from a shift in personal convictions.

To elicit normative expectations, we ask participants, “How socially appropriate do you think most others find reporting ‘⚃′ when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in?” on a six-point scale from “very socially inappropriate” to “‘very socially appropriate,” coded from -1 to 1 in equal distances. To elicit empirical expectations, we ask, "What percentage of participants do you think reports '⚃' when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in?" The mean answers within all treatments are shown in Figure 5.2 for normative and empirical expectations. The two graphs paint a similar and convincing picture. First, we observe a significant shift toward dishonesty in the set of *dishonesty* treatments, confirming the effectiveness of our treatment manipulation: in the dishonesty treatments (T3 & T4), lying is perceived to be significantly more socially appropriate (public: $p < 0.001$, private: $p < 0.001$; Mann-Whitney U test) and participants expect a lower percentage of people to report truthfully (public: $p < 0.001$, private: $p < 0.001$; t-test). This is also reflected in an analysis of the feedback from A players to B players about their behavior: Lying in the honest treatments evokes strongly negative feedback, while lying in the dishonest treatments has no (numerical feedback) or less negative (verbal feedback) repercussions (see Figure 5.6 in the Appendix). Another important observation we can see from these results on social norms, however, is that lying is never perceived as socially appropriate. Nevertheless, after a dishonesty nudge, lying on average is perceived as neither socially appropriate nor socially inappropriate, but rather neutral. Finally, it is reassuring that our norm-nudge treatments are indeed effective in shifting normative and empirical expectations, while they remain unaffected by varying observability, as we find no differences between the respective *public* and *private* treatment conditions. Linear regression estimates confirm these initial results (see Table 5.2).[14]

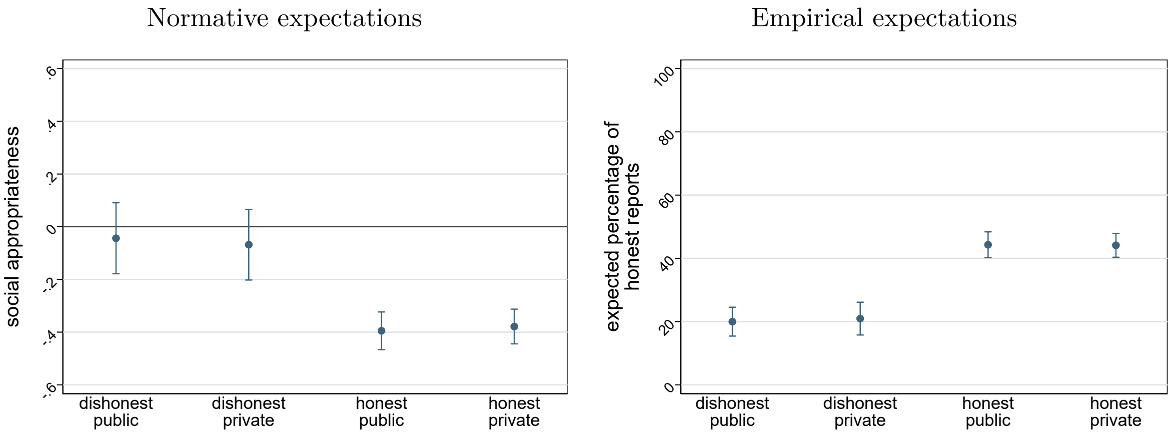


Figure 5.2: **Normative and empirical expectations across treatments.**

This figure presents the mean values and 95% confidence intervals of participants' normative expectations (left panel) and their empirical expectations (right panel) across all four treatments. Normative expectations refers to the social appropriateness of a given action from "very socially inappropriate" (-1) to "very socially appropriate" (1). Empirical expectations refers to the expected percentage of honest reports in a given situation. Note that we elicited the expected percentage of dishonest reports but present the expected percentage of honest reports ($100 - E(\textit{percentage of dishonest reports})$) for comparability.

[14] Observations for normative expectations, empirical expectations, and personal norms differ slightly, as a small number of participants dropped out before completing all survey pages.

	Normative		Empirical		Personal	
	(1)	(2)	(3)	(4)	(5)	(6)
Private	0.006	−0.024	−0.121	−0.960	0.042	0.004
	(0.044)	(0.087)	(2.271)	(4.445)	(0.049)	(0.093)
Honest	−0.332***	−0.351***	−23.777***	−24.312***	−0.103*	−0.127
	(0.050)	(0.069)	(2.580)	(3.547)	(0.055)	(0.078)
Private × Honest		0.041		1.136		0.051
		(0.101)		(5.172)		(0.110)
Constant	−0.057	−0.044	79.645***	80.019***	−0.282***	−0.265***
	(0.048)	(0.058)	(2.428)	(2.966)	(0.052)	(0.067)
Observations	725	725	728	728	704	704
R^2	0.057	0.057	0.106	0.106	0.006	0.006

Table 5.2: **Linear regression: Normative expectations, empirical expectations, and personal norms.**

This table shows the estimated coefficients from linear regressions of normative expectations (1, 2), empirical expectations (3, 4), and personal norms (5, 6) on binary variables indicating the particular treatment. Normative expectations refers to a given action's social appropriateness. Empirical expectations refers to the expected percentage of honest reports in a given situation. Personal norms refers to participants' personal views on a given action's social appropriateness. "Private" takes the value 1 for the set of *private* treatments (T2 and T4) and 0 otherwise; "Honest" takes the value 1 for the set of *honest* treatments (T1 and T2) and 0 otherwise. No controls were included. Robust standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To elicit participants' personal norms, we asked, "How socially appropriate do you find reporting '⚃′ when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in?" on a six-point scale from "very socially inappropriate" to "'very socially appropriate," which was coded from -1 to 1 in equal distances. While the mean values of public and private conditions paired with the dishonesty nudge are almost identical, the respective means in honesty-nudge treatments differ slightly and are both lower than in the dishonesty treatments. Nevertheless, none of these differences are found to be statistically significant after administering non-parametric Mann-Whitney U tests (all $p > 0.100$), whereas linear regression estimates reveal weak evidence for a dishonesty shift with regard to participants' personal norms. As shown in Column (5), the difference between honest and dishonest nudge is significant at the 10% level ($p = 0.059$), pooling private and public, but when disentangling the two, as seen in Column (6), any significance disappears. Comparing these results with the ones for *social* norms, we find almost identical values with regard to appropriateness of lying in the honesty conditions, whereas there is a considerably larger shift toward approval of dishonesty in social norms than in personal norms. This result confirms prior evidence by Bicchieri et al. (2022), in that empirical and normative expectations can be quite malleable with respect to being exposed to peer behavior and observability, whereas personal norms are usually not as easily swayed.

5.4.2 Observability

In our second treatment variation, we manipulated whether B players' decisions are observable by the respective A players to induce social image concerns. To test the effectiveness of this treatment variation, we elicited two sets of survey questions. The first set measures the degree to which participants are concerned about what others think about them ("Social Image Concern"). The second set specifically measures whether they wanted others to think they are honest ("Honest Image"). We then construct two measures of social and honest image concerns by calculating the respective averages over each set of survey questions.

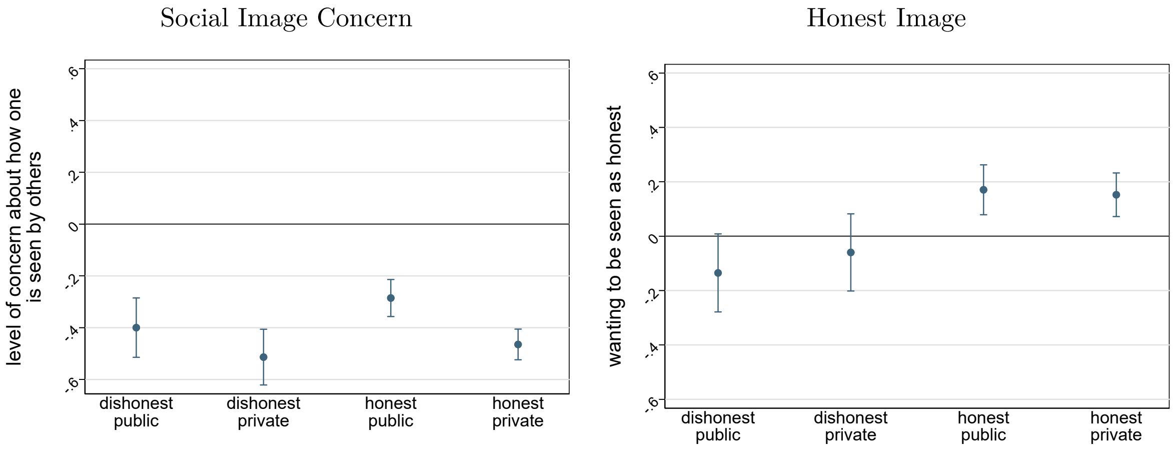


Figure 5.3: **Social image concerns across treatments.**

This figure shows the mean values and 95% confidence intervals of the variables "Social Image Concern" (left panel) and "Honest Image" (right panel) across all four treatments. The variables refer to whether a participant is concerned about how others see them or to whether they want to be viewed as honest in particular and represent the averages of participants' responses to the respective survey items (see footnote 10).

Figure 5.3 shows the mean values for each treatment. All values of the social image concern measure are in the negative domain, suggesting that, overall, participants tend to disagree with the statements that they are concerned about their social image. However, we do observe differences between private and public treatments: In the honesty condition, participants are significantly more concerned about others' opinions in the public treatment than in the private treatment ($p < 0.001$, Mann-Whitney U tests). This difference is not significant in the dishonesty condition ($p = 0.208$, although this lack in significance can also be due to lower sample sizes in the dishonesty conditions see table 5.5 in the Appendix). The linear regression estimate for the private treatment is highly significant (see Table 5.3), while adding the interaction term in Column (2) renders the private coefficient statistically insignificant. However, the combined coefficients of private treatment and interaction term *are* significant ($p < 0.001$, Wald test), indicating that the variation of observability in our setting primarily affects social image concerns in the honesty treatments, but not in the dishonesty treatments.

With regard to our honest image measure, we find that the averages switch from general agreement in the honest treatments to general disagreement in the dishonest treatments, whereas disagreement is strongest in the *dishonest-public* treatment. This indicates that being viewed as honest is more important to participants who were confronted with other *honest* player's decisions beforehand. The differences between honest and dishonest treatments are highly statistically significant (public: $p < 0.001$, private: $p = 0.010$; Mann-Whitney U test). Linear regression estimates confirm these non-parametric results (see Table 5.3).

	Social Image Concern		Honest Image	
	(1)	(2)	(3)	(4)
Private	-0.162^{***}	-0.114	0.006	0.075
	(0.040)	(0.079)	(0.053)	(0.103)
Honest	0.083^{*}	0.114^{*}	0.261^{***}	0.306^{***}
	(0.046)	(0.064)	(0.060)	(0.083)
Private × Honest		-0.066		-0.094
		(0.092)		(0.120)
Constant	-0.378^{***}	-0.400^{***}	-0.104^{*}	-0.135^{*}
	(0.043)	(0.053)	(0.057)	(0.069)
Observations	711	711	711	711
R^2	0.025	0.026	0.026	0.027

Table 5.3: **Linear regression: Social image concerns.**

This table shows the estimated coefficients from linear regressions of the variables "Social Image Concern" (1, 2) and "Honest Image" (3, 4) on binary variables indicating the particular treatment. "Social Image Concern" refers to whether a participant is concerned about how others see them; "Honest Image" refers to whether they want to be viewed as honest. "Private" takes the value 1 for the set of *private* treatments (T2 and T4) and 0 otherwise; "Honest" takes the value 1 for the set of *honest* treatments (T1 and T2) and 0 otherwise. No controls were included. Robust standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

From the analysis of the two survey measures, we conclude that, on average, participants' image concerns are indeed affected by our treatment variations. Through the social image concern measure, we find evidence for the public condition increasing image concerns compared to the private condition. Under the honest condition, however, we observe a stronger desire to appear honest, while participants under the dishonest condition tend to disagree with such a motive. Combining these two findings, we assess social image concerns to be more pronounced when participants' decisions are observed by a reference group of other participants. Additionally, one's desire to appear honest differs distinctly between honest and dishonest treatments.

5.4.3 Honest Behavior

Having established how our treatment variations shift social norms and social image concerns, we go on to examine the extent to which they influence lying behavior. Figure 5.4 provides an initial overview of honesty rates across the four treatments. These honesty rates describe the percentage of subjects who chose to be honest about the outcome of the die roll. Comparing the mean values between *public* and *private* treatments, we find no visible difference in honesty rates for the *dishonest* treatments, and only a marginal difference in the *honest* treatments, for which the rates differ by four percentage points. This difference is not statistically significant ($p = 0.242$, χ^2 test). A noticeable difference that can be seen in Figure 5.4, however, is the difference in honesty rates between *honest* and *dishonest* treatments, holding public and private treatments fixed. Moving from the *dishonest* to *honest* treatment under the *public* condition increases the honesty rate by 19 percentage points ($p < 0.001$). Similarly, under the *private* condition, the mean honesty rate is 15 percentage points higher in the *honest* treatment compared to the *dishonest* treatment ($p = 0.016$).

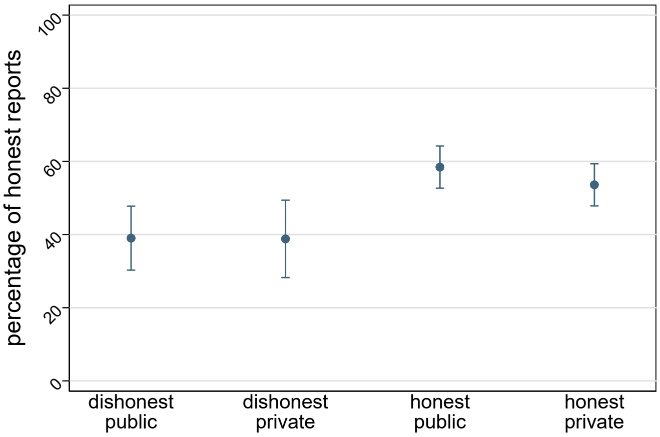


Figure 5.4: **Honesty rates across treatments.**

This figure shows the mean values and 95% confidence intervals of honesty rates across all four treatments. Honesty rates refers to the percentage of honest reports in the die-rolling game in a given treatment.

Table 5.4 shows the estimates from the regression of reporting honestly on binary treatment variables and their respective interaction. We first estimate a linear probability model (LPM) and then check for robustness of our findings using a logistic regression model (Logit).

Prob(honest report)	(1) LPM	(2) Logit	(3) LPM	(4) Logit
Honest	0.174^{***}	0.706^{***}	0.194^{***}	0.788^{***}
	(0.040)	(0.166)	(0.053)	(0.221)
Private	−0.036	−0.149	−0.002	−0.008
	(0.036)	(0.146)	(0.069)	(0.289)
Private × Honest			−0.046	−0.188
			(0.080)	(0.335)
Constant	0.404^{***}	-0.389^{**}	0.390^{***}	-0.446^{**}
	(0.037)	(0.154)	(0.044)	(0.185)
Observations	783	783	783	783
R^2	0.024		0.024	

Table 5.4: **Linear probability models and Logit regressions: Probability of an honest report.**

This table shows the estimated coefficients from linear probability models (LPM, models (1) and (3)) as well as from Logit regressions (models (2) and (4)) on binary variables indicating the particular treatment. The dependent variable takes the value 1 if a participant has reported the true number shown on the computerized die roll and 0 otherwise. “Private” takes the value 1 for the set of *private* treatments (T2 and T4) and 0 otherwise; “Honest” takes the value 1 for the set of *honest* treatments (T1 and T2) and 0 otherwise. No controls were included. Robust standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Column (1) shows the estimates of the LPM, including binary dummy variables for each treatment, where “Honest” takes the value 1 for the set of *honest* treatments and 0 otherwise, and “Private” takes the value 1 for the set of *private* treatments and 0 otherwise. The estimates confirm the observations found in Figure 5.4, as the “Honest” coefficient is significantly positive and suggests an honesty rate that is 17.4 percentage points higher in the set of *honest* treatments compared with the set of *dishonest* treatments ($p < 0.001$). Logit estimates, shown in Column (2), are consistent with LPM estimates. This result further supports H1 (“People behave more honestly when there is a social norm for honesty than when there is a social norm for dishonesty.”). On the other hand, the coefficient for “Private” is not statistically significant in any of our regression models. Thus, we cannot confirm H2 (“People’s decisions to behave honestly or dishonestly are affected by social image concerns.”)

To test H3 (“Social image concerns increase honesty under an honest norm, yet decrease honesty under a dishonest norm.”), we include an interaction term between the binary treatment variables “Honest” and “Private” in our estimation. Again, only the “Honest” coefficient is statistically significant at the 1% level. The LPM suggests the set of *honest* treatments increases the probability of reporting honestly by 19.4 percentage points. Nevertheless, neither the interaction coefficient nor the “Private” coefficient yields statistical significance. In addition to this statistical insignificance, the respective effect size is rather small. Hence, our findings cannot support H3, as we find no statistically significant interaction effect between our social norm and social image concern treatment variations on the probability of reporting honestly.

5.4.4 Robustness: Separating Social Norms from Behavior

Our results demonstrate that presenting social information about others' reporting behavior affects normative and empirical expectations, which in turn influence (dis)honest behavior. At least two questions remain, however. First, the presented information only comprises the decisions of three participants, which raises the question of how representative this behavior is seen as a whole. Second, the normative and empirical expectations presented so far were elicited from the same participants who completed the die roll task themselves. This suggests the possibility that our norm elicitation might thus be affected by the choices of the participants.[15]

To examine the robustness of our results on how the social-norm-nudge treatments affect participants' normative and empirical expectations, we collected additional data on a new group of 301 participants from Prolific, which allows us to identify social norms separately from behavior (Krupka and Weber, 2013; also see Huber and Huber, 2020 for social norms in the context of (dis)honest behavior). In this complementary data collection, we applied the same inclusion criteria as in the main study and used the same incentivized task to elicit normative and empirical expectations. Instead of having completed the die-roll task themselves, however, participants were only presented with a description of a B player's decision situation. We then elicited their normative and empirical expectations next to their personal norms in randomized order.

Analog to our main study, treatments only differed in the social information about the behavior of three other participants in the described situation: In the *dishonest* treatment, we provided information on three dishonestly reporting participants ($n = 100$); in the *honest* treatment, we provided information on three honestly reporting participants ($n = 100$); and, as an add-on, we examined a third treatment without any social information provision (*no info*, $n = 101$).

Figure 5.5 shows an overview of the results of this complementary experiment, in which the red triangles represent the respective mean values for each treatment. For comparison, the blue circles depict the mean values of the two *dishonest* treatments from the main experiment pooled together and the two *honest* treatments from the main experiment pooled together.

We first observe that both normative and empirical expectations are similar between the two experiments. This indicates that the norms elicited in the main experiment are not driven by the fact that behavior and norm assessments were elicited from the same participants; if

[15] Participants might form motivated beliefs (e.g., Bénabou and Tirole, 2016), in the sense of a self-serving belief distortion; that is, strategically expecting many liars in the overall population can reduce the psychological costs of lying and subsequently justify dishonest behavior (Bicchieri et al., 2023).

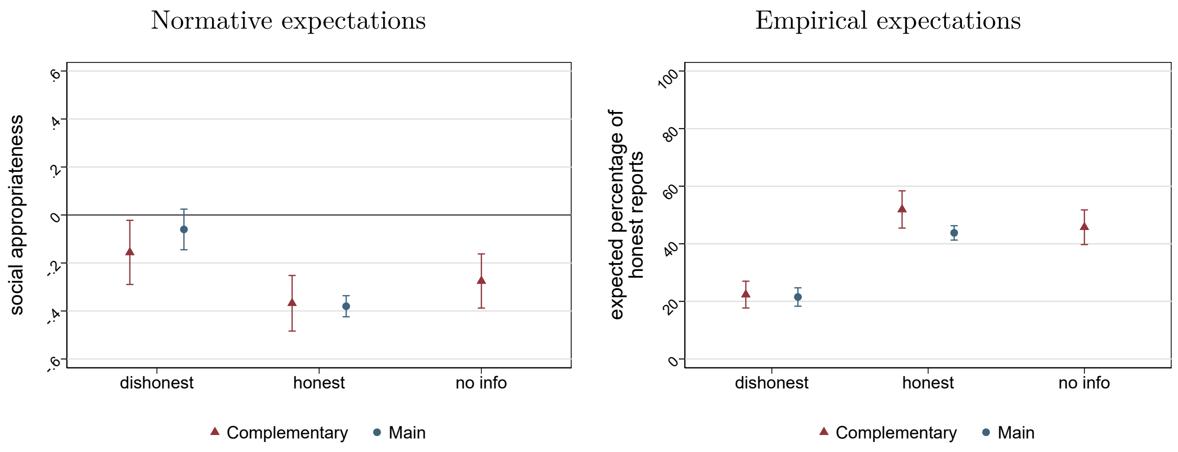


Figure 5.5: **Normative and empirical expectations across treatments and experiments.** This figure presents the mean values and 95% confidence intervals of participants' normative expectations (left panel) and their empirical expectations (right panel) across all three treatments of the complementary experiment (red triangles) and across the two pooled dishonest treatments and two pooled honest treatments of the main experiment (blue circles). See the notes to Figure 5.2 for further details.

anything, it seems as if participants expected even more honesty in the *honest* treatment of the complementary experiment ($p = 0.021$; all other pairwise comparisons between the two experiments yield $p > 0.100$). Also, comparing empirical expectations in the *honest* treatment of the main study with these in the *no info* treatment of the complementary study does not reveal a significant difference.

