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

DISCUSSION PAPER SERIES

IZA DP No. 16954

Honesty of Groups: Effects of Size and Gender Composition

Gerd Muehlheusser
Timo Promann
Andreas Roider
Niklas Wallmeier

APRIL 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16954

Honesty of Groups: Effects of Size and Gender Composition

Gerd Muehlheusser
University of Hamburg, IZA and CESifo

Timo Promann
University of Hamburg

Andreas Roider
University of Regensburg, CEPR, IZA and CESifo

Niklas Wallmeier
arq decisions GmbH

APRIL 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Honesty of Groups: Effects of Size and Gender Composition*

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.

JEL Classification: C92, J16, D70

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

Corresponding author:
Andreas Roider
Department of Economics
University of Regensburg
Universitätsstr. 31
93040 Regensburg
Germany
E-mail: andreas.roider@ur.de

* The project was preregistered at the AEA RCT Registry (AEARCTR-0008564). We gratefully acknowledge comments and suggestions by Jean-Michel Benkert, Karen Bernhardt-Walther, Frauke von Bieberstein, Christoph Engel, Eberhard Feess, Jana Gallus, Rubén Poblete, Simeon Schudy, Ben Weidmann, Stephane Wolton, as well as from seminar audiences at the workshop "At the Crossroads of Experimental Law and Economics" (Rotterdam), the "Organizational Economics Summer Symposium" (Ohlstadt), and at the Universities of Bern, Graz, Hamburg, Lüneburg, Passau, and Regensburg. Jan-Patrick Mayer, and Eugen Tereschenko provided excellent research assistance. We gratefully acknowledge financial support by the German Research Foundation (DFG, Grant 658143 and Research Training Group 2503/1 "Collective Decision-Making").

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?

The answers to these questions might very well depend on the type of decision, i.e. whether a given group task is *intellective* or *judgmental* in nature (see e.g. Kugler, Kausel, and Kocher, 2012). Intellective decisions (such as solving complex, possibly non-routine problems) mainly rely on the group members' cognitive skills and effort (for a recent field experiment with groups, see e.g. Englmaier, Grimm, Schindler, and Schudy, 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.[1] 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 1996–2004 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

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

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.

To the best of our knowledge, there is so far no systematic evidence (i.e. within the same study) on the effects of group size. For example, do the observed effects for dyads and triads extend to larger groups, and is unethical behavior increasing or decreasing in group size? 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 average team sizes to be around 3.4 with standard deviations of 1.2-1.5 (see e.g. Haleblian and Finkelstein, 1993; Amason and Sapienza, 1997).[2] Moreover, according to estimates by the Association of Certified Fraud Examiners (2022), groups of three or more perpetrators are responsible for 38% of cases of corporate fraud, while the numbers for dyads and for individuals are 20% and 42%, respectively.

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.[3]

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.[4] 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).[5] Overall, males seem to be somewhat more prone to unethical behavior than females.[6] Much less is known about how (unethical) group

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

[3] Evidence on the former (in the context of delegation) is provided by Bartling and Fischbacher (2012). In their survey, Abeler, Nosenzo, and Raymond (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).

[4] For example, the focus on the group gender composition is highlighted by a recent, successful influence campaign by the "Big Three" asset managers in the US to increase female representation on corporate boards (Gormley, Gupta, Matsa, Mortal, and Yang, 2023).

[5] While reality is more complex, most of the literature on gender effects focusses on potential behavioral differences between females and males. 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.

[6] 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

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 explanations would lead to different predictions for the effect of the group gender composition in larger groups.

Framework To the best of our knowledge, in the economics literature in group decisions there does not yet exist a systematic analysis (i.e. within the same study) of the effects of group size and gender composition in the domain of unethical decisions. In this paper, we present results 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.[7] 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.[8] They are given ample time to discuss their report using the video chat. If a group fails to reach an agreement till the end of the discussion period, the payoff from the group task is zero for each group member.

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 five that lie is more than twice as high compared to the case of groups of two. We also find that the group size has

employees, managers, and owners.

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

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

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 stand out regarding their proclivity to lie, and that the prevalence of lying in groups does not strictly decrease with the number of females in the group. Rather, it seems to be the *first* female in the group that causes an honesty shift.

We then study the effects of group size and 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 that do not contain any cheaters we show that the frequency of lying is still 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 after the group task was completed. 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 2 discusses the related literature. Section 3 explains the experimental design and implementation. Section 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, the robustness of the results in a regression analysis, and groups that did not reach an agreement. Section 5 discusses our findings and concludes. The Appendix provides additional empirical results, and the experimental instructions.