Next, we test whether the differences between the *honest* and *dishonest* treatments still hold true in the complementary experiment. As we can see from the red triangles and confidence intervals in Figure 5.5, both normative and empirical expectations indeed differ significantly between the *honest* and *dishonest* treatments (normative: $p = 0.027$, empirical: $p < 0.001$). This result also holds when controlling for the order of questions in linear regression analyses (see Table 5.7 in the Appendix). Looking at empirical expectations in particular, we see that, on average, participants expect only 22.3% of other participants to be honest in the *dishonest* treatment, while this percentage increases to 51.9% in the *honest* treatment (also see Bicchieri and Dimant, 2019, who reported norm nudges to primarily affect empirical rather than normative expectations). This considerable shift indicates that the provided social information led to vastly different expectations in terms of others' honesty, thereby suggesting that participants did consider the provided information about others as representative of the overall population.

Finally, the additional *no info* treatment allows for a comparison of our social-norm-nudge treatments to a neutral baseline without social information. Here we can see that lying seems to be regarded as slightly less socially appropriate, and participants expect overall fewer liars in the *honest* treatment. However, these differences are not statistically significant, as the *honest* treatment in our experiment is hardly able to shift social norms (all $p > 0.100$). Being presented with a number of dishonest decisions in the *dishonest* treatment, in contrast, leads to a significant drop in the expected percentage of honest reports ($p < 0.001$).

5.5 Discussion and Conclusion

In this experimental study, we aimed to shed light on the effect(s) of observability and social norms on (dis)honest behavior, as well as on how these two variables interact. In four distinct treatments, we induced either an honest or a dishonest social norm nudge and varied whether participants' decisions were observable and open to judgment by a reference group of other participants. Overall, we found strong evidence for the proposition that people behave more honestly when they have seen other people behave honestly than when they have seen other people behave *dis*honestly. In particular, observing others lie increased the probability of participants lying by almost 20 percentage points compared to observing others reporting truthfully. This effect was driven by a shift in social norms. Our results show that normative and empirical expectations were both significantly affected by the induced norm nudge: In the honest condition, the expectation was that lying would be regarded as less socially appropriate and that this would translate into actual behavior; on average, people expected fewer liars in the honest condition. By contrast, we cannot support the proposition that social image concerns significantly affected lying behavior in our particular setup. A post-experimental survey suggests that whether participants' actions were observable and open to judgment by a reference group affected participants' social image concerns. However, the average answer in all four treatments was to disagree with having social image concerns; only the level of disagreement differed between treatments. Hence, we found no significant difference in honesty rates in the incentivized lying task when varying observability.

These findings underscore the importance of social norms but might question the role of social image concerns in lying behavior. Two important limitations to this conclusion apply, however. First, we conducted the experiment in an one-shot online setup as closed laboratories during the COVID-19 pandemic did not allow for in-person experiments. While such a setting can mimic some forms of online interactions, which have increased substantially since the beginning of the pandemic, naturally occurring interactions are often less anonymous and more personal, even in online meetings (e.g., via video chat). Moreover, the vast number of real-life interactions are repeated in nature, while our anonymous online-setting with globally recruited participants saw the probability of interacting with the same person again converge toward zero. As the presence of actual humans can enhance social image concerns, and thereby increases honesty (Cohn et al., 2022), an anonymous online setting such as the one in the present study might not be able to induce, and thus accurately capture, these concerns. As a second potential limitation of the presented results, on average, participants expected lying to be socially inappropriate, rather than appropriate, even in the condition with a dishonest norm nudge. A potential interaction effect between observability and social norms would more likely occur if the expectation of lying being socially (in)appropriate differed between social norm treatments. Experimentally, an expectation of lying being considered socially appropriate might be established by adding a positive externality, such that a third party benefits from one's dishonest behavior, for example.

Policymakers and managers of organizations often aim to implement policies to increase honesty. We contribute to findings concerning the effects of observability and social norms on people's decision to behave (dis)honestly. We found support that norm nudges work. However, in our online setting, we found no evidence supporting an impact of observability nor an interactive effect of observability and social norms on (dis)honest behavior. These findings enhance our understanding of influential factors in people's decisions to act (dis)honestly and thereby help policymakers and managers to implement effective policies.

Appendix

5.A Additional Tables and Figures

A. Main experiment					
	N	Female	Student	Age	Income*
A player	393	40%	49%	29.78	15
B player					
dishonest-public	128	44%	54%	28.14	15
dishonest-private	88	39%	46%	30.16	15
honest-public	304	36%	53%	27.78	15
honest-private	314	35%	48%	29.96	15
B. Complementary experiment					
no info	101	50%	48%	28.38	–
dishonest	100	42%	45%	29.28	–
honest	100	47%	48%	27.84	–

Table 5.5: **Summary of participant demographics.**

This table summarizes participant demographics across player roles, treatments, and experiments. Player roles and treatments were randomly allocated. The self-reported income is elicited in brackets and we report the respective mid-points such that income bracket 15, for example, means that a participant's yearly income is between USD 10.000 and USD 19.999 per year. Note that N represents the total number of observations in our sample, while for calculating the summary statistics a few participants' demographics are missing in the data set.

A. Main experiment

			B players			
	A players		dishonest		honest	
Personal norm	−0.46	(0.59)	−0.26	(0.66)	−0.33	(0.66)
Normative expectation	−0.28	(0.56)	−0.06	(0.65)	−0.38	(0.57)
Empirical expectation	42.03	(30.92)	21.51	(24.56)	43.78	(32.63)
Social image concern	−0.29	(0.53)	−0.42	(0.56)	−0.36	(0.54)
Honest image	0.34	(0.64)	−0.02	(0.70)	0.19	(0.70)

B. Complementary experiment

	no info		dishonest		honest	
Personal norm	−0.35	(0.61)	−0.30	(0.61)	−0.37	(0.63)
Normative expectation	−0.28	(0.58)	−0.16	(0.68)	−0.37	(0.59)
Empirical expectation	45.74	(30.85)	22.34	(23.78)	52.05	(32.88)

Table 5.6: **Summary of social norms and social image concerns across treatments and experiments.**

This table show the means (and standard deviations) for our measures of personal norms, normative expectations, empirical expectations, social image concerns, and concerns for an honest image as described in the main text (also see Tables 5.2 and 5.3). Panel A shows the values from the main experiment, where column *dishonest* contains the pooled values from treatments dishonest-private and dishonest-public, and column *honest* contains the pooled values from treatments honest-private and honest-public. Panel B shows the respective values from the complementary experiment (in the complementary experiment, social and honest image concerns were not elicited).

	Normative		Empirical		Personal	
	(1)	(2)	(3)	(4)	(5)	(6)
Honest	-0.212^{**}	-0.231^{**}	29.58^{***}	28.87^{***}	-0.0720	-0.0648
	(0.0903)	(0.0897)	(4.069)	(4.061)	(0.0876)	(0.0890)
Constant	-0.156^{**}	-0.259^{*}	22.34^{***}	29.95^{***}	-0.296^{***}	-0.420^{***}
	(0.0682)	(0.137)	(2.378)	(5.512)	(0.0607)	(0.116)
Order control	No	Yes	No	Yes	No	Yes
Observations	200	200	200	200	200	200
R^2	0.027	0.075	0.211	0.247	0.003	0.037

Table 5.7: **Linear regression: Normative expectations, empirical expectations, and personal norms (complementary experiment).**

This table shows the estimated coefficients from linear regressions of normative expectations (1, 2), empirical expectations (3, 4), and personal norms (5, 6) on binary variables indicating the particular treatment while the *dishonest* treatment acts the baseline. The randomized order of the respective questions is included as an independent variable in specifications (2), (4), and (6). See the notes to Table 5.2 for further details. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

	(1)	(2)	(3)
	Normative	Empirical	Personal
Honest	-0.117	6.704	-0.0241
	(0.0808)	(4.470)	(0.0881)
Dishonest	0.120	-22.92^{***}	0.0448
	(0.0898)	(3.919)	(0.0864)
Constant	-0.240^{**}	48.27^{***}	-0.360^{***}
	(0.116)	(4.917)	(0.107)
Order control	Yes	Yes	Yes
Observations	301	301	301
R^2	0.053	0.184	0.018

Table 5.8: **Linear regression: Normative expectations, empirical expectations, and personal norms (complementary experiment).**

This table shows the estimated coefficients from linear regressions of normative expectations (1), empirical expectations (2), and personal norms (3), on binary variables indicating the particular treatment while the *no info* treatment acts as the baseline. The randomized order of the respective questions is included as an independent variable in all specifications. See the notes to Table 5.2 for further details. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The main study was pre-registered on the Open Science Framework (OSF; DOI: 10.17605/OSF.IO/U9HRQ) and includes a pre-analysis plan (PAP). While our procedures and analyses adhere closely to the PAP in general, they deviate occasionally. To increasing transparency, we thus follow Haushofer and Shapiro (2016) and report all of these deviations, and the reasons for them, in Appendix Table 5.9 below.

Pre-analysis plan	Modification	Location
Terminology	We use the term *(social) image concerns* instead of *reputational concerns* to unambiguously distinguish self-image concerns (when behavior remains unobserved by others), social image concerns (when behavior is observed by others but payouts remain independent), and reputation concerns (when behavior is observed by others and entails interrelated payouts) according to Bolton et al. (2021).	—
Sample recruitment	We recruited a total of 1,629 participants instead of 2,000 due to a high number of dropouts who still were eligible for partial payment and a higher number of participants falling under our pre-registered exclusion criteria. With a final sample size of 834 B players we are still able, however, to detect small- to medium-sized main effects between $d = 0.23$ and $d = 0.39$ at a 5% significance level with 80% power in pairwise comparisons.	—
Exclusion criteria	99 observations from participants who took part in two sessions of the experiment due to a technical error were excluded on top of the pre-registrered exclusion criteria. For those participants, only data from their first participation was included.	—
Main analyses	Use of linear probability models (LPM) in addition to the pre-registered Logit models	Table 5.4
	Use of two-sided tests for all hypotheses for robustness (the pre-registration mentions only a one-sided test for Hypothesis 1)	Table 5.4
Exploratory analyses	Personal preferences other than image concerns and perceived social norms have been omitted	Omitted
	Additional analysis: impact of treatment manipulations on normative expectations, empirical expectations, and personal norms	Table 5.2
	Additional analysis: analysis of feedback given from A players to B players; added in the review process	Figure 5.6
Complementary experiment	We conducted an additional, complementary, experiment on a new set of 301 participants, in which we we separate the elicitation of social norms from behavioral decisions; added in the review process.	Section 5.4.4

Table 5.9: **Pre-Analysis Plan Discrepancies.**

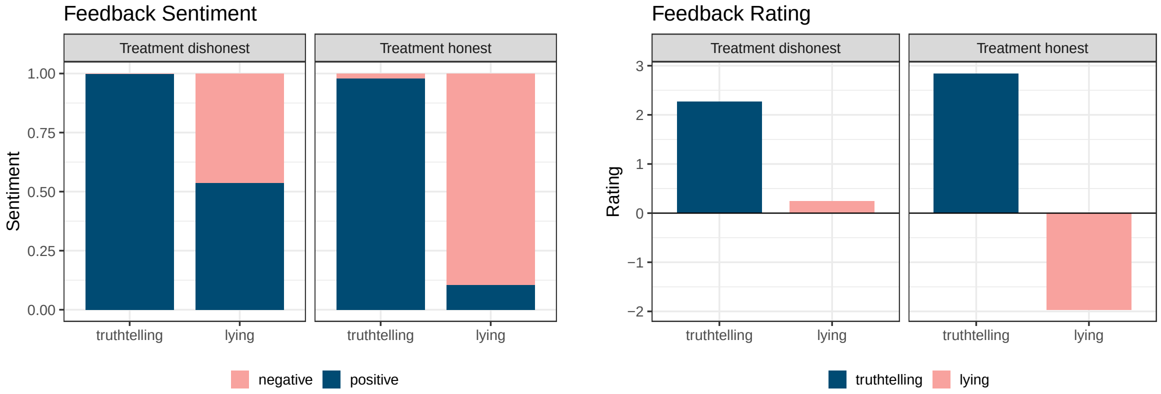
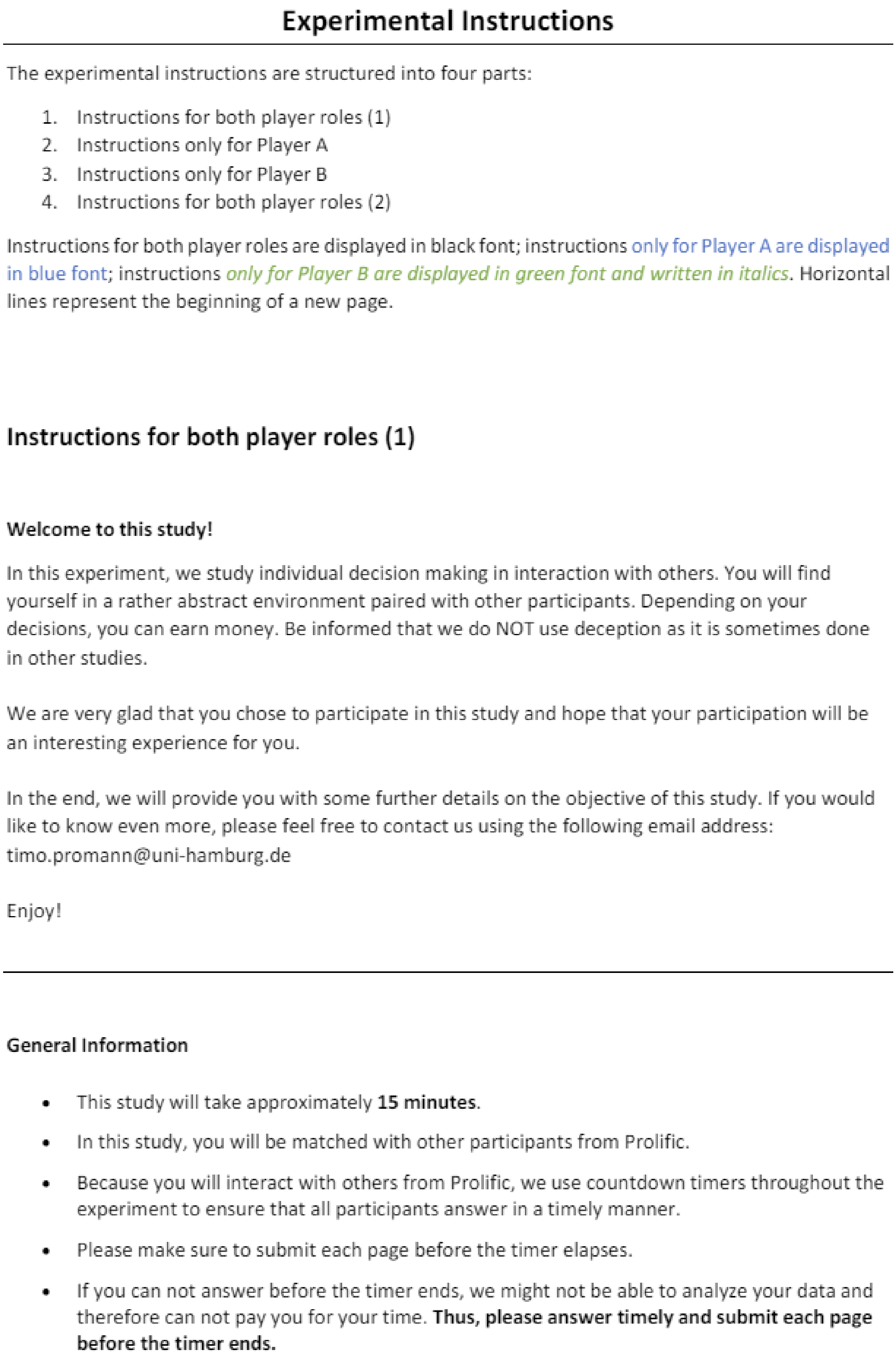


Figure 5.6: **Feedback by A players to B players.**

This figure shows data on the feedback given by A players to B players in the two *public* treatments after having observed their decision in the die roll task. *Left panel:* Sentiment analysis of verbal feedback. The figure shows the percentage of positive and negative feedback separately for B players who decided honestly (*truthtelling* or dishonestly (*lying*) in the two norm-nudge treatments (*dishonest* or *honest*). Feedback words are categorized in a binary fashion, as either positive or negative, based on Hu and Liu (2004). *Right panel:* Mean numerical feedback rating (measured on a scale from "very negative" (−3) to "very positive" (3)) separated by honest and dishonest B players in the two norm-nudge treatments (*dishonest* or *honest*).

5.B Experimental Instructions

The following pages contain the experimental instructions. For screenshots of each page of the experimental software, we refer to the supplementary OSF repository: osf.io/hswz5.



Experimental Instructions

The experimental instructions are structured into four parts:

1. Instructions for both player roles (1)
2. Instructions only for Player A
3. Instructions only for Player B
4. Instructions for both player roles (2)

Instructions for both player roles are displayed in black font; instructions only for Player A are displayed in blue font; instructions *only for Player B are displayed in green font and written in italics*. Horizontal lines represent the beginning of a new page.

Instructions for both player roles (1)

Welcome to this study!

In this experiment, we study individual decision making in interaction with others. You will find yourself in a rather abstract environment paired with other participants. Depending on your decisions, you can earn money. Be informed that we do NOT use deception as it is sometimes done in other studies.

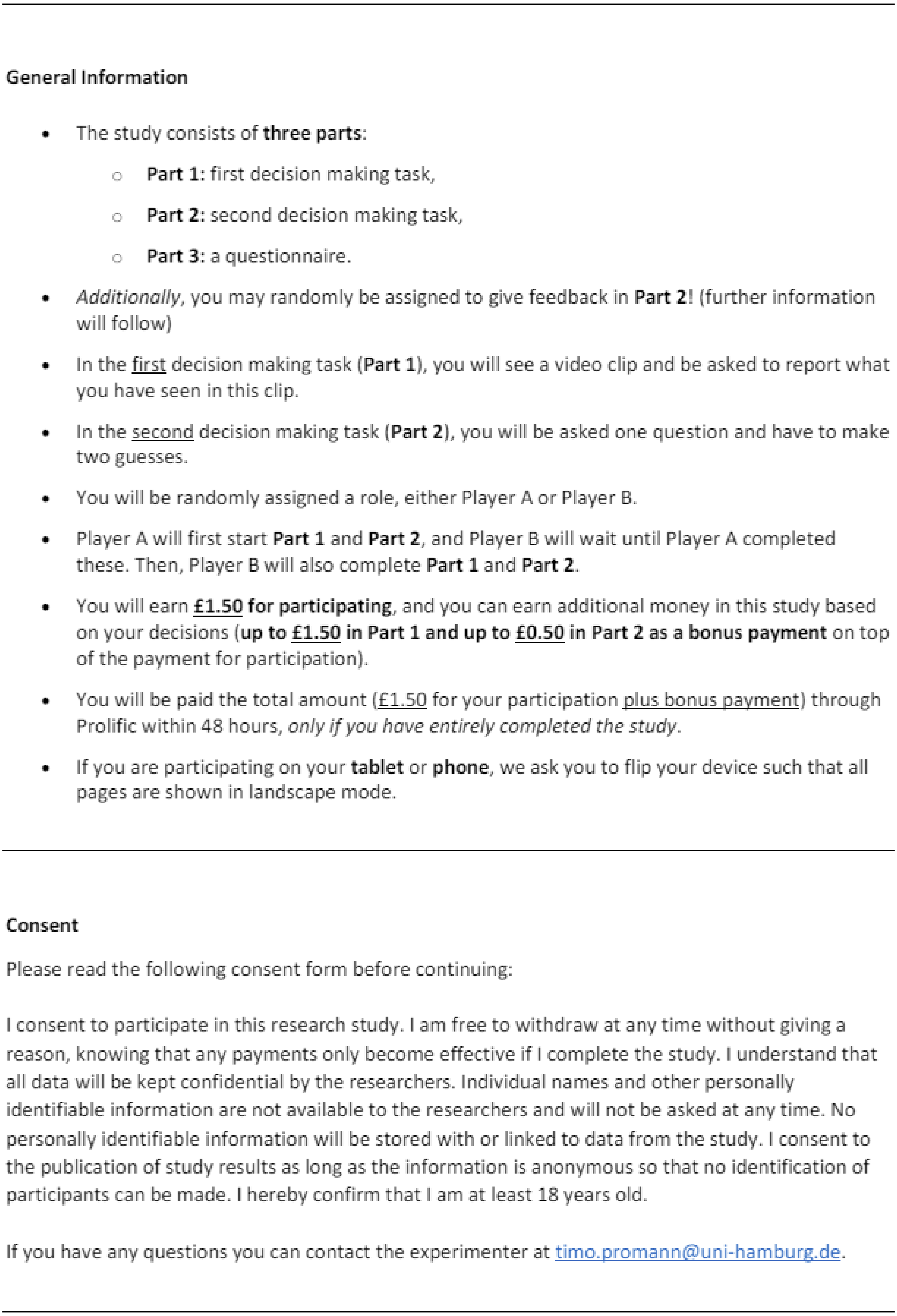
We are very glad that you chose to participate in this study and hope that your participation will be an interesting experience for you.

In the end, we will provide you with some further details on the objective of this study. If you would like to know even more, please feel free to contact us using the following email address: timo.promann@uni-hamburg.de

Enjoy!

General Information

- This study will take approximately **15 minutes**.
- In this study, you will be matched with other participants from Prolific.
- Because you will interact with others from Prolific, we use countdown timers throughout the experiment to ensure that all participants answer in a timely manner.
- Please make sure to submit each page before the timer elapses.
- If you can not answer before the timer ends, we might not be able to analyze your data and therefore can not pay you for your time. **Thus, please answer timely and submit each page before the timer ends.**



General Information

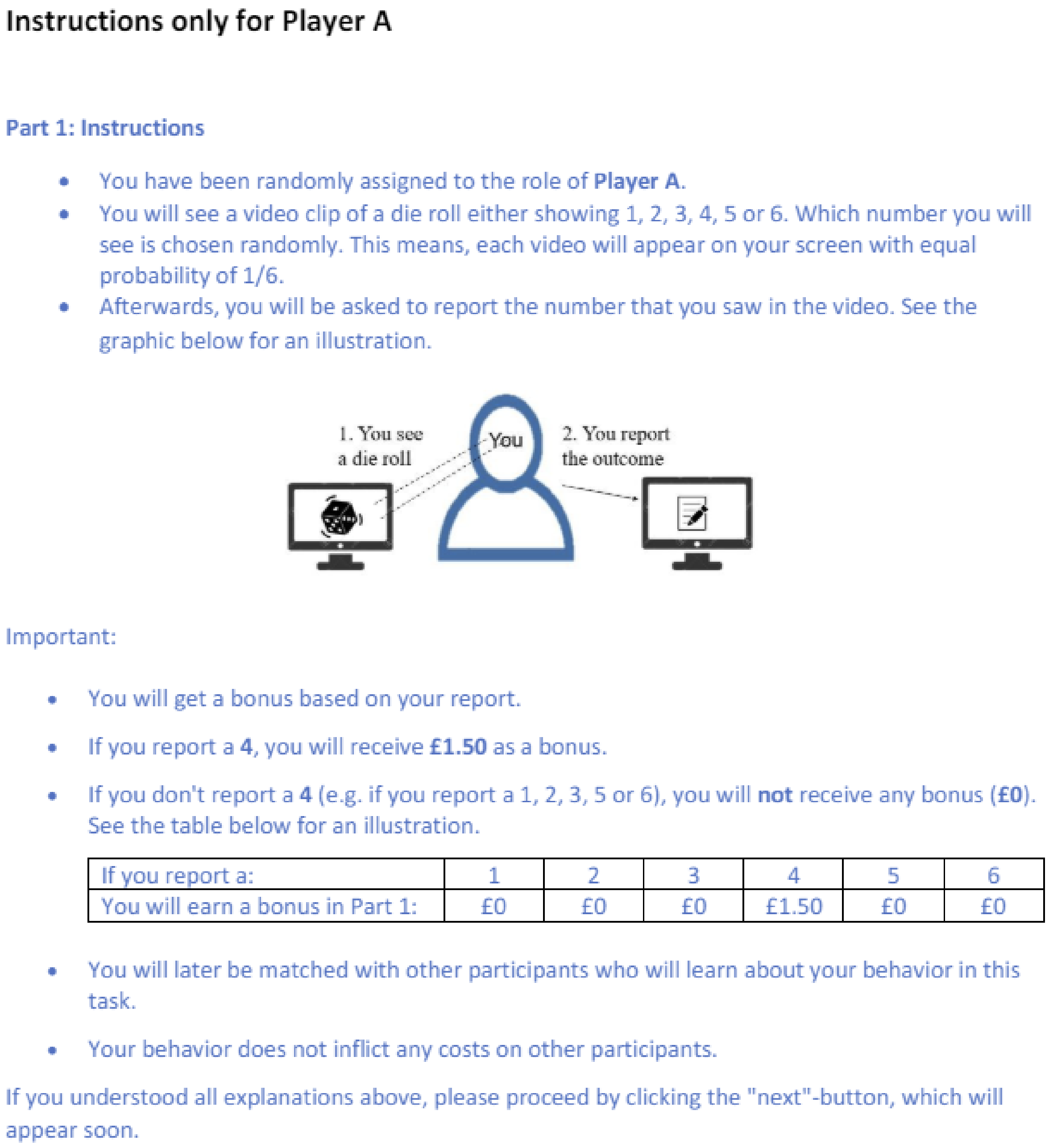
- The study consists of **three parts**:
  - **Part 1:** first decision making task,
  - **Part 2:** second decision making task,
  - **Part 3:** a questionnaire.
- *Additionally*, you may randomly be assigned to give feedback in **Part 2**! (further information will follow)
- In the first decision making task (**Part 1**), you will see a video clip and be asked to report what you have seen in this clip.
- In the second decision making task (**Part 2**), you will be asked one question and have to make two guesses.
- You will be randomly assigned a role, either Player A or Player B.
- Player A will first start **Part 1** and **Part 2**, and Player B will wait until Player A completed these. Then, Player B will also complete **Part 1** and **Part 2**.
- You will earn **£1.50 for participating**, and you can earn additional money in this study based on your decisions (**up to £1.50 in Part 1 and up to £0.50 in Part 2 as a bonus payment** on top of the payment for participation).
- You will be paid the total amount (£1.50 for your participation plus bonus payment) through Prolific within 48 hours, *only if you have entirely completed the study*.
- If you are participating on your **tablet** or **phone**, we ask you to flip your device such that all pages are shown in landscape mode.

Consent

Please read the following consent form before continuing:

I consent to participate in this research study. I am free to withdraw at any time without giving a reason, knowing that any payments only become effective if I complete the study. I understand that all data will be kept confidential by the researchers. Individual names and other personally identifiable information are not available to the researchers and will not be asked at any time. No personally identifiable information will be stored with or linked to data from the study. I consent to the publication of study results as long as the information is anonymous so that no identification of participants can be made. I hereby confirm that I am at least 18 years old.

If you have any questions you can contact the experimenter at timo.promann@uni-hamburg.de.



Instructions only for Player A

Part 1: Instructions

- You have been randomly assigned to the role of **Player A**.
- You will see a video clip of a die roll either showing 1, 2, 3, 4, 5 or 6. Which number you will see is chosen randomly. This means, each video will appear on your screen with equal probability of 1/6.
- Afterwards, you will be asked to report the number that you saw in the video. See the graphic below for an illustration.

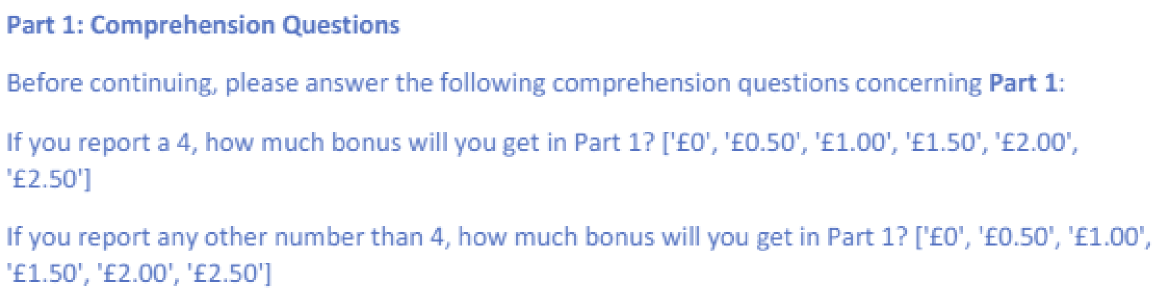
Important:

- You will get a bonus based on your report.
- If you report a **4**, you will receive **£1.50** as a bonus.
- If you don't report a **4** (e.g. if you report a 1, 2, 3, 5 or 6), you will **not** receive any bonus (**£0**). See the table below for an illustration.

If you report a:	1	2	3	4	5	6
You will earn a bonus in Part 1:	£0	£0	£0	£1.50	£0	£0

- You will later be matched with other participants who will learn about your behavior in this task.
- Your behavior does not inflict any costs on other participants.

If you understood all explanations above, please proceed by clicking the "next"-button, which will appear soon.

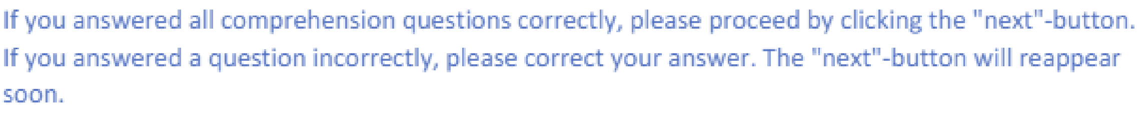


Part 1: Comprehension Questions

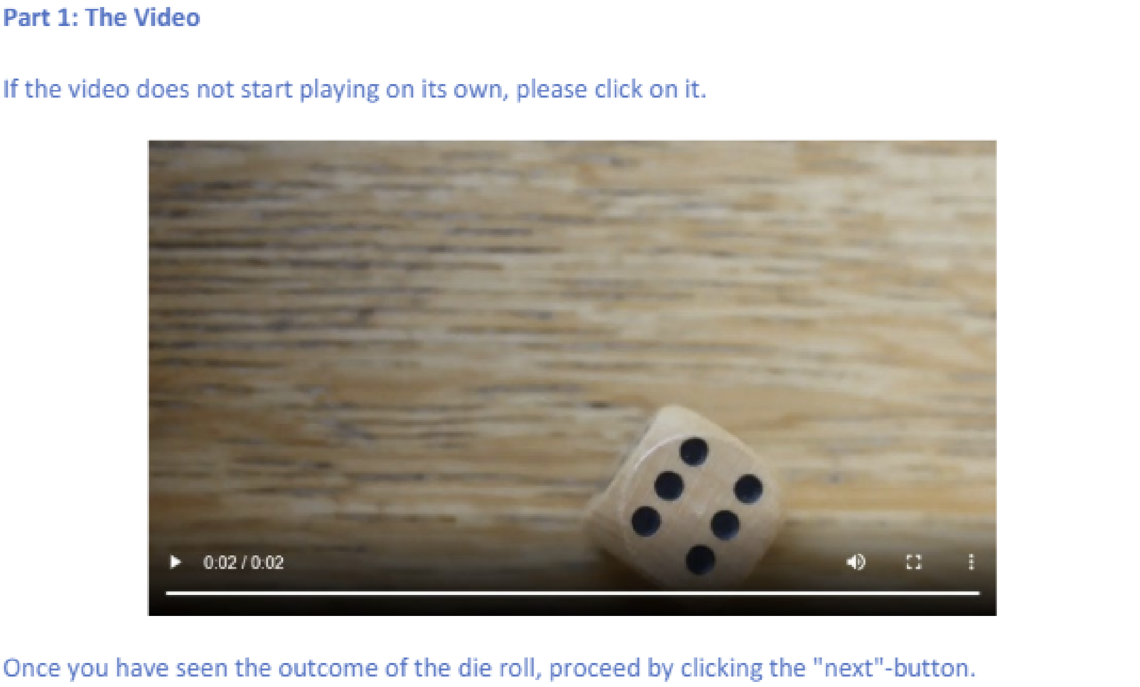
Before continuing, please answer the following comprehension questions concerning **Part 1**:

If you report a 4, how much bonus will you get in Part 1? ['£0', '£0.50', '£1.00', '£1.50', '£2.00', '£2.50']

If you report any other number than 4, how much bonus will you get in Part 1? ['£0', '£0.50', '£1.00', '£1.50', '£2.00', '£2.50']



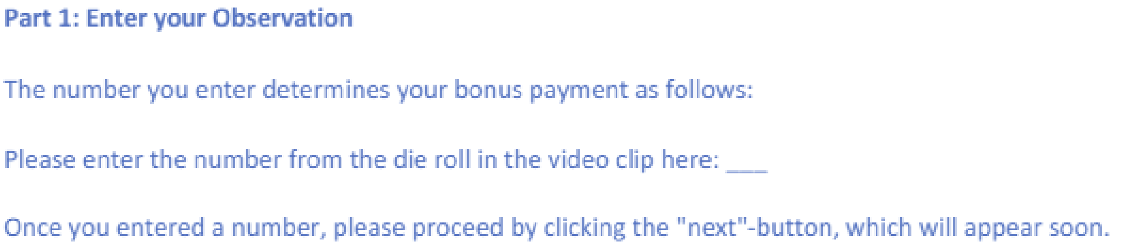
If you answered all comprehension questions correctly, please proceed by clicking the "next"-button. If you answered a question incorrectly, please correct your answer. The "next"-button will reappear soon.



Part 1: The Video

If the video does not start playing on its own, please click on it.

Once you have seen the outcome of the die roll, proceed by clicking the "next"-button.

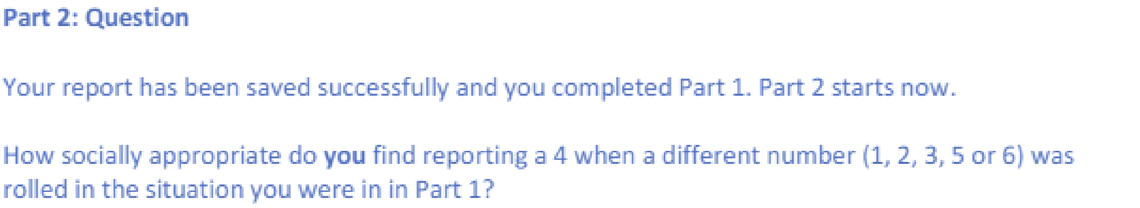


Part 1: Enter your Observation

The number you enter determines your bonus payment as follows:

Please enter the number from the die roll in the video clip here: ___

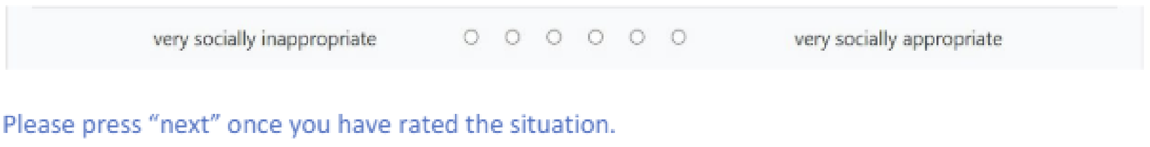
Once you entered a number, please proceed by clicking the "next"-button, which will appear soon.



Part 2: Question

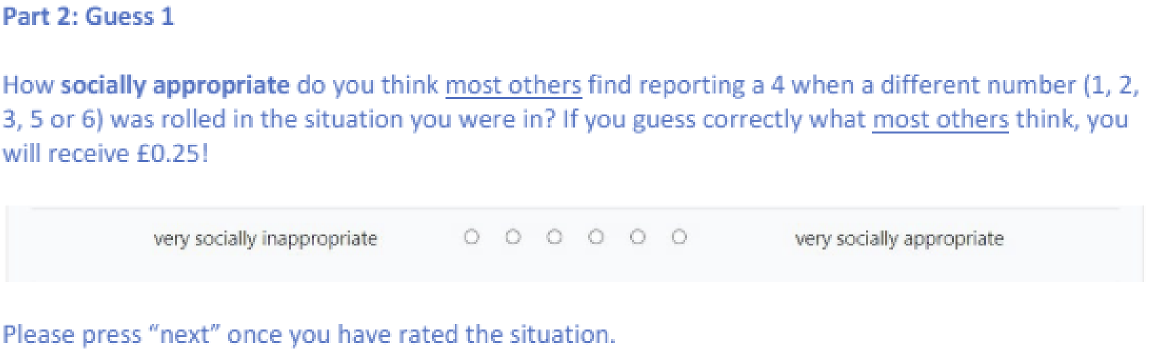
Your report has been saved successfully and you completed Part 1. Part 2 starts now.

How socially appropriate do **you** find reporting a 4 when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in in Part 1?



very socially inappropriate ○ ○ ○ ○ ○ ○ very socially appropriate

Please press "next" once you have rated the situation.

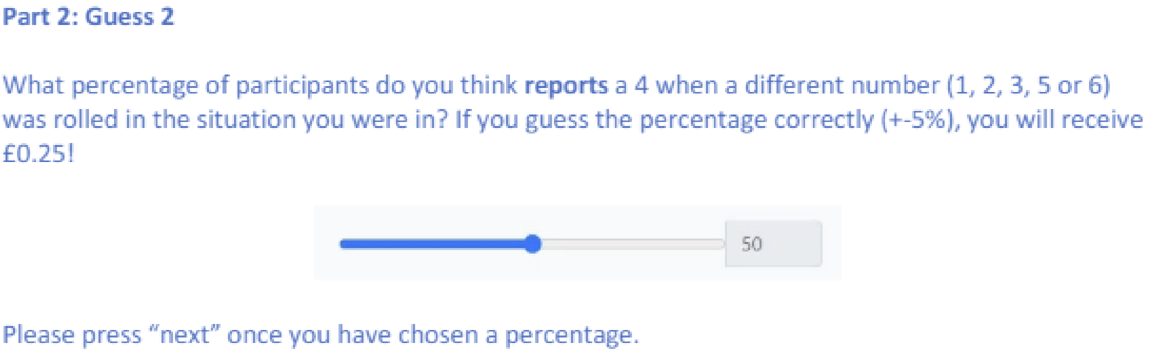


Part 2: Guess 1

How **socially appropriate** do you think most others find reporting a 4 when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in? If you guess correctly what most others think, you will receive £0.25!

very socially inappropriate ○ ○ ○ ○ ○ ○ very socially appropriate

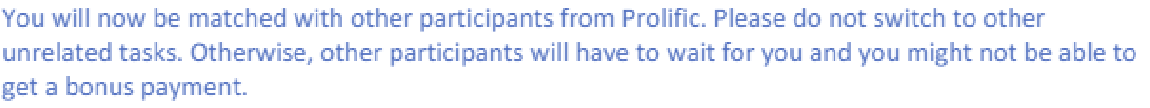
Please press "next" once you have rated the situation.



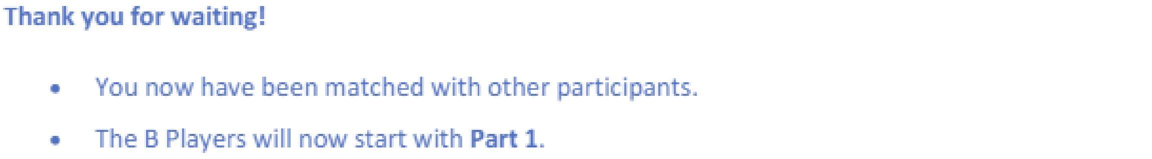
Part 2: Guess 2

What percentage of participants do you think **reports** a 4 when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in? If you guess the percentage correctly (+-5%), you will receive £0.25!

Please press "next" once you have chosen a percentage.