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.[9]

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 preferences (i) for being honest and (ii) for 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, recent 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 Fis-

[9] Using non-incentivized experiments, the effects of group size are also studied in psychology (see e.g., Laughlin, Hatch, Silver, and Boh, 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).

chbacher and Föllmi-Heusi (2013) to group settings, where group members have to make a joint decision (as does the present paper).[10] For example, Kocher, Schudy, and Spantig (2018) study lying by individuals and triads, where they focus on the role of within-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 it is not the experimenter that 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 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, Irlenbusch, Rilke, and Walkowitz (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, Mussweiler, Saxler, Shalvi, and Weiss (2020) study the role of feelings of similarity and a code of conduct in a setting where members of a dyad observe and report a die-roll outcome sequentially.

Apart from the die-roll 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 the form of cheap talk games (Gneezy, 2005).[11] 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, Gunia, Kim-Jun, and Murnighan, 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.[12]

The third area to which our paper contributes is the **literature on how group decisions are**

[10] 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).

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

[12] 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.

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 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. One exception 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 Experiment

In this section, we describe the design (Section 3.1) and the implementation (Section 3.2) of the experiment. The instructions are provided in Appendix B.

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 there are no treatment variations. They serve to elicit individual preferences and characteristics, and they are discussed in more detail in Sections 4.4 and 4.5 below.

In Task 1, each group needs to reach a decision on which number to report. 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, i.e. the group might act honestly or dishonestly. To obtain a positive payoff, a group must reach a unanimous agreement concerning the number they jointly report.[13] 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 the true outcome of the die roll is 5, a group can increase its payoff by agreeing to

[13] 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).

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 3.2, 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.[14]

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. $3 + 4 + 5 + 6$).

3.2 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 an 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. 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.[15]

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.[16] Thereby, it is also possible to track each group member's communication patterns (e.g. the frequency, volume, and duration of contributions).[17]

As illustrated in Figure 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

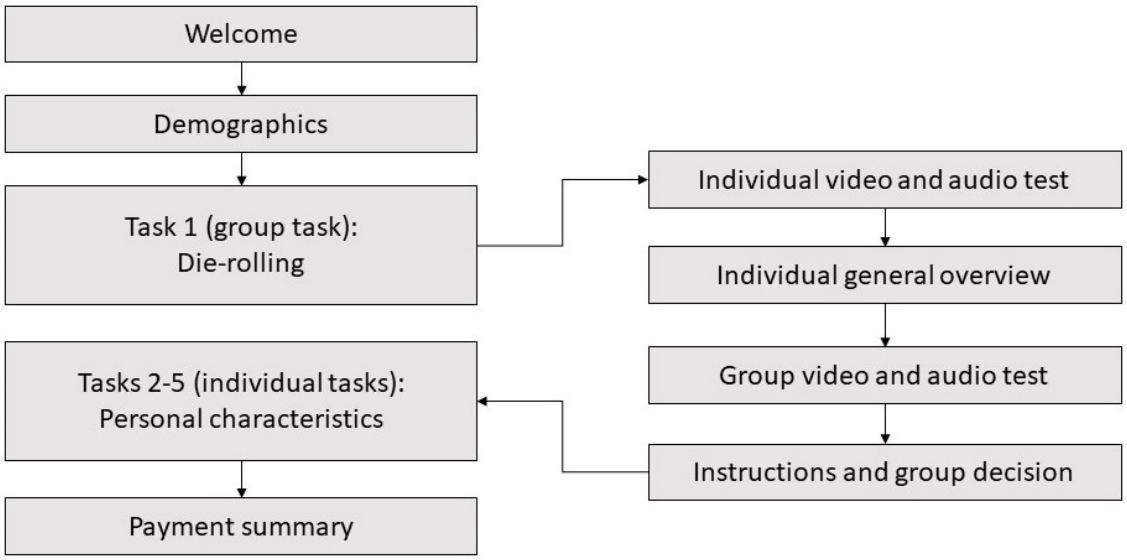
[14] In their study of gender effects in group leadership decisions, Born, Ranehill, and Sandberg (2022) also emphasize the desirability of groups being able to interact and deliberate face-to-face.

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

[16] 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.

[17] 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.