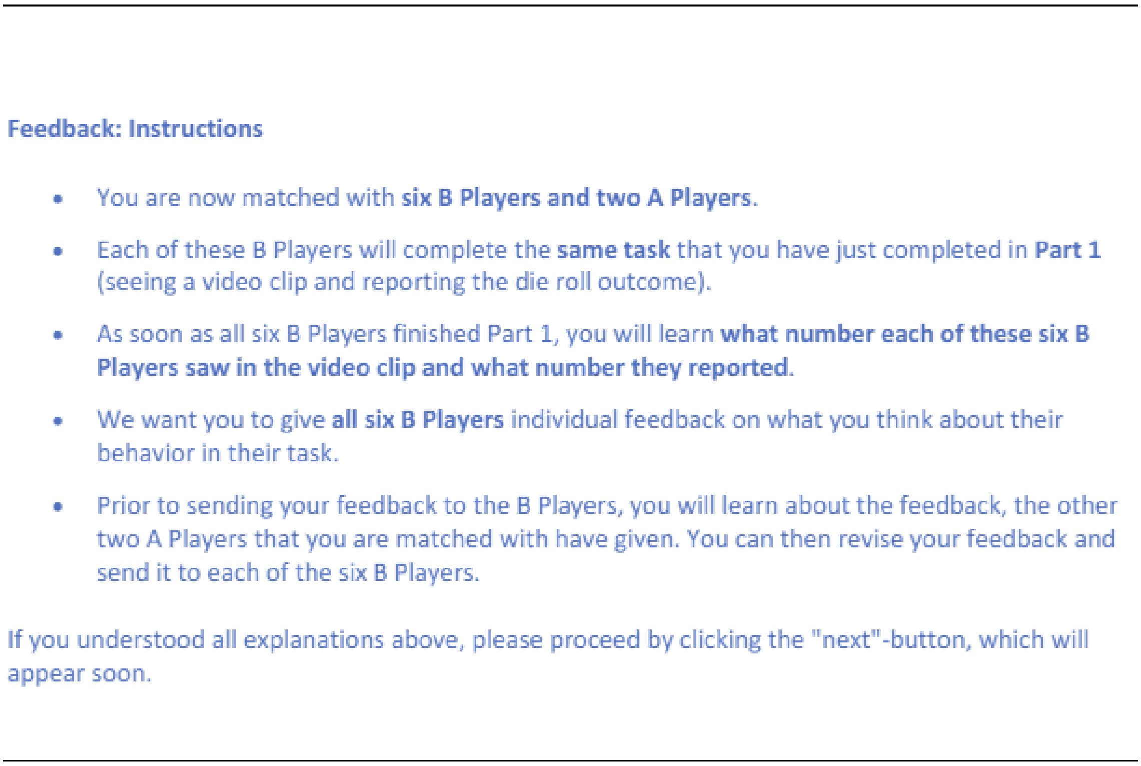


You will now be matched with other participants from Prolific. Please do not switch to other unrelated tasks. Otherwise, other participants will have to wait for you and you might not be able to get a bonus payment.



Thank you for waiting!

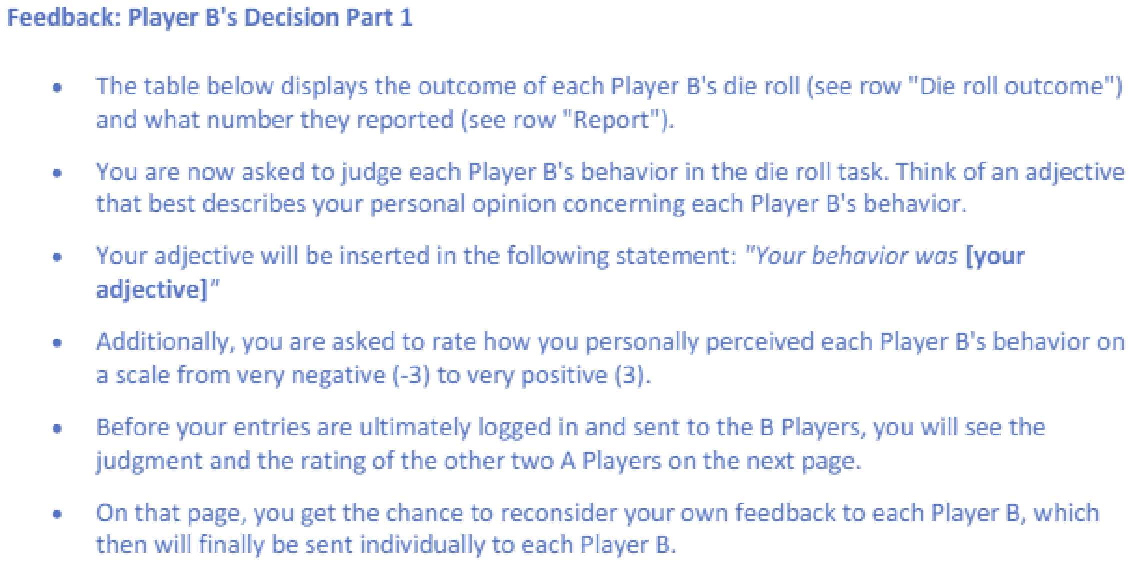
- You now have been matched with other participants.
- The B Players will now start with **Part 1**.



Feedback: Instructions

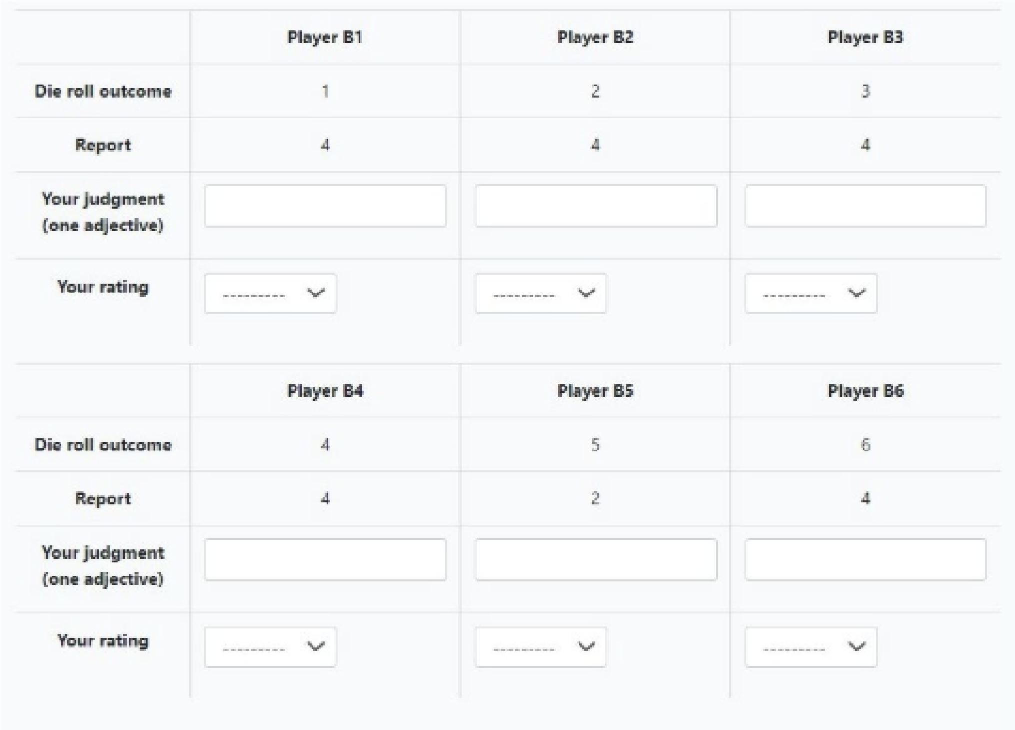
- You are now matched with **six B Players and two A Players**.
- Each of these B Players will complete the **same task** that you have just completed in **Part 1** (seeing a video clip and reporting the die roll outcome).
- As soon as all six B Players finished Part 1, you will learn **what number each of these six B Players saw in the video clip and what number they reported**.
- We want you to give **all six B Players** individual feedback on what you think about their behavior in their task.
- Prior to sending your feedback to the B Players, you will learn about the feedback, the other two A Players that you are matched with have given. You can then revise your feedback and send it to each of the six B Players.

If you understood all explanations above, please proceed by clicking the "next"-button, which will appear soon.



Feedback: Player B's Decision Part 1

- The table below displays the outcome of each Player B's die roll (see row "Die roll outcome") and what number they reported (see row "Report").
- You are now asked to judge each Player B's behavior in the die roll task. Think of an adjective that best describes your personal opinion concerning each Player B's behavior.
- Your adjective will be inserted in the following statement: *"Your behavior was* **[your adjective]***"*
- Additionally, you are asked to rate how you personally perceived each Player B's behavior on a scale from very negative (-3) to very positive (3).
- Before your entries are ultimately logged in and sent to the B Players, you will see the judgment and the rating of the other two A Players on the next page.
- On that page, you get the chance to reconsider your own feedback to each Player B, which then will finally be sent individually to each Player B.

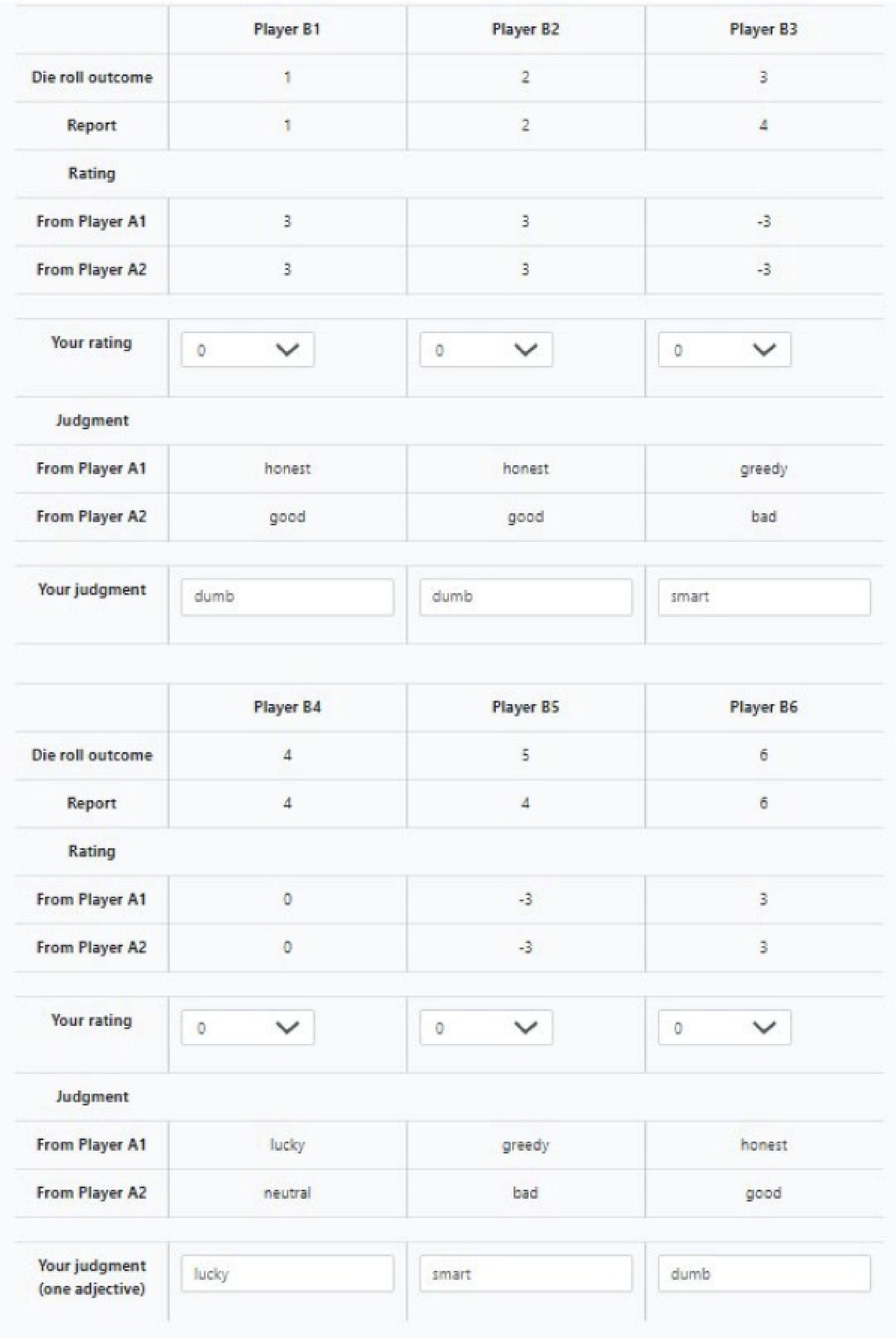


	Player B1	Player B2	Player B3
Die roll outcome	1	2	3
Report	4	4	4
Your judgment (one adjective)			
Your rating	---------	---------	---------

	Player B4	Player B5	Player B6
Die roll outcome	4	5	6
Report	4	2	4
Your judgment (one adjective)			
Your rating	---------	---------	---------

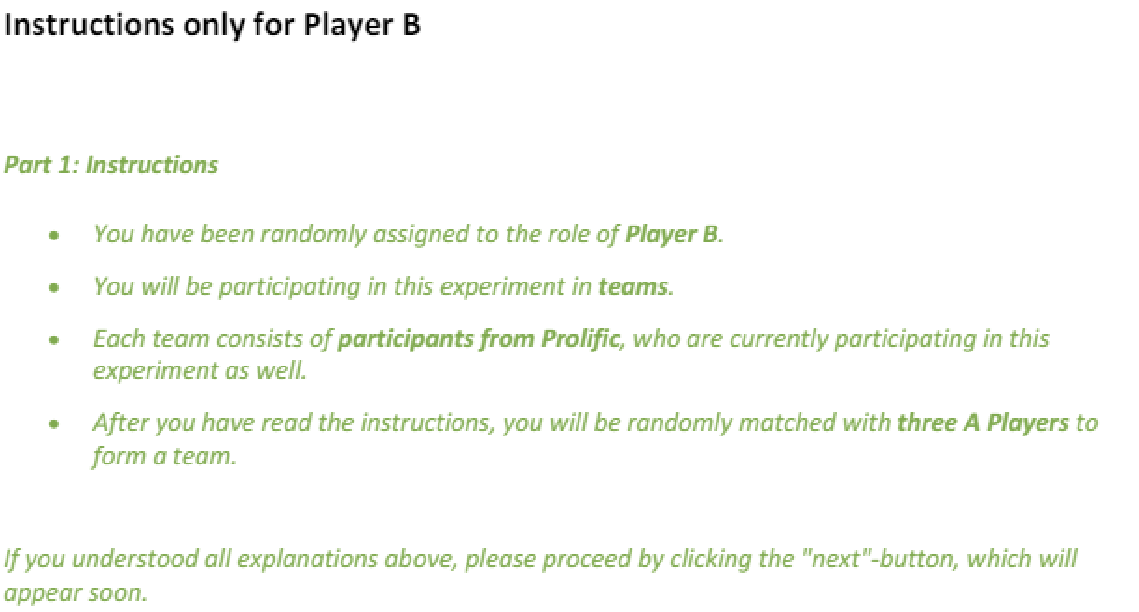
Feedback: Player B's Decision Part 1

- The table below includes the feedback from the other two A Players (termed Player A2 and Player A3).
- In "Rating from Player A2 / A3" you can see their rating of the respective Player B.
- In "Judgment from Player A2 / A3" you can see their adjective concerning the respective Player B.
- You now have the chance to revise your own feedback to each Player B.



	Player B1	Player B2	Player B3
Die roll outcome	1	2	3
Report	1	2	4
Rating			
From Player A1	3	3	-3
From Player A2	3	3	-3
Your rating	0	0	0
Judgment			
From Player A1	honest	honest	greedy
From Player A2	good	good	bad
Your judgment	dumb	dumb	smart

	Player B4	Player B5	Player B6
Die roll outcome	4	5	6
Report	4	4	6
Rating			
From Player A1	0	-3	3
From Player A2	0	-3	3
Your rating	0	0	0
Judgment			
From Player A1	lucky	greedy	honest
From Player A2	neutral	bad	good
Your judgment (one adjective)	lucky	smart	dumb

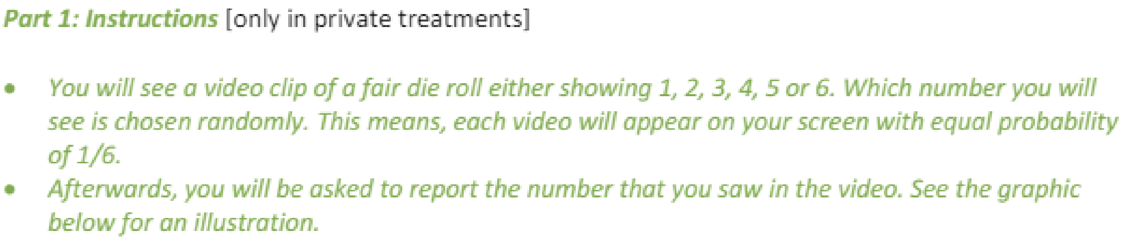


Instructions only for Player B

Part 1: Instructions

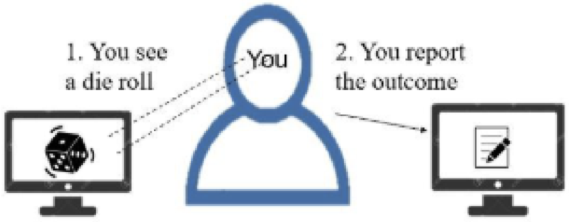
- *You have been randomly assigned to the role of **Player B**.*
- *You will be participating in this experiment in **teams**.*
- *Each team consists of **participants from Prolific**, who are currently participating in this experiment as well.*
- *After you have read the instructions, you will be randomly matched with **three A Players** to form a team.*

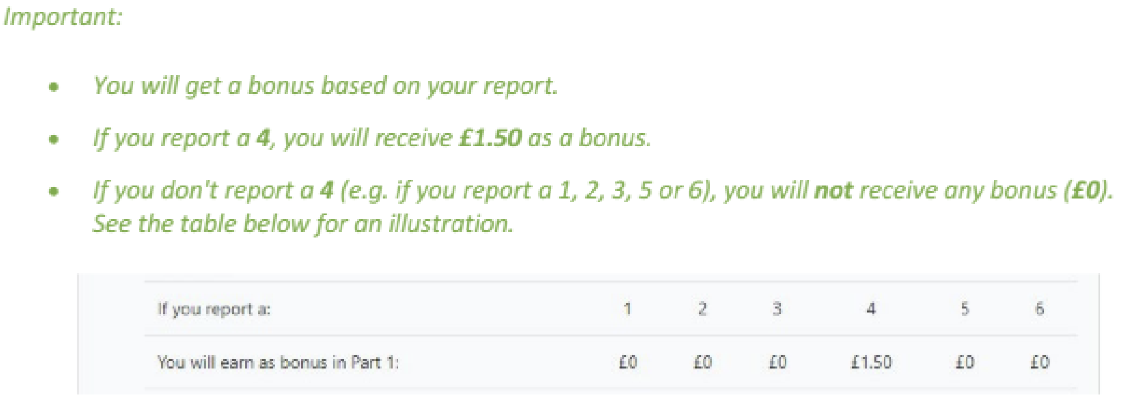
If you understood all explanations above, please proceed by clicking the "next"-button, which will appear soon.



Part 1: Instructions [only in private treatments]

- *You will see a video clip of a fair die roll either showing 1, 2, 3, 4, 5 or 6. Which number you will see is chosen randomly. This means, each video will appear on your screen with equal probability of 1/6.*
- *Afterwards, you will be asked to report the number that you saw in the video. See the graphic below for an illustration.*





Important:

- *You will get a bonus based on your report.*
- *If you report a **4**, you will receive **£1.50** as a bonus.*
- *If you don't report a **4** (e.g. if you report a 1, 2, 3, 5 or 6), you will **not** receive any bonus **(£0)**. See the table below for an illustration.*

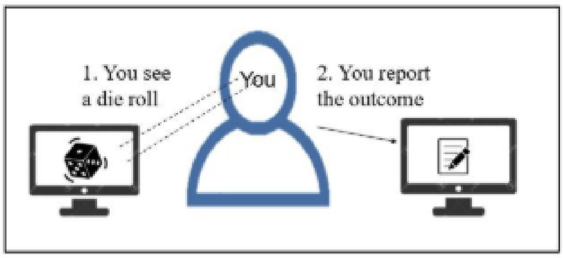
If you report a:	1	2	3	4	5	6
You will earn as bonus in Part 1:	£0	£0	£0	£1.50	£0	£0

- *Your bonus is* ***independent*** *of the bonus of the other participants in your team. This means, your report* ***does not influence*** *the bonus of other team members while your bonus is also* ***not affected by*** *other reports.*
- *Also, your behavior does not inflict any costs on other participants.*
- *Currently, several A Players are completing the task you will be completing later (seeing a video clip and reporting the die roll outcome). You will be matched with three A Players after these instructions.* ***You will learn the outcome of the die roll and what number each Player A in your team reported.***
- *The number that* ***you*** *will see in the video and the number that you will report is* ***private****. This means, nobody in your team will know what you saw and what you reported.*

If you understood all explanations above, please proceed by clicking the "next"-button, which will appear soon.