Figure 1: Sequence of events in the experiment



1) would be performed in a group. We informed the subjects that the details on each task (including information on the payment scheme) are 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 1, and for screenshots see Figures 6 to 10 in Appendix B). 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 6 in Appendix B).[18] Subjects who successfully completed the functionality test then proceeded to the screen "Individual general overview", where we gave them some basic information regarding the structure of the upcoming group interaction (see Figure 8 in Appendix B). 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

[18] 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 microphone and rates them on scores ranging from 0 to 4.5 (for an example, see Figure 7 in Appendix B). 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.

hear all other group members before being able to proceed (see Figure 9 in Appendix B).[19]

Finally, on the screen "Instructions and group decision", subjects first had three minutes to read the instructions for Task 1 (see Figure 10 in Appendix B). 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 roll 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 all of the group members (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.[20] 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 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 1. In the preregistration, we specified the experimental design and that we aimed to study the effects of group size and group gender composition through our 18 treatments.[21] Given the lack of clear theoretical predictions, we did not preregister any directed hypotheses.

[19] 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.

[20] 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 automatically proceeded to Task 2.

[21] 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 1.

Table 1: Number of group observations per treatment

	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

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.[22] In particular, we report on how lying behavior is affected by group size (Section 4.1) and by the group gender composition as measured by the number of females in the group (Section 4.2).[23] 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 4.3).[24] In Section 4.4 we investigate how group members' individual honesty affects group lying behavior. In Section 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 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 members of such groups).

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.

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

The result is illustrated in Figure 2 and supported by a highly statistically significant Jonckheere-Terpstra test for the presence of a trend ($p = 0.002$).[25] For example, groups of four and five are

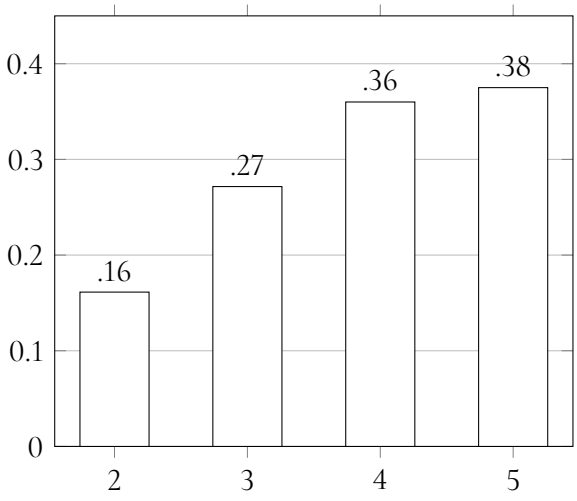
[22] For an overview of how these group observations are distributed across treatments, see Table 3 in Appendix A.

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

[24] 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.

[25] A positive effect of group size on the frequency of lying is also confirmed in a regression analysis (see Table 2 below).

Figure 2: Frequency of lying by group size



Notes: The figure is based on 62, 81, 100 and 120 observations for group size $n = 2, 3, 4, 5$, respectively.

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).

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).[26] This reasoning can also rationalize the absence of partial lying in our online group setting, where the die roll is observed by the experimenter.

4.2 The effect of the group gender composition on lying behavior

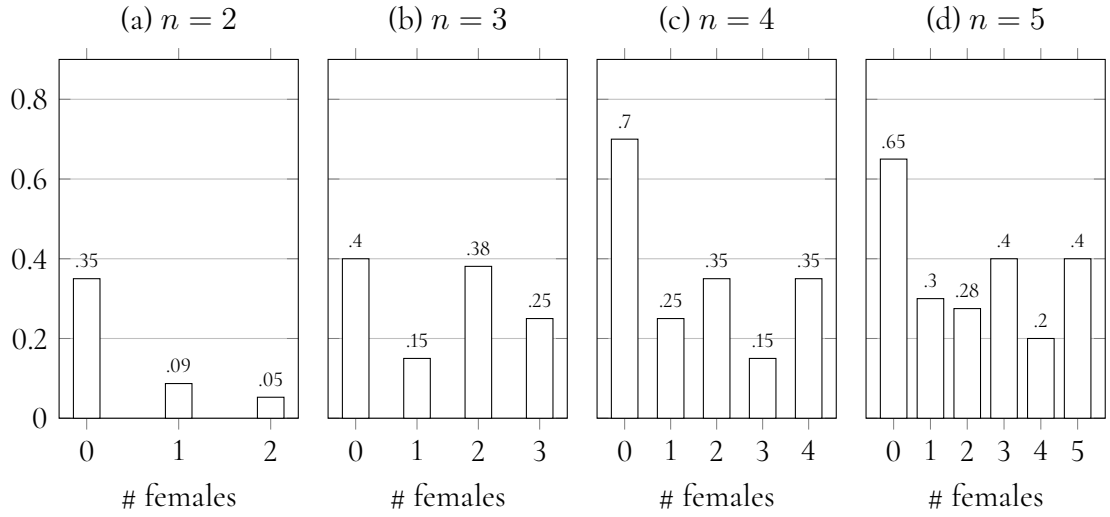
We now turn to the question how the lying behavior of groups is affected by the group gender composition. First evidence on this issue has been provided by Muehlheusser, Roider, and Wallmeier (2015), who find that all-male dyads lie significantly more than all-female dyads. The present study replicates this finding as can be seen in Figure 3(a), where the left-most and right-most bars correspond to all-male groups and all-female groups, respectively. In addition, the other panels of Figure 3 show that this finding is not specific to dyads, but holds across group sizes:

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

In each of the four panels of Figure 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

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

Figure 3: Frequency of lying by group size and number of females in the group



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.