Part 1: Instructions [only in private treatments]

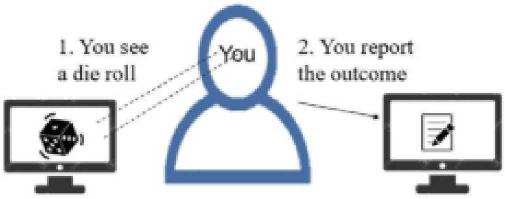
- *The number that you will see in the video and the number that you will report is* ***private****. This means, nobody in your team will know what you saw and what you reported.*
- *After seeing the video clip, you will be asked to report the number that you saw in the video. See the graphic below for an illustration.*



If you understood all explanations above, please proceed by clicking the "next"-button, which will appear soon.

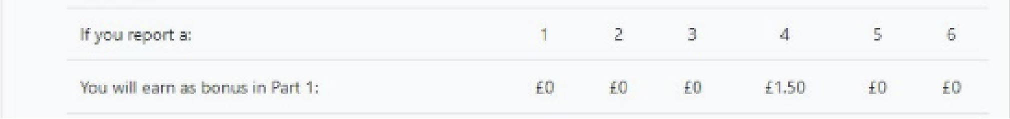
Part 1: Instructions [only in public treatments]

- *You will see a video clip of a fair die roll either showing 1, 2, 3, 4, 5 or 6. Which number you will see is chosen randomly. This means, each video will appear on your screen with equal probability of 1/6.*
- *Afterwards, you will be asked to report the number that you saw in the video. See the graphic below for an illustration.*



Important:

- *You will get a bonus based on your report.*
- *If you report a* ***4****, you will receive* ***£1.50*** *as a bonus.*
- *If you don't report a* ***4*** *(e.g. if you report a 1, 2, 3, 5 or 6), you will* ***not*** *receive any bonus* ***(£0)****. See the table below for an illustration.*



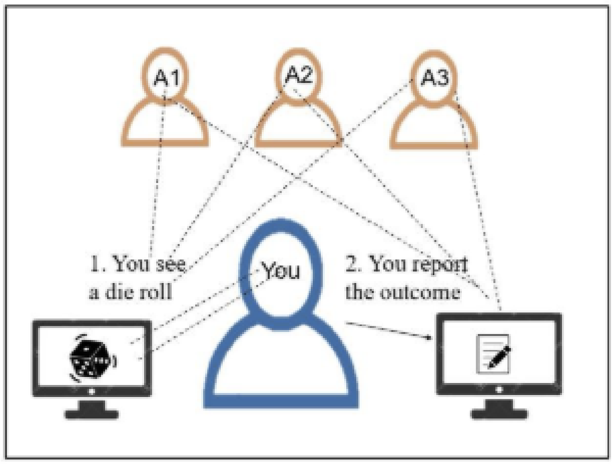
If you report a:	1	2	3	4	5	6
You will earn as bonus in Part 1:	£0	£0	£0	£1.50	£0	£0

- *Your bonus is* ***independent*** *of the bonus of the other participants in your team. This means, your report* ***does not influence*** *the bonus of other team members while your bonus is also* ***not affected by*** *other reports.*
- *Also, your behavior does not inflict any costs on other participants.*
- *Currently, several A Players are completing the task you will be completing later (seeing a video clip and reporting the die roll outcome). You will be matched with three A Players after these instructions.* ***You will learn the outcome of the die roll and what number each Player A in your team reported.***

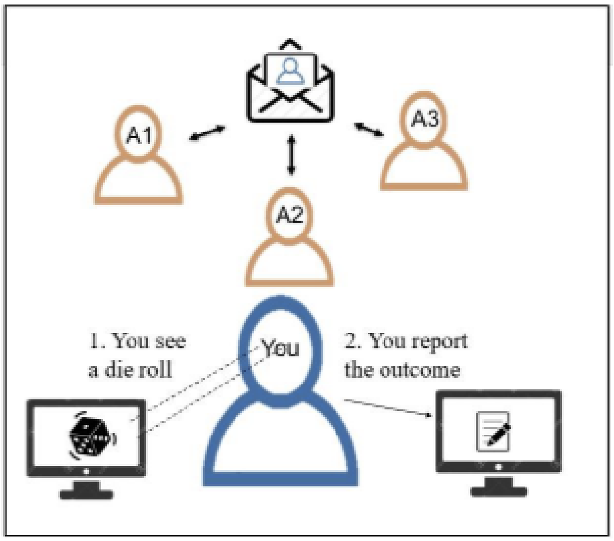
If you understood all explanations above, please proceed by clicking the "next"-button, which will appear soon.

Part 1: Instructions [only in public treatments]

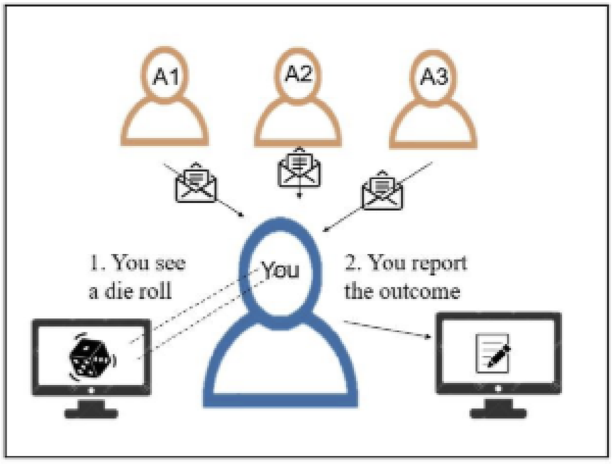
- *The number that you will see in the video and the number that you will report is* ***public****. This means, the A Players in your team will know what you saw and what you reported. See the graphic below for an illustration.*



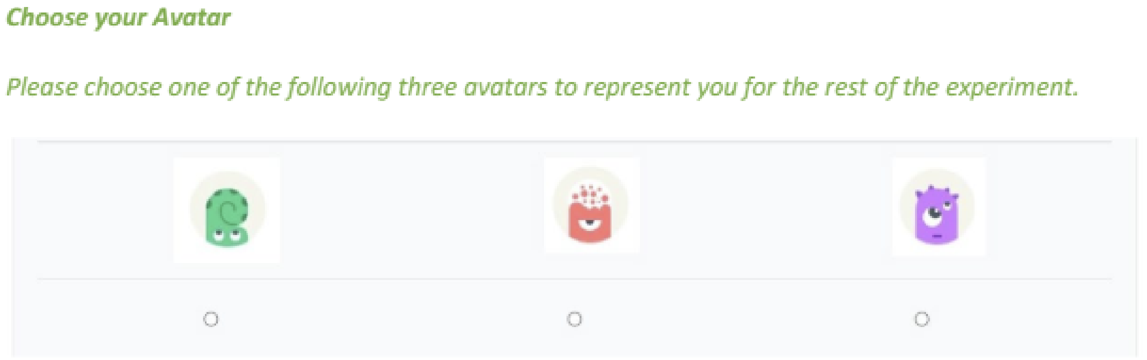
- *The three A Players in your team can then exchange their opinion about your decision. See the graphic below for an illustration.*



- *Lastly, each Player A will send you feedback about your behavior in your task. See the graphic below for an illustration.*



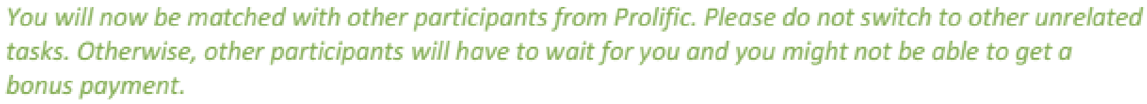
If you understood all explanations above, please proceed by clicking the "next"-button, which will appear soon.



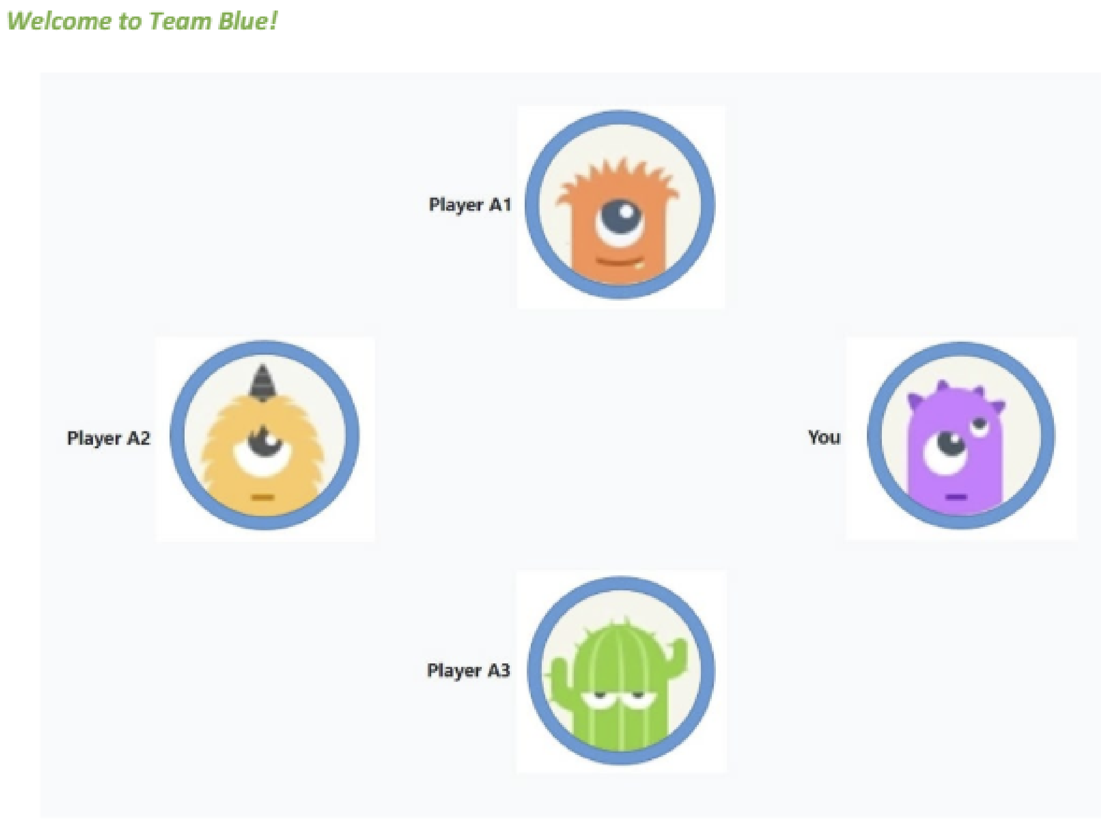
Choose your Avatar

Please choose one of the following three avatars to represent you for the rest of the experiment.

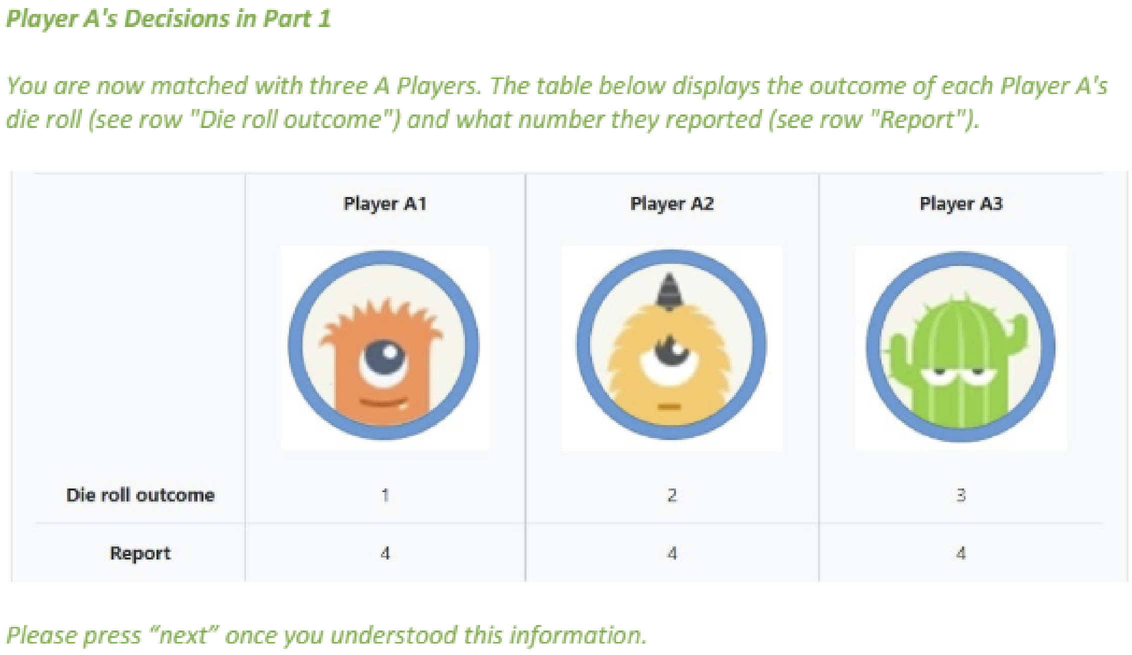
○ ○ ○



You will now be matched with other participants from Prolific. Please do not switch to other unrelated tasks. Otherwise, other participants will have to wait for you and you might not be able to get a bonus payment.



Welcome to Team Blue!

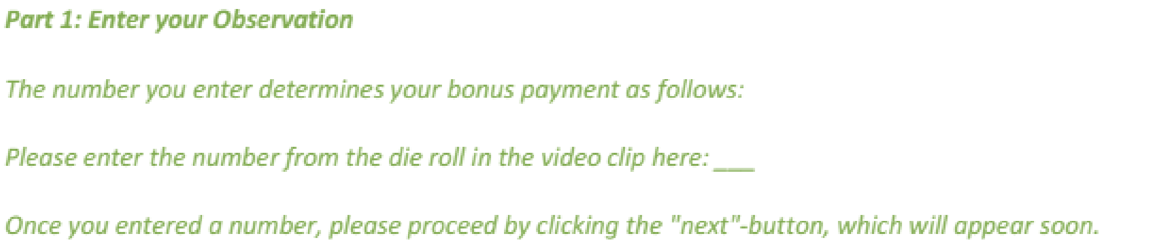


Player A's Decisions in Part 1

You are now matched with three A Players. The table below displays the outcome of each Player A's die roll (see row "Die roll outcome") and what number they reported (see row "Report").

	Player A1	Player A2	Player A3
Die roll outcome	1	2	3
Report	4	4	4

Please press "next" once you understood this information.

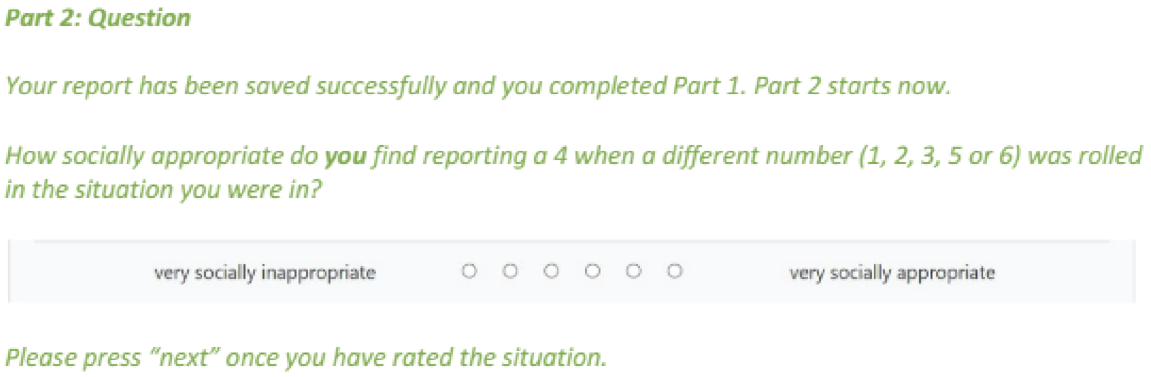


Part 1: Enter your Observation

The number you enter determines your bonus payment as follows:

Please enter the number from the die roll in the video clip here: ___

Once you entered a number, please proceed by clicking the "next"-button, which will appear soon.



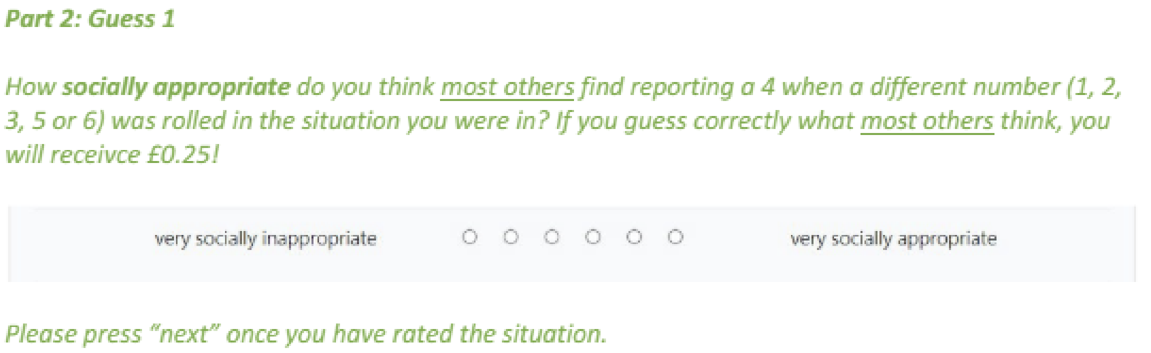
Part 2: Question

Your report has been saved successfully and you completed Part 1. Part 2 starts now.

*How socially appropriate do **you** find reporting a 4 when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in?*

very socially inappropriate ○ ○ ○ ○ ○ ○ very socially appropriate

Please press "next" once you have rated the situation.

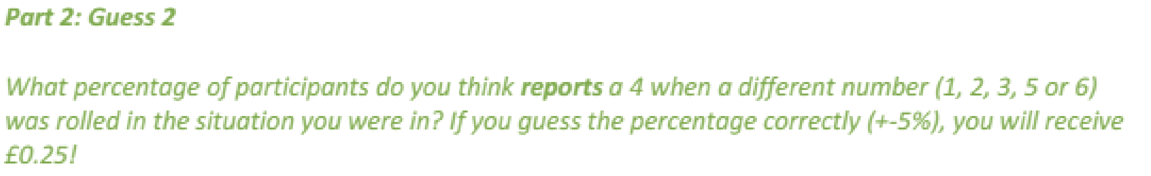


Part 2: Guess 1

*How **socially appropriate** do you think most others find reporting a 4 when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in? If you guess correctly what most others think, you will receivce £0.25!*

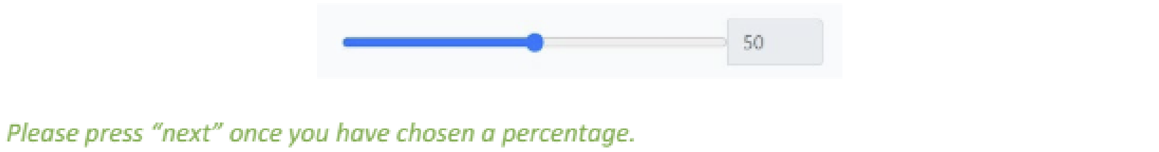
very socially inappropriate ○ ○ ○ ○ ○ ○ very socially appropriate

Please press "next" once you have rated the situation.



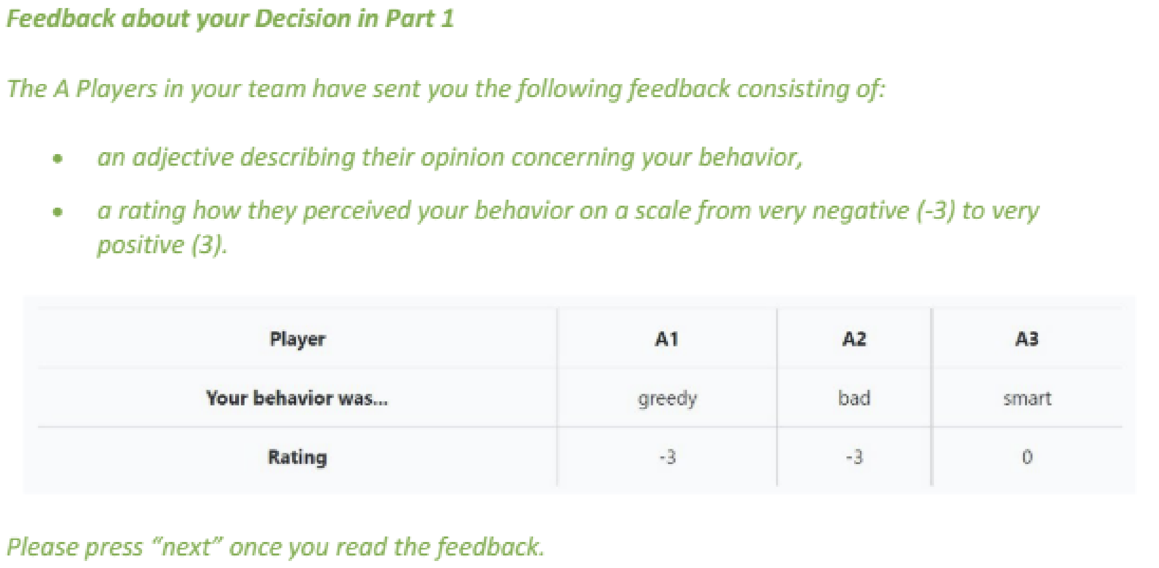
Part 2: Guess 2

*What percentage of participants do you think **reports** a 4 when a different number (1, 2, 3, 5 or 6) was rolled in the situation you were in? If you guess the percentage correctly (+-5%), you will receive £0.25!*



50

Please press "next" once you have chosen a percentage.



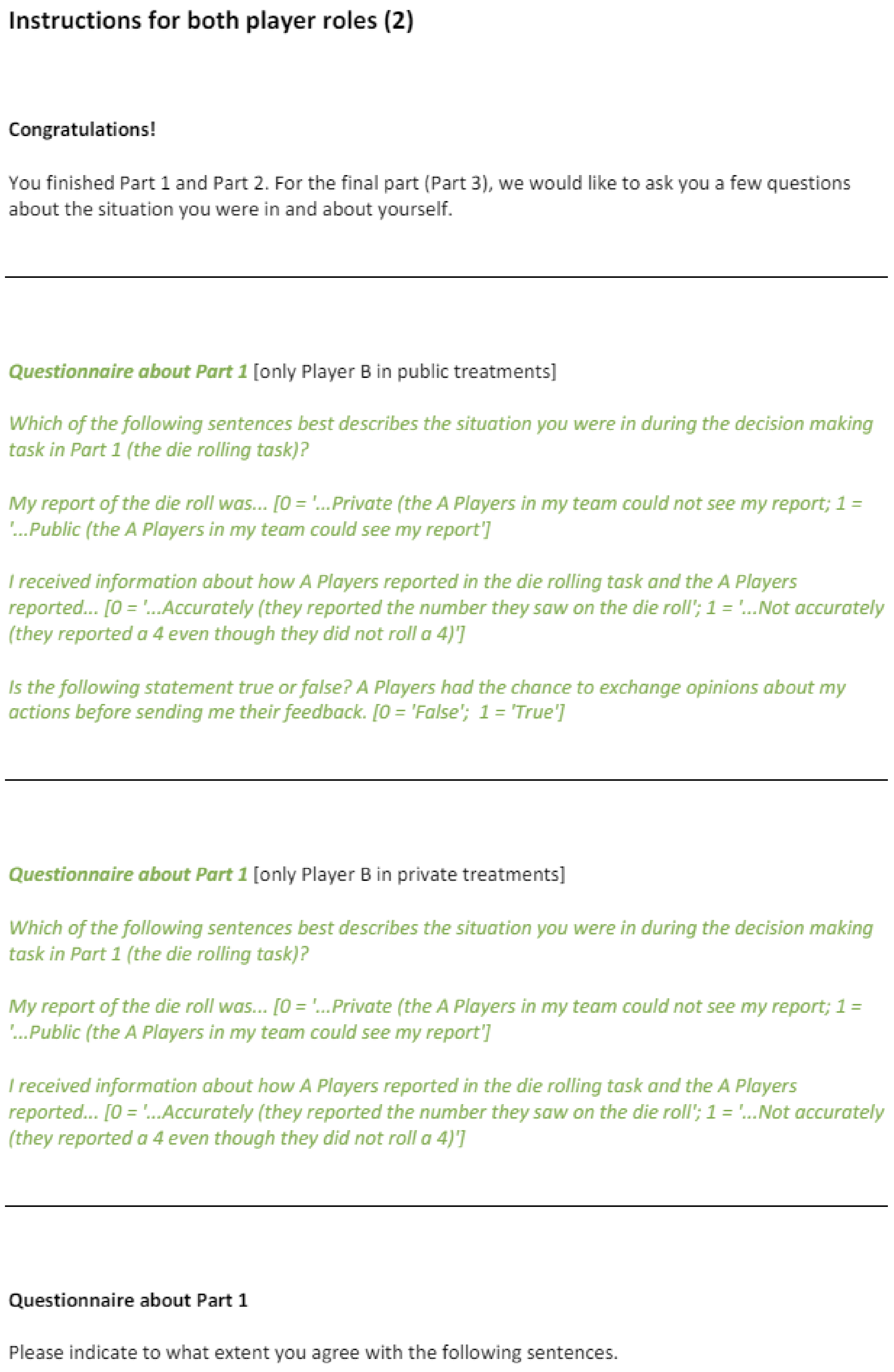
Feedback about your Decision in Part 1

The A Players in your team have sent you the following feedback consisting of:

- *an adjective describing their opinion concerning your behavior,*
- *a rating how they perceived your behavior on a scale from very negative (-3) to very positive (3).*

Player	**A1**	**A2**	**A3**
Your behavior was...	greedy	bad	smart
Rating	-3	-3	0

Please press "next" once you read the feedback.



Instructions for both player roles (2)

Congratulations!

You finished Part 1 and Part 2. For the final part (Part 3), we would like to ask you a few questions about the situation you were in and about yourself.

Questionnaire about Part 1 [only Player B in public treatments]

Which of the following sentences best describes the situation you were in during the decision making task in Part 1 (the die rolling task)?

My report of the die roll was... [0 = '...Private (the A Players in my team could not see my report; 1 = '...Public (the A Players in my team could see my report']

I received information about how A Players reported in the die rolling task and the A Players reported... [0 = '...Accurately (they reported the number they saw on the die roll'; 1 = '...Not accurately (they reported a 4 even though they did not roll a 4)']

Is the following statement true or false? A Players had the chance to exchange opinions about my actions before sending me their feedback. [0 = 'False'; 1 = 'True']

Questionnaire about Part 1 [only Player B in private treatments]

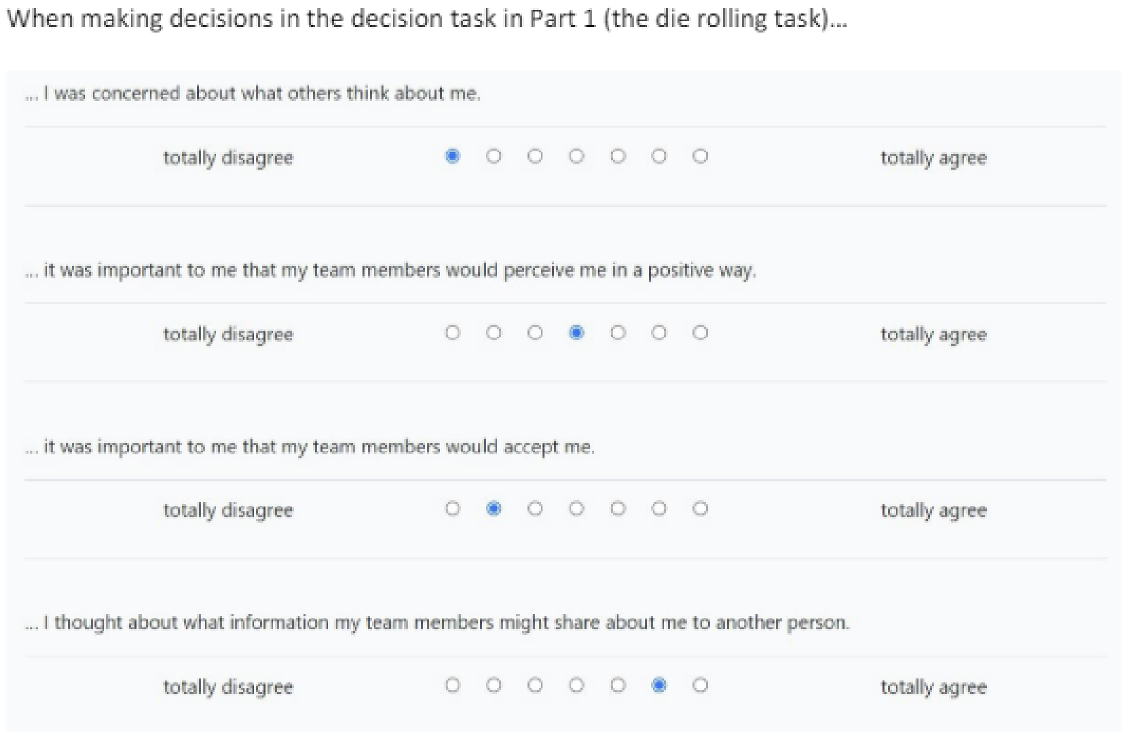
Which of the following sentences best describes the situation you were in during the decision making task in Part 1 (the die rolling task)?

My report of the die roll was... [0 = '...Private (the A Players in my team could not see my report; 1 = '...Public (the A Players in my team could see my report']

I received information about how A Players reported in the die rolling task and the A Players reported... [0 = '...Accurately (they reported the number they saw on the die roll'; 1 = '...Not accurately (they reported a 4 even though they did not roll a 4)']

Questionnaire about Part 1

Please indicate to what extent you agree with the following sentences.



When making decisions in the decision task in Part 1 (the die rolling task)...

... I was concerned about what others think about me.

totally disagree ● ○ ○ ○ ○ ○ ○ totally agree

... it was important to me that my team members would perceive me in a positive way.

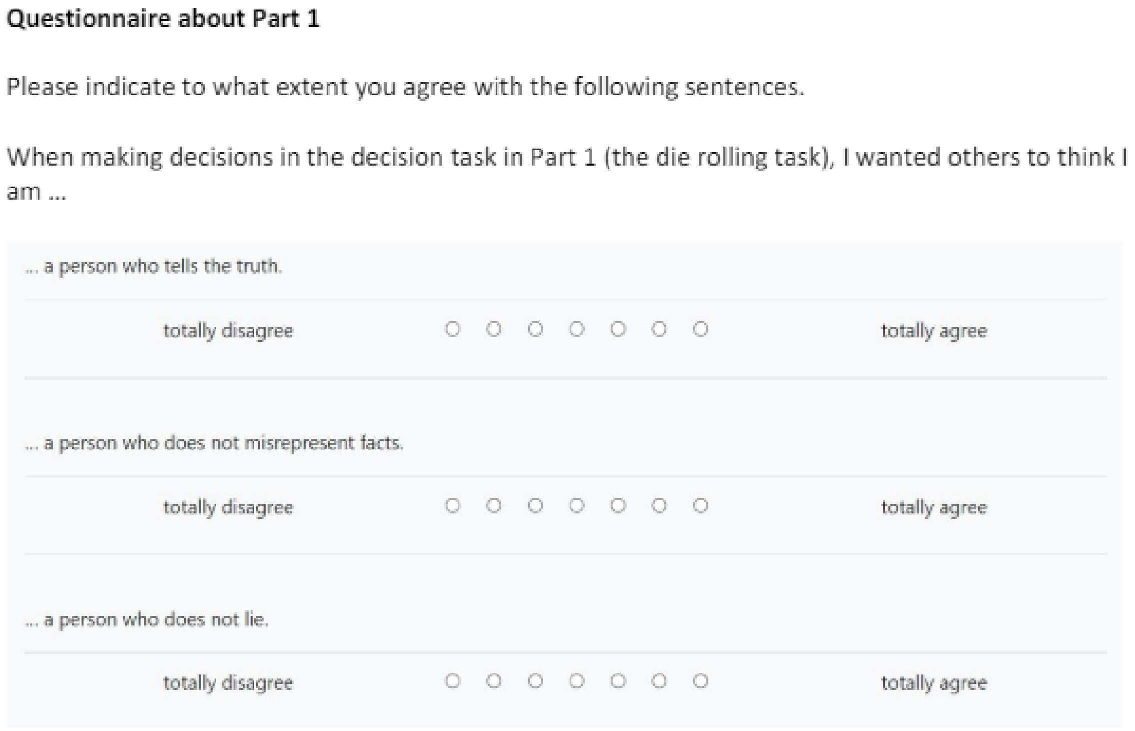
totally disagree ○ ○ ○ ● ○ ○ ○ totally agree

... it was important to me that my team members would accept me.

totally disagree ○ ● ○ ○ ○ ○ ○ totally agree

... I thought about what information my team members might share about me to another person.

totally disagree ○ ○ ○ ○ ○ ● ○ totally agree



Questionnaire about Part 1

Please indicate to what extent you agree with the following sentences.

When making decisions in the decision task in Part 1 (the die rolling task), I wanted others to think I am ...

... a person who tells the truth.

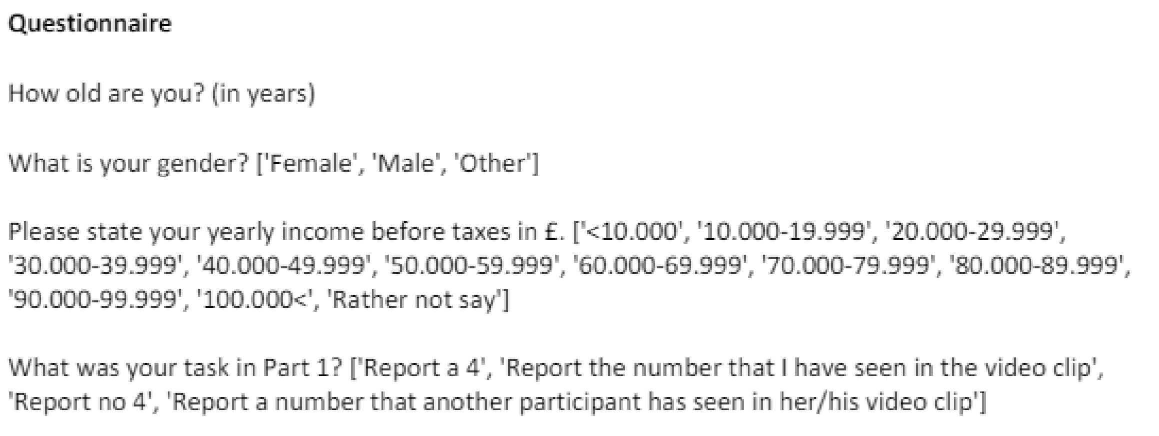
totally disagree ○ ○ ○ ○ ○ ○ ○ totally agree

... a person who does not misrepresent facts.

totally disagree ○ ○ ○ ○ ○ ○ ○ totally agree

... a person who does not lie.

totally disagree ○ ○ ○ ○ ○ ○ ○ totally agree



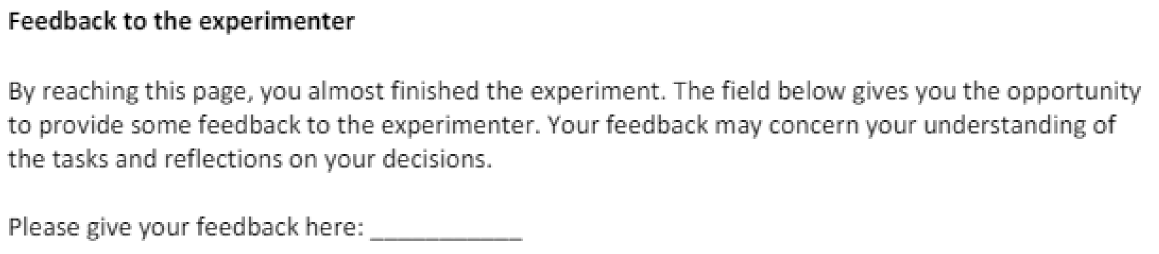
Questionnaire

How old are you? (in years)

What is your gender? ['Female', 'Male', 'Other']

Please state your yearly income before taxes in £. ['<10.000', '10.000-19.999', '20.000-29.999', '30.000-39.999', '40.000-49.999', '50.000-59.999', '60.000-69.999', '70.000-79.999', '80.000-89.999', '90.000-99.999', '100.000<', 'Rather not say']

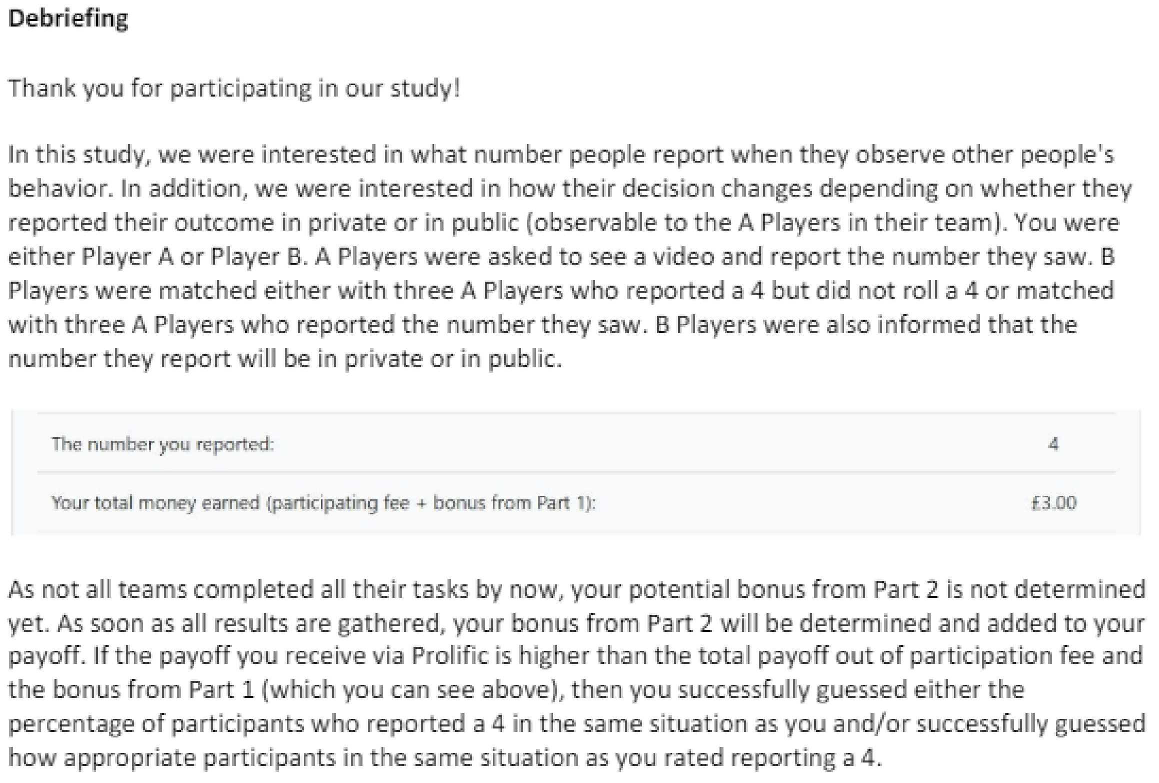
What was your task in Part 1? ['Report a 4', 'Report the number that I have seen in the video clip', 'Report no 4', 'Report a number that another participant has seen in her/his video clip']



Feedback to the experimenter

By reaching this page, you almost finished the experiment. The field below gives you the opportunity to provide some feedback to the experimenter. Your feedback may concern your understanding of the tasks and reflections on your decisions.

Please give your feedback here: ___________



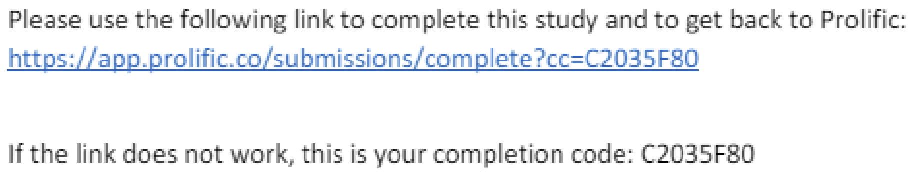
Debriefing

Thank you for participating in our study!