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(a), this is also the case in our experiment. Moreover, recall from Section 2 the observation that female individuals tend to lie less than male individuals. One might thus hypothesize that, more generally, the number of females in a group has a (weakly) monotone effect on group dishonesty. However, panels (b)-(d) of Figure 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 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 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 female group members). While Figure 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 ($p = 0.368$).[27]

Our finding that the frequency of lying is higher in all-male groups compared to almost-all-male groups, raises the question of how, more generally, the number of females in a group affects lying. In fact, our next result 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(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, 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.

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.[28] 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

[27] 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 ($p = 0.338$).

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

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 1).[29]

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 4. As shown in panel (a), and not surprisingly, larger groups take 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 compared to dyads. Moreover, we find no evidence that almost-all-male groups exhibit longer group discussions. As illustrated in Figure 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 communication 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.

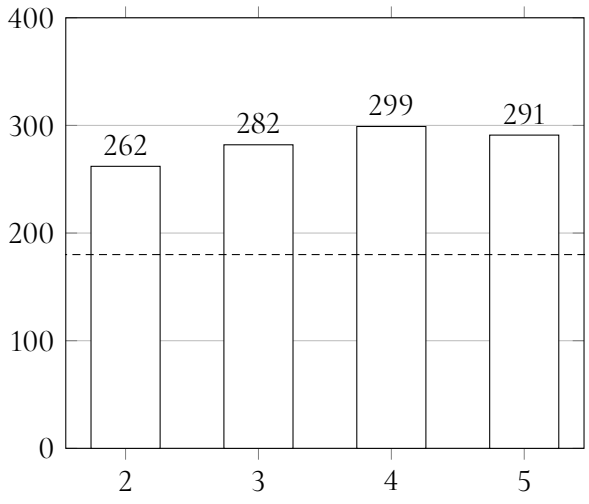
Overall, in almost-all-male groups the sole female's average share of total talking time in the group is *lower* than $1/n$ (i.e. the share that would result from identical talking times of all n group members). In particular, we examine the behavior in almost-all-male groups partitioned according to their group size ($n = 2, 3, 4, 5$) and whether or not they lied. In seven out of the eight resulting 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.[30]

[29] 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.

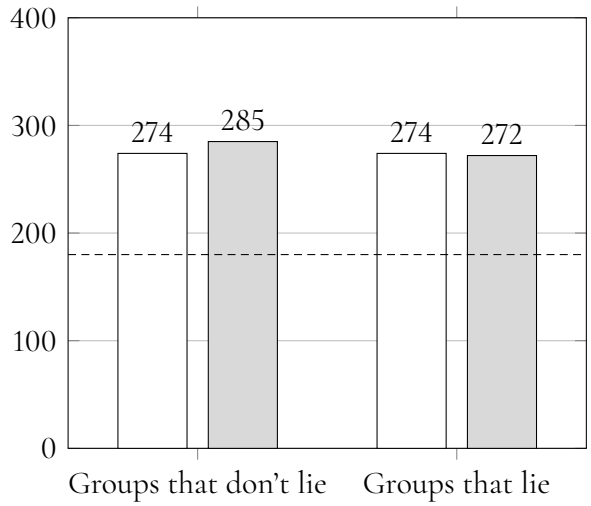
[30] In their experiment on team performance, Hardt, Mayer, and Rincke (2022) study groups of four. They find that, in gender-balanced teams, males talk significantly more than males assigned to all-male teams. Interestingly, females exhibit the opposite pattern, that is, they talk less in mixed teams 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

Figure 4: Groups' decision times (in seconds)

(a) By group size



(b) For all-male groups (white bars) and almost-all-male groups (gray bars) by lying decision



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 as elicited in a separate task. 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

cited therein.

the internet. For example, one question asks in which year the French composer Claude Debussy was born.[31] 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.[32] 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.

In our experiment, 31 percent of male subjects are cheaters compared 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 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.[33] 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(a) above).