In this study, we were interested in what number people report when they observe other people's behavior. In addition, we were interested in how their decision changes depending on whether they reported their outcome in private or in public (observable to the A Players in their team). You were either Player A or Player B. A Players were asked to see a video and report the number they saw. B Players were matched either with three A Players who reported a 4 but did not roll a 4 or matched with three A Players who reported the number they saw. B Players were also informed that the number they report will be in private or in public.

The number you reported:	4
Your total money earned (participating fee + bonus from Part 1):	£3.00

As not all teams completed all their tasks by now, your potential bonus from Part 2 is not determined yet. As soon as all results are gathered, your bonus from Part 2 will be determined and added to your payoff. If the payoff you receive via Prolific is higher than the total payoff out of participation fee and the bonus from Part 1 (which you can see above), then you successfully guessed either the percentage of participants who reported a 4 in the same situation as you and/or successfully guessed how appropriate participants in the same situation as you rated reporting a 4.



Please use the following link to complete this study and to get back to Prolific:
https://app.prolific.co/submissions/complete?cc=C2035F80

If the link does not work, this is your completion code: C2035F80

List of Figures

2.1 **Answer Distribution for Example Question.** 13

2.2 **Treatment Overview.** 15

2.3 **Flow Diagram.** 16

2.4 **Sum of points for all treatments.** 19

2.5 **Sum of points for all team treatments split up between WiSo-Lab and Prolific.** 31

2.6 **Sum of points for all treatments with Prolific pre-study data set.** 32

2.7 **Sum of points for all team treatments split up between WiSo-Lab and Prolific with Prolific pre-study data set.** 32

2.8 **Sum of ranks for all treatments.** 33

2.9 **Sum of ranks for all treatments with Prolific pre-study data set.** 33

2.10 **Sum of ranks for all team treatments split up between WiSo-Lab and Prolific.** 34

2.11 **Sum of ranks for all team treatments split up between WiSo-Lab and Prolific with Prolific pre-study data set.** 34

2.12 **Welcome Page.** 38

2.13 **Video & Audio Test for Team Treatments.** 39

2.14 **Choice of Avatar - Individuals.** 40

2.15 **Choice of Avatar - Teams.** 40

2.16 **Instructions.** 41

2.17 **Main Task - Individuals.** 42

2.18 **Main Task - Teams.** 43

2.19 **Demographics.** 44

2.20 **Debriefing.** 44

3.1 **Sequence of events in the experiment.** 55

3.2 **Frequency of lying by group size.** 58

3.3 **Frequency of lying by group size and number of females in the group.** 60

3.4 **Groups' decision times (in seconds).** 63

3.5 **Frequency of group lying by number of cheaters in group.** 66

3.6 **Screenshot of the page "Individual video and audio test".** 74

3.7 **Screenshot of external test page.** 75

3.8 **Screenshot of the page "Individual general overview".** 76

3.9 **Screenshot of the page "Group video and audio test".** 77

3.10 **Screenshot of the page "Instructions and group decision".** 78

4.1 **Video chat test.** 91

5.1 **Overview of the experimental procedure and design.** 103

5.2 **Normative and empirical expectations across treatments.** 106

5.3 **Social image concerns across treatments.** 108

5.4 **Honesty rates across treatments.** 110

5.5 **Normative and empirical expectations across treatments and experiments.** 113

5.6 **Feedback by A players to B players.** 119

List of Tables

2.1 Treatment Overview. 14
2.2 Payoffs. 17
2.3 Number of observations per treatment. 18
2.4 Average enjoyment over all treatments. 23
2.5 Average fairness perception over all treatments. 24
2.6 Summary of participant demographics. 30

3.1 Number of group observations per treatment. 57
3.2 Robustness of main results. 67
3.4 Treatments with die roll outcome $\neq 5$. 86
3.5 Robustness of main results (all coefficients). 87

5.1 Treatment overview. 101
5.2 Linear regression: Normative expectations, empirical expectations, and personal norms. 107
5.3 Linear regression: Social image concerns. 109
5.4 Linear probability models and Logit regressions: Probability of an honest report. 111
5.5 Summary of participant demographics. 115
5.6 Summary of social norms and social image concerns across treatments and experiments. 116
5.7 Linear regression: Normative expectations, empirical expectations, and personal norms (complementary experiment). 117
5.8 Linear regression: Normative expectations, empirical expectations, and personal norms (complementary experiment). 117
5.9 Pre-Analysis Plan Discrepancies. 118

Bibliography

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for truth-telling," *Econometrica*, 87, 1115–1153.

Abraham, D., B. Greiner, and M. Stephanides (2023): "On the Internet you can be anyone: An experiment on strategic avatar choice in online marketplaces," *Journal of Economic Behavior & Organization*, 206, 251–261.

Acemoglu, D. and M. O. Jackson (2015): "History, expectations, and leadership in the evolution of social norms," *The Review of Economic Studies*, 82, 423–456.

Acharya, V. V. and T. C. Johnson (2010): "More insiders, more insider trading: Evidence from private-equity buyouts," *Journal of Financial Economics*, 98, 500–523.

Ajzenman, N. (2021): "The Power of Example: Corruption Spurs Corruption," *American Economic Journal: Applied Economics*, 13, 230–257.

Akerlof, G. A. and R. E. Kranton (2005): "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, 19, 9–32.

Alexander, L. and M. Moore (2016): "Deontological ethics," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, Stanford University.

Amason, A. and H. Sapienza (1997): "The effects of top management team size and interaction norms on cognitive and affective conflict," *Journal of Management*, 23, 495–516.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar (2009): "Large stakes and big mistakes," *The Review of Economic Studies*, 76, 451–469.

Association of Certified Fraud Examiners (2022): *Occupational Fraud 2022: A Report to the Nations*, https://legacy.acfe.com/report-to-the-nations/2022/.

Azmat, G. and B. Petrongolo (2015): "Gender and the labor market: What have we learned from field and lab experiments?" *Labour Economics*, 30, 32–40.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017): "Does the gender composition of scientific committees matter?" *American Economic Review*, 107, 1207–1238.

Bagues, M. F. and B. Esteve-Volart (2010): "Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment," *The Review of Economic Studies*, 77, 1301–1328.

Balafoutas, L. and M. Sutter (2012): "Affirmative action policies promote women and do not harm efficiency in the laboratory," *science*, 335, 579–582.

Baron-Cohen, S., S. Wheelwright, J. Hill, Y. Raste, and I. Plumb (2001): "The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism," *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42, 241–251.

Barron, K., A. Kajackaite, and S. Saccardo (2021): "Lying for image," *Available at SSRN 3977054*.

Bartling, B., E. Fehr, and Y. Özdemir (2023): "Does market interaction erode moral values?" *Review of Economics and Statistics*, 105, 226–235.

Bartling, B. and U. Fischbacher (2012): "Shifting the blame: On delegation and responsibility," *Review of Economic Studies*, 79, 67–87.

Bartling, B., R. Weber, and L. Yao (2015): "Do markets erode social responsibility?" *Quarterly Journal of Economics*, 130, 219–266.

Bašić, Z. and E. Verrina (2023): “Personal norms—and not only social norms—shape economic behavior,” *MPI Collective Goods Discussion Paper.*

Bašić, Z. and S. Quercia (2022): “The influence of self and social image concerns on lying,” *Games and Economic Behavior*, 133, 162–169.

Becker, G. S. (1968): “Crime and Punishment: An Economic Approach,” *Journal of Political Economy*, 76, 169–217.

Behnk, S., L. Hao, and E. Reuben (2022): “Shifting normative beliefs: On why groups behave more antisocially than individuals,” *European Economic Review*, 145, 104116.

Bénabou, R., A. Falk, L. Henkel, and J. Tirole (2020): “Eliciting moral preferences: Theory and experiment,” *Princeton University.*

Bénabou, R. and J. Tirole (2006): “Incentives and Prosocial Behavior,” *American Economic Review*, 96, 1652–1678.

Bénabou, R. and J. Tirole (2016): “Mindful economics: The production, consumption, and value of beliefs,” *Journal of Economic Perspectives*, 30, 141–64.

Berge, L., K. Juniwaty, and L. Sekei (2016): “Gender composition and group dynamics: Evidence from a laboratory experiment with microfinance clients,” *Journal of Economic Behavior & Organization*, 131, 1–20.

Bertrand, M. (2010): “New perspectives on gender,” in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 4, 1543–1590.

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

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

Bicchieri, C. and E. Dimant (2019): “Nudging with care: The risks and benefits of social information,” *Public Choice*, 1–22.

Bicchieri, C., E. Dimant, S. Gächter, and D. Nosenzo (2022): “Social proximity and the erosion of norm compliance,” *Games and Economic Behavior*, 132, 59–72.

Bicchieri, C., E. Dimant, and S. Sonderegger (2023): “It’s Not A Lie if You Believe the Norm Does Not Apply: Conditional Norm-Following and Belief Distortion,” *Games and Economic Behavior*, forthcoming.

Bicchieri, C. and R. Muldoon (2011): “Social Norms,” in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, Stanford University, mar 1, 2011. URL: https://plato.stanford.edu/archives/spr2011/entries/social-norms/, accessed Dec 16, 2021.

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

Bick, A., A. Blandin, and K. Mertens (2023): “Work from home before and after the COVID-19 outbreak,” *American Economic Journal: Macroeconomics*, 15, 1–39.

Bo, E. E., J. Slemrod, and T. O. Thoresen (2015): “Taxes on the Internet: Deterrence Effects of Public Disclosure,” *American Economic Journal: Economic Policy*, 7, 35–62.

Bock, O., I. Baetge, and A. Nicklisch (2014): “hroot: Hamburg registration and organization online tool,” *European Economic Review*, 71, 117–120.

Bodenschatz, A. and B. Irlenbusch (2019): “Do two bribe less than one? An experimental study on the four-eyes-principle,” *Applied Economics Letters*, 26, 191–195.

Bolton, G., E. Dimant, and U. Schmidt (2021): “Observability and social image: On the robustness and fragility of reciprocity,” *Journal of Economic Behavior & Organization*, 191, 946–964.

Born, A., E. Ranehill, and A. Sandberg (2022): "Gender and willingness to lead: Does the gender composition of teams matter?" *Review of Economics and Statistics*, 104, 259–275.

Bos, N., D. Gergle, J. S. Olson, and G. M. Olson (2001): "Being there versus seeing there: Trust via video," in *CHI'01 extended abstracts on human factors in computing systems*, 291–292.

Boyd, C. L., L. Epstein, and A. D. Martin (2010): "Untangling the Causal Effects of Sex on Judging," *American Journal of Political Science*, 54, 389–411.

Brosig, J. and J. Weimann (2003): "The effect of communication media on cooperation," *German Economic Review*, 4, 217–241.

Byrne, K. A., C. D. Silasi-Mansat, and D. A. Worthy (2015): "Who chokes under pressure? The Big Five personality traits and decision-making under pressure," *Personality and individual differences*, 74, 22–28.

Calsamiglia, C., J. Franke, and P. Rey-Biel (2013): "The incentive effects of affirmative action in a real-effort tournament," *Journal of Public Economics*, 98, 15–31.

Castillo, G., L. Choo, and V. Grimm (2022): "Are groups always more dishonest than individuals? The case of salient negative externalities," *Journal of Economic Behavior & Organization*, 198, 598–611.

Charness, G., U. Gneezy, and A. Henderson (2018): "Experimental methods: Measuring effort in economics experiments," *Journal of Economic Behavior & Organization*, 149, 74–87.

Charness, G., E. Karni, and D. Levin (2010): "On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda," *Games and Economic Behavior*, 68, 551–556.

Charness, G., A. Samek, and J. van de Ven (2022): "What is considered deception in experimental economics?" *Experimental Economics*, 25, 385–412.

Charness, G. and M. Sutter (2012): "Groups make better self-interested decisions," *Journal of economic perspectives*, 26, 157–176.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree–An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Childs, J. (2012): "Gender differences in lying," *Economics Letters*, 114, 147–149.

Chowdhury, S. M., P. Esteve-González, and A. Mukherjee (2023): "Heterogeneity, leveling the playing field, and affirmative action in contests," *Southern Economic Journal*, 89, 924–974.

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

Coffman, K., C. B. Flikkema, and O. Shurchkov (2021): "Gender stereotypes in deliberation and team decisions," *Games and Economic Behavior*, 129, 329–349.

Cohen, T. R., B. C. Gunia, S. Y. Kim-Jun, and J. K. Murnighan (2009): "Do groups lie more than individuals? Honesty and deception as a function of strategic self-interest," *Journal of Experimental Social Psychology*, 45, 1321–1324.

Cohn, A., T. Gesche, and M. A. Maréchal (2022): "Honesty in the digital age," *Management Science*, 68, 827–845.

Cohn, A. and M. A. Maréchal (2018): "Laboratory measure of cheating predicts school misconduct," *The Economic Journal*, 128, 2743–2754.

Cohn, A., M. A. Maréchal, and T. Noll (2015): "Bad boys: How criminal identity salience affects rule violation," *Review of Economic Studies*, 82, 1289–1308.

Conrads, J., B. Irlenbusch, R. M. Rilke, A. Schielke, and G. Walkowitz (2014): "Honesty in tournaments," *Economics Letters*, 123, 90–93.

Conrads, J., B. Irlenbusch, R. M. Rilke, and G. Walkowitz (2013): "Lying and team incentives," *Journal of Economic Psychology*, 34, 1–7.

Corchón, L. C., M. Serena, et al. (2018): "Contest theory," *Handbook of game theory and industrial organization*, 2, 125–146.

Crede, A.-K. and F. von Bieberstein (2020): "Reputation and lying aversion in the die roll paradigm: Reducing ambiguity fosters honest behavior," *Managerial and decision economics*, 41, 651–657.

Croson, R. and U. Gneezy (2009): "Gender differences in preferences," *Journal of Economic Literature*, 47, 448–474.

Dai, Z., F. Galeotti, and M. C. Villeval (2018): "Cheating in the lab predicts fraud in the field: An experiment in public transportation," *Management Science*, 64, 1081–1100.

Dana, J., R. Weber, and J. X. Kuang (2007): "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 33, 67–80.

Dannenberg, A. and E. Khachatryan (2020): "A comparison of individual and group behavior in a competition with cheating opportunities," *Journal of Economic Behavior & Organization*, 177, 533–547.

De Oliveira, A., R. Croson, and C. Eckel (2015): "One bad apple? Heterogeneity and information in public good provision," *Experimental Economics*, 18, 116–135.

DeCaro, M. S., R. D. Thomas, N. B. Albert, and S. L. Beilock (2011): "Choking under pressure: multiple routes to skill failure." *Journal of experimental psychology: general*, 140, 390.

Dechenaux, E., D. Kovenock, and R. M. Sheremeta (2015): "A survey of experimental research on contests, all-pay auctions and tournaments," *Experimental Economics*, 18, 609–669.

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

Dimant, E. and T. Gesche (2023): "Nudging enforcers: How norm perceptions and motives for lying shape sanctions," *PNAS nexus*, 2, pgad224.

Dimant, E. and S. Shalvi (2022): "Meta-nudging honesty: Past, present, and future of the research frontier," *Current Opinion in Psychology*, 101426.

Dimant, E., G. A. Van Kleef, and S. Shalvi (2020): "Requiem for a nudge: Framing effects in nudging honesty," *Journal of Economic Behavior & Organization*, 172, 247–266.

Dimmock, S. G., W. C. Gerken, and N. P. Graham (2018): "Is Fraud Contagious? Coworker Influence on Misconduct by Financial Advisors," *Journal of Finance*, 73, 1417–1450.

Dohmen, T. J. (2008): "Do professionals choke under pressure?" *Journal of economic behavior & organization*, 65, 636–653.

Dreber, A. and M. Johannesson (2008): "Gender differences in deception," *Economics Letters*, 99, 197–199.

Drugov, M. and D. Ryvkin (2017): "Biased contests for symmetric players," *Games and Economic Behavior*, 103, 116–144.

——— (2022): "Hunting for the discouragement effect in contests," *Review of Economic Design*, 1–27.

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

Dufwenberg, M. and A. Muren (2006): "Gender composition in teams," *Journal of Economic Behavior & Organization*, 61, 50–54.

Dyck, A., A. Morse, and L. Zingales (2023): "How pervasive is corporate fraud?" *Review of Accounting Studies*, 1–34.

Economist (2020): “The Number of the Best,” *25 January: 53.*

Ellingsen, T. and M. Johannesson (2004): “Promises, threats and fairness,” *The Economic Journal*, 114, 397–420.

Engl, F. (2022): “Ideological Motives and Group Decision-Making,” *Available at SSRN 3738759.*

Englmaier, F., S. Grimm, D. Schindler, and S. Schudy (2024): “The effect of incentives in non-routine analytical teams tasks-evidence from a field experiment,” *Journal of Political Economy, forthcoming.*

Erat, S. and U. Gneezy (2012): “White lies,” *Management Science*, 58, 723–733.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2023): “The preference survey module: A validated instrument for measuring risk, time, and social preferences,” *Management Science*, 69, 1935–1950.

Falk, A., T. Neuber, and N. Szech (2020): “Diffusion of being pivotal and immoral outcomes,” *Review of Economic Studies*, 87, 2205–2229.

Falk, A. and N. Szech (2013): “Morals and markets,” *Science*, 340, 707–711.

Farhang, S. and G. Wawro (2004): “Institutional Dynamics on the US Court of Appeals. Minority Representation under Panel Decision Making,” *Journal of Law, Economics, and Organization*, 20, 299–330.

Feddersen, T. and W. Pesendorfer (1998): “Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting,” *American Political Science Review*, 92, 23–35.

Feess, E., F. Kerzenmacher, and G. Muehlheusser (2023): “Morally questionable decisions by groups: Guilt sharing and its underlying motives,” *Games and Economic Behavior*, 140, 380–400.

Feess, E., F. Kerzenmacher, and Y. Timofeyev (2022): “Utilitarian or deontological models of moral behaviorWhat predicts morally questionable decisions?” *European Economic Review*, 149, 104264.

Feess, E., G. Muehlheusser, and M. Walzl (2008): “Unfair contests,” *Journal of Economics*, 267–291.

Fischbacher, U. (1999): “Z-tree: A toolbox for readymade economic experiments,” Tech. rep., IEW Working paper 21, University of Zurich.

——— (2007): “z-Tree: Zurich toolbox for ready-made economic experiments,” *Experimental Economics*, 10, 171–178.

Fischbacher, U. and F. Föllmi-Heusi (2013): “Lies in disguise—an experimental study on cheating,” *Journal of the European Economic Association*, 11, 525–547.

Franke, J., C. Kanzow, W. Leininger, and A. Schwartz (2013): “Effort maximization in asymmetric contest games with heterogeneous contestants,” *Economic Theory*, 52, 589–630.

Franke, J., W. Leininger, and C. Wasser (2018): “Optimal favoritism in all-pay auctions and lottery contests,” *European Economic Review*, 104, 22–37.

Fries, T., U. Gneezy, A. Kajackaite, and D. Parra (2021): “Observability and lying,” *Journal of Economic Behavior & Organization*, 189, 132–149.

Gächter, S. and J. F. Schulz (2016): “Intrinsic honesty and the prevalence of rule violations across societies,” *Nature*, 531, 496–499.

Gibson, R., C. Tanner, and A. F. Wagner (2013): “Preferences for truthfulness: Heterogeneity among and within individuals,” *American Economic Review*, 103, 532–548.

Gill, D. and V. Prowse (2012): “A structural analysis of disappointment aversion in a real effort competition,” *American Economic Review*, 102, 469–503.

Gneezy, U. (2005): “Deception: The role of consequences,” *American Economic Review*, 95, 384–394.

Gneezy, U., A. Kajackaite, and J. Sobel (2018): “Lying aversion and the size of the lie,” *American Economic Review*, 108, 419–453.

Gneezy, U., B. Rockenbach, and M. Serra-Garcia (2013): “Measuring lying aversion,” *Journal of Economic Behavior and Organization*, 93, 293–300.

Goldberg, L. R. (1992): “The development of markers for the Big-Five factor structure.” *Psychological Assessment*, 4, 26.

Gormley, T. A., V. K. Gupta, D. A. Matsa, S. C. Mortal, and L. Yang (2023): “The Big Three and Board Gender Diversity: The Effectiveness of Shareholder Voice,” *Journal of Financial Economics*, 149, 323–348.

Grözinger, N., B. Irlenbusch, K. Laske, and M. Schröder (2020): “Innovation and communication media in virtual teams–An experimental study,” *Journal of Economic Behavior & Organization*, 180, 201–218.

Haleblian, J. and S. Finkelstein (1993): “Top management team size, CEO dominance, and firm performance: The moderating roles of environmental turbulence and discretion,” *Academy of Management Journal*, 36, 844–863.

Hanna, R. and S.-Y. Wang (2017): “Dishonesty and selection into public service: Evidence from India,” *American Economic Journal: Economic Policy*, 9, 262–90.

Hardt, D., L. Mayer, and J. Rincke (2023): “Who Does the Talking Here? The Impact of Gender Composition on Team Interactions,” *CESifo Working Paper No. 10550.*

Haushofer, J. and J. Shapiro (2016): “The short-term impact of unconditional cash transfers to the poor: Experimental evidence from Kenya,” *Quarterly Journal of Economics*, 131, 1973–2042.

Hertwig, R. and A. Ortmann (2008): “Deception in Experiments: Revisiting the Arguments in Its Defense,” *Ethics & Behavior*, 18, 59–92.

Hey, J. D. (1998): “Experimental economics and deception: A comment,” *Journal of Economic Psychology*, 19, 397–401.

Heyman, J. and D. Ariely (2004): “Effort for payment: A tale of two markets,” *Psychological science*, 15, 787–793.

Holt, C. A. and S. K. Laury (2002): “Risk aversion and incentive effects,” *American Economic Review*, 92, 1644–1655.

Houser, D., J. List, M. Piovesan, A. Samek, and J. Winter (2016): “Dishonesty: From parents to children,” *European Economic Review*, 82, 242–254.

Hu, M. and B. Liu (2004): “Mining and summarizing customer reviews,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177.

Huber, C. and J. Huber (2020): “Bad bankers no more? Truth-telling and (dis)honesty in the finance industry,” *Journal of Economic Behavior & Organization*, 180, 472–493.

Huber, C., C. Litsios, A. Nieper, and T. Promann (2023): “On social norms and observability in (dis) honest behavior,” *Journal of Economic Behavior & Organization*, 212, 1086–1099.

Hugh-Jones, D. (2016): “Honesty, beliefs about honesty, and economic growth in 15 countries,” *Journal of Economic Behavior & Organization*, 127, 99–114.

Irlenbusch, B., T. Mussweiler, D. Saxler, S. Shalvi, and A. Weiss (2020): “Similarity increases collaborative cheating,” *Journal of Economic Behavior & Organization*, 178, 148–173.

Irlenbusch, B. and D. J. Saxler (2019): “The role of social information, market framing, and diffusion of responsibility as determinants of socially responsible behavior,” *Journal of Behavioral and Experimental Economics*, 80, 141–161.

Isler, O. and S. Gächter (2022): “Conforming with peers in honesty and cooperation,” *Journal of Economic Behavior & Organization*, 195, 75–86.

Jensen, L. A., J. J. Arnett, S. S. Feldman, and E. Cauffman (2002): “It’s wrong, but everybody does it: Academic dishonesty among high school and college students,” *Contemporary Educational Psychology*, 27, 209–228.

Kajackaite, A. and U. Gneezy (2017): “Incentives and cheating,” *Games and Economic Behavior*, 102, 433–444.

Karpowitz, C., S. O’Connell, J. Preece, and O. Stoddard (2024): “Strength in Numbers? Gender Composition, Leadership, and Women’s Influence in Teams,” *Journal of Political Economy, forthcoming.*

Kartik, N. (2009): “Strategic communication with lying costs,” *The Review of Economic Studies*, 76, 1359–1395.

Keck, S. and W. Tang (2018): “Gender composition and group confidence judgment: The perils of all-male groups,” *Management Science*, 64, 5877–5898.

Khalmetski, K. and D. Sliwka (2019): “Disguising lies – Image concerns and partial lying in cheating games,” *American Economic Journal: Microeconomics*, 11, 79–110.

Kirkegaard, R. (2012): “Favoritism in asymmetric contests: Head starts and handicaps,” *Games and Economic Behavior*, 76, 226–248.

Kocher, M. G., S. Schudy, and L. Spantig (2018): “I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups,” *Management Science*, 64, 3995–4008.

Konrad, K. A. (2009): *Strategy and dynamics in contests*, Oxford University Press.

Kroll (2016): *Global Fraud Report: Vulnerability on the Rise*, http://www.kroll.com/en-us/global-fraud-report.

Krupka, E. L. and R. A. Weber (2013): “Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?” *Journal of the European Economic Association*, 11, 495–524.

Kugler, T., E. E. Kausel, and M. G. Kocher (2012): “Are groups more rational than individuals? A review of interactive decision making in groups,” *Wiley interdisciplinary reviews: Cognitive science*, 3, 471–482.

Köbis, N. C., J.-W. van Prooijen, F. Righetti, and P. A. M. Van Lange (2015): ““Who Doesn’t?”—The Impact of Descriptive Norms on Corruption,” *PLOS ONE*, 10, e0131830.

Laughlin, P., E. Hatch, J. Silver, and L. Boh (2006): “Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size.” *Journal of Personality and Social Psychology*, 90, 644–651.

Lazear, E. P. and K. L. Shaw (2007): “Personnel economics: The economist’s view of human resources,” *Journal of Economic Perspectives*, 21, 91–114.

Lee, J. J. and J. M. McCabe (2021): “Who speaks and who listens: Revisiting the chilly climate in college classrooms,” *Gender & Society*, 35, 32–60.

Lois, G. and M. Wessa (2020): “Honest mistake or perhaps not: The role of descriptive and injunctive norms on the magnitude of dishonesty,” *Journal of Behavioral Decision Making*, 34, 20–34.

Mas, A. and E. Moretti (2009): “Peers at work,” *American Economic Review*, 99, 112–145.

Matsa, D. and A. Miller (2013): “A female style in corporate leadership? Evidence from quotas,” *American Economic Journal: Applied Economics*, 5, 136–69.

Mealem, Y. and S. Nitzan (2016): “Discrimination in contests: a survey,” *Review of Economic Design*, 20, 145–172.

Mobius, M. M. and T. S. Rosenblat (2006): "Why beauty matters," *American Economic Review*, 96, 222–235.

Muehlheusser, G., T. Promann, A. Roider, and N. Wallmeier (2024): "Honesty of groups: The effects of group size and group gender composition," In preparation.

Muehlheusser, G., A. Roider, and N. Wallmeier (2015): "Gender Differences in Honesty: Groups versus Individuals," *Economics Letters*, 128, 25–29.

Mukhopadhaya, K. (2003): "Jury size and the free rider problem," *Journal of Law, Economics, and Organization*, 19, 24–44.

Niederle, M. (2016): "Gender," in *The Handbook of Experimental Economics*, ed. by J. Kagel and A. Roth, Princeton University Press, vol. 2, 481–563.

Niederle, M., C. Segal, and L. Vesterlund (2013): "How costly is diversity? Affirmative action in light of gender differences in competitiveness," *Management Science*, 59, 1–16.

Niederle, M. and L. Vesterlund (2007): "Do women shy away from competition? Do men compete too much?" *The Quarterly Journal of Economics*, 122, 1067–1101.

Palan, S. and C. Schitter (2018): "Prolific.ac – A subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, 17, 22–27.

Peresie, J. L. (2005): "Female Judges Matter. Gender and Collegial Decisionmaking in the Federal Appellate Courts," *Yale Law Journal*, 114, 1759–1790.

Plott, C. and V. Smith (2008): *Handbook of experimental economics results*, vol. 1, Elsevier.

Potters, J. and J. Stoop (2016): "Do cheaters in the lab also cheat in the field?" *European Economic Review*, 87, 26–33.

Promann, T. (2024): "Unfair Contests - Experimental Evidence for Individuals and Teams," In preparation.

Radbruch, J. and A. Schiprowski (2023): "Committee Deliberation and Gender Differences in Influences," *CRC TRR 190 Discussion Paper No. 398.*

Raven, J. (1995): *Advanced progressive matrices*, Oxford Psychologists Press.

Rothenhäusler, D., N. Schweizer, and N. Szech (2018): "Guilt in voting and public good games," *European Economic Review*, 101, 664–681.

Schmidt, R., F. Heinicke, and C. König-Kersting (2022): "Using coordination games to measure beliefs," *Economics Letters*, 219, 110821.

Schotter, A. and K. Weigelt (1992): "Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results," *The Quarterly Journal of Economics*, 107, 511–539.

Serdarevic, N. (2021): "Licence to lie and the social (In)appropriateness of lying," *Economics Letters*, 199, 109716.

Siegel, R. (2010): "Head starts in contests," *Available at SSRN 1743795.*

——— (2014a): "Asymmetric all-pay auctions with interdependent valuations," *Journal of Economic Theory*, 153, 684–702.

——— (2014b): "Asymmetric contests with head starts and nonmonotonic costs," *American Economic Journal: Microeconomics*, 6, 59–105.

Sinnott-Armstrong, W. (1988): *Moral dilemmas*, Blackwell Publishers.

Slemrod, J. (2007): "Cheating Ourselves: The Economics of Tax Evasion," *Journal of Economic Perspectives*, 21, 25–48.

Sutter, M. (2005): "Are four heads better than two? An experimental beauty-contest game with teams of different size," *Economics Letters*, 88, 41–46.

——— (2009): "Deception through telling the truth?! Experimental evidence from individuals and teams," *The Economic Journal*, 119, 47–60.

Swire-Thompson, B., U. K. Ecker, S. Lewandowsky, and A. J. Berinsky (2020): "They might be a liar but theyre my liar: Source evaluation and the prevalence of misinformation," *Political Psychology*, 41, 21–34.

Thompson, L. (2017): *Making the team: A guide for managers (6th edition)*, Pearson.

Useem, J. (2006): "How to build a great team," *FORTUNE Magazine*, retrieved from https://money.cnn.com/2006/05/31/magazines/fortune/intro_greatteams_fortune_061206/index.htm.

Vojnović, M. (2015): *Contest theory: Incentive mechanisms and ranking methods*, Cambridge University Press.

Von Gaudecker, H.-M., R. Holler, L. Janys, B. Siflinger, and C. Zimpelmann (2020): "Labour supply in the early stages of the CoViD-19 Pandemic: Empirical Evidence on hours, home office, and expectations," *IZA Discussion Paper No. 13158.*

Weibull, J. W. and E. Villa (2005): "Crime, punishment and social norms," Tech. rep., SSE/EFI Working Paper Series in Economics and Finance.

Weidmann, B. and D. J. Deming (2021): "Team players: How social skills improve team performance," *Econometrica*, 89, 2637–2657.

Weisel, O. and S. Shalvi (2015): "The collaborative roots of corruption," *Proceedings of the National Academy of Sciences of the United States of America*, 112, 10651–10656.

Wilson, B. J. (2016): "The meaning of deceive in experimental economic science," in *The Oxford Handbook of Professional Economic Ethics*, Oxford University Press.

Wu, J., D. Balliet, and P. A. Van Lange (2015): "When does gossip promote generosity? Indirect reciprocity under the shadow of the future," *Social Psychological and Personality Science*, 6, 923–930.

Zhu, F. (2021): "On optimal favoritism in all-pay contests," *Journal of Mathematical Economics*, 95, 102472.

Anhang der Dissertation

Zusammenfassungen

Chapter 2: *Unfair Contests - Experimental Evidence for Individuals and Teams*

I study the impact of head starts in contests on the performance of individuals and teams. In an online experiment, participants compete in a cognitive effort task for a doubled payoff. Individuals show their best performance when they are given a large head start. They perform second best when the potential head start for either competitor is unknown. Individual participants perform the worst when the contest is designed completely fair. Teams distinctly outperform individuals but remain virtually unaffected by head starts. In summation, my findings contradict central findings of contest theory, namely that fair contests always induce the highest performance.

Ich untersuche den Einfluss von Wettbewerbsvorteilen auf die Leistung von Individuen und Teams. In einem Online-Experiment konkurrieren die Teilnehmenden in einer kognitiven Aufgabe um die Verdopplung ihrer Auszahlung. Individuen, die vorab einen großen Vorteil erhalten, zeigen die stärkste Leistung. Die zweitstärkste Leistung liefern Individuen, denen die Bevorteilung für sie oder ihre Gegenspielerin nicht bekannt ist. Wenn keiner der beiden Gegenspieler einen Vorteil erhält, zeigen Individuen ihre schwächste Leistung. Teams übertreffen Individuen in der gegebenen Aufgabe deutlich, zeigen jedoch kaum eine Reaktion auf die Bevorteilung von ihnen oder ihrem Gegnerteam. Zusammengenommen stellen meine Ergebnisse einige zentrale Erkenntnisse der Wettbewerbstheorie in Frage, insbesondere, dass ein fairer Wettbewerb immer zur höchsten Leistung der Wettbewerber führt.

Chapter 3: *Honesty of groups: Effects of size and gender composition*

We examine the effects of size and gender composition on a groups' honesty in an online experiment. We vary the group size between 2, 3, 4 and 5 while investigating every possible gender composition of each group size. Groups have to form a unanimous decision whether to behave honestly or dishonestly. Participants communicate via a novel video chat tool that is directly integrated in the experiment. We find that all-male groups lie significantly more than all other gender compositions. This effect completely disappears when one female is added to the group. Furthermore, dishonesty significantly increases with group size. Our findings add interesting insights to

the active debate concerning gender quotas and strongly advise against all-male groups whenever ethical behavior is relevant.

Wir analysieren die Effekte von Größe und geschlechtlicher Zusammensetzung auf die Ehrlichkeit einer Gruppe. Hierzu variieren wir die Gruppengröße zwischen 2, 3, 4 und 5 während wir jede mögliche geschlechtliche Zusammensetzung dieser Gruppengrößen untersuchen. Jede Gruppe muss eine einstimmige Entscheidung treffen, ob sie sich ehrlich oder unehrlich verhalten möchte. Die Kommunikation innerhalb einer Gruppe findet über ein neu aufgebautes Video Chat Tool statt, welches direkt im Experiment integriert ist. Das hervorstechenste Ergebnis ist, dass reine Männergruppen deutlich häufiger lügen als jede andere geschlechtliche Zusammensetzung einer Gruppe. Eine Frau in der Gruppe reicht jedoch aus, um diesen Effekt verschwinden zu lassen. Zusätzlich nimmt Unehrlichkeit mit steigender Gruppengröße zu. Unsere Ergebnisse liefern interessante Einsichten zur Diskussion bezüglich Geschlechter-Quoten und raten stark davon ab, reine Männergruppen in Bereichen einzusetzen, in denen ethisches Verhalten von Bedeutung ist.

Chapter 4: *ChaTree - video chat integration in oTree*

We build a video chat tool that can be integrated in oTree. Thereby, participants of an online experiment can directly access a video conference by entering the designated web-page. This short article provides instructions and access to code such that an easy integration of the video chat tool in any oTree-based online experiment is possible. The ability to use a video conference in online experiments strongly facilitates the exploration of online face-to-face communication. Especially during but also after the COVID-19 pandemic, working from home distinctly increased, rendering the investigation of online- communication and behavior very relevant.

Wir haben ein Video Chat Tool geschaffen, dass in oTree-Experimente integriert werden kann. Teilnehmende an Online-Experimenten können am Video Chat automatisch teilnehmen, sobald sie die dafür vorgesehene Website des Experiments betreten. Dieser kurze Artikel stellt eine Anleitung sowie Zugang zum Code zur Verfügung, sodass das Video Chat Tool in jedes oTree-basierte Online-Experiment integriert werden kann. Einen Video Chat in Online-Experimenten zu nutzen, ermöglicht zunehmend digitale face-to-face Kommunikation zu untersuchen. Da insbesondere während, jedoch auch nach der COVID-19 Pandemie, die Arbeit aus dem Home Office deutlich zugenommen hat, enthält die verstärkte Untersuchung von Online-Kommunikation sowie -Verhalten hohe Relevanz.

Chapter 5: *On Social Norms and Observability in (Dis)honest Behavior*

We conduct an online experiment to investigate (interacting) effects of social norms and reputational concerns on (dis)honest behavior. Participants take part in a variation of the die roll game. The treated participants are confronted with an overview of decisions by other participants before they take their own decision. Participants who observed honest decisions behave significantly more honest than participants who observed dishonest decisions. Increased observabilty of the treated participants had no impact on their honesty. Therefore, also the interaction between social norm and observability did not affect honesty. Hence, our study confirms that norm nudges can be used to impact (dis)honest behavior but finds no effect of reputational concerns in a comparably anonymous online environment.

Mithilfe eines Online-Experiments untersuchen wir etwaige (Interaktions-)Effekte von Sozialen Normen und der Sorge um die eigene Reputation auf das Ehrlichkeitsverhalten von Probanden. Die Testpersonen nehmen an einem Würfelspiel teil und werden mit den Entscheidungen anderer Spieler konfrontiert. Teilnehmende, die die Entscheidungen ehrlicher Spielerinnen beobachtet haben, verhalten sich deutlich ehrlicher als Teilnehmende, die die Entscheidung von unehrlichen Spielerinnen beobachteten. Die Entscheidungen der Teilnehmenden an weitere Spieler zu kommunizieren hatte keinen Effekt auf das Ehrlichkeitsverhalten. Somit gibt es in unserer Studie ebenfalls keinen Interaktionseffekt von Sozialen Normen und Reputationssorgen auf das eigene Ehrlichkeitsverhalten. Als Ergebnis ist festzuhalten, dass sich Ehrlichkeitsverhalten in einer vergleichsweise anonymen Online-Umgebung über die soziale Norm jedoch nicht über Reputationseffekte beeinflussen lässt.

Liste der aus dieser Dissertation hervorgegangenen Veröffentlichungen

Huber, C., Litsios, C., Nieper, A., & Promann, T. (2023). On social norms and observability in (dis) honest behavior. *Journal of Economic Behavior & Organization, 212*, 1086-1099.

Selbstdeklaration bei kumulativen Promotionen

Konzeption / Planung: Formulierung des grundlegenden wissenschaftlichen Problems, basierend auf bisher unbeantworteten theoretischen Fragestellungen inklusive der Zusammenfassung der generellen Fragen, die anhand von Analysen oder Experimenten / Untersuchungen beantwortbar sind. Planung der Experimente / Analysen und Formulierung der methodischen Vorgehensweise, inklusive Wahl der Methode und unabhängige methodologische Entwicklung.

Durchführung: Grad der Einbindung in die konkreten Untersuchungen bzw. Analysen.

Manuskripterstellung: Präsentation, Interpretation und Diskussion der erzielten Ergebnisse in Form eines wissenschaftlichen Artikels.

Die Einschätzung des geleisteten Anteils erfolgt mittels Punkteinschätzung von 1-100%.

Für einen der vorliegenden Artikel (chapter 2) liegt die Eigenleistung bei 100%.

Für einen dritten Artikel (chapter 3) liegt die Eigenleistung für

das Konzept / die Planung bei	30%
die Durchführung bei	60%
der Manuskripterstellung bei	20%

Für einen vierten Artikel (chapter 4) liegt die Eigenleistung für

das Konzept / die Planung bei	60%
die Durchführung bei	70%
der Manuskripterstellung bei	95%

Für einen zweiten Artikel (chapter 5) liegt die Eigenleistung für

das Konzept / die Planung bei	30%
die Durchführung bei	45%
der Manuskripterstellung bei	30%

Die vorliegende Einschätzung in Prozent über die von mir erbrachte Eigenleistung wurde mit den am Artikel beteiligten Koautoren einvernehmlich abgestimmt.

Hamburg, 22.04.2024 T. Pro
Ort/Datum Doktorand/in

Erklärung

Hiermit erkläre ich, Timo Promann, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die Arbeit wurde nicht schon einmal in einem früheren Promotionsverfahren angenommen oder als ungenügend beurteilt.

Hamburg, 22.04.2024 T. Pro
Ort/Datum Doktorand/in

Eidesstattliche Versicherung

Ich, Timo Promann, versichere an Eides statt, dass ich die Dissertation mit dem Titel:

"*Experimental Evidence on Individual and Collective Decision-Making in Ethical Dilemmas and Unfair Contests*"

selbst und bei einer Zusammenarbeit mit anderen Wissenschaftlerinnen oder Wissenschaftlern gemäß den beigefügten Darlegungen nach § 6 Abs. 3 der Promotionsordnung der Fakultät für Wirtschafts- und Sozialwissenschaften vom 18. Januar 2017 verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht benutzt.

Hamburg, 22.04.2024 T. Pro
Ort/Datum Doktorand/in