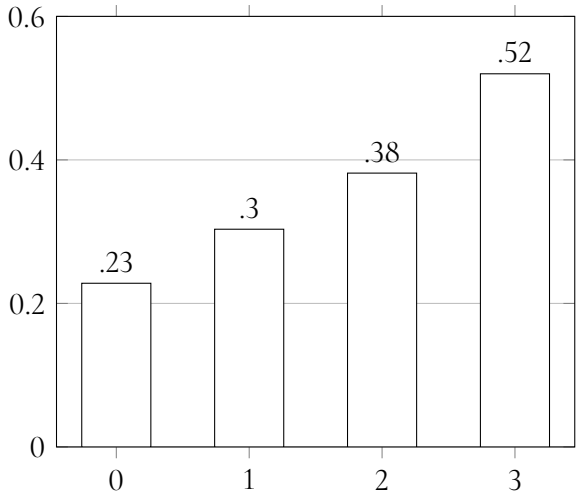
In a next step, we study how the *number* of dishonest group members affects the lying behavior of groups. In particular, we investigate whether group behavior is mainly affected by the presence

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

[32] 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).

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

Figure 5: Frequency of group lying by number of cheaters in the group



Notes: 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.

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 is increasing in the number of cheaters in the group.

The result is illustrated in Figure 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 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 5 (but they are included when testing). As shown in the regression analysis of Section 4.5, this result is robust when controlling for group size.

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.

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 that is zero otherwise). We again confine attention to those 363 groups that had an incentive to lie to their advantage (i.e. that observed 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 2, which displays the coefficients of our main variables of interest. Table 4 in Appendix A 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 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.[34] 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' extraversion, agreeableness, conscientiousness, emotional stability, and intellect/imagination.[35] 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, Wheelwright, Hill, Raste, and Plumb, 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 (for an example, see Appendix B). In

[34] 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 B.

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

Table 2: Robustness of main results

	(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

Notes: 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.

both, the Raven and the RMET tasks, 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 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.

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 in Appendix A) 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 3.1 above).[36] As explained in Section 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).

[36] 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.

These group types account for 40.5 percent out of the 385 groups (see Table 3). 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 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.

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, various 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) 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, our findings do not seem to be consistent with 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 might 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 gender-specific image concerns) could potentially be disentangled in a follow-up treatment in which group decisions are taken by simple majority. To illustrate this, 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 voting (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 out of fear for their reputation).

References

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

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.

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," *Review of Economic Studies*, 77, 1301–1328.

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.

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.

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. and J. Tirole (2006): "Incentives and prosocial behavior," *American Economic Review*, 96, 1652–1678.

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.

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.

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.

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., 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. and M. Sutter (2012): "Groups make better self-interested decisions," *Journal of Economic Perspectives*, 26, 157–176.

Chen, D., 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.

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. 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.

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.

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.

Dreber, A. and M. Johannesson (2008): "Gender differences in deception," *Economics Letters*, 99, 197–199.

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.*

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 behavior—What predicts morally questionable decisions?" *European Economic Review*, 149, 104264.

Fischbacher, U. and F. Föllmi-Heusi (2013): "Lies in disguise—an experimental study on cheating," *Journal of the European Economic Association*, 11, 525–547.

Gächter, S. and J. F. Schulz (2016): "Intrinsic honesty and the prevalence of rule violations across societies," *Nature*, 531, 496–499.

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.

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.

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, 262–290.

Hardt, D., L. Mayer, and J. Rincke (2022): "Who Does the Talking Here? The Impact of Gender Composition on Team Interactions," *CESifo Working Paper No. 10550*.

Houser, D., J. List, M. Piovesan, A. Samek, and J. Winter (2016): "Dishonesty: From parents to children," *European Economic Review*, 82, 242–254.

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.

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*.

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.

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.

Kroll (2016): *Global Fraud Report: Vulnerability on the Rise*, http://www.kroll.com/en-us/global-fraud-report.

Kugler, T., E. Kausel, and M. Kocher (2012): "Are groups more rational than individuals? A review of interactive decision making in groups," *Wiley Interdisciplinary Reviews: Cognitive Science*, 3, 471–482.

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.

Matsa, D. and A. Miller (2013): "A female style in corporate leadership? Evidence from quotas," *American Economic Journal: Applied Economics*, 5, 136–69.

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.

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.

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.

Sinnott-Armstrong, W. (1988): *Moral dilemmas*, Blackwell Publishers.

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.

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.

Weidmann, B. and D. J. Deming (2021): "Team players: How social skills improve team performance," *Econometrica*, 89, 2637–2657.

Appendix

A Additional figures and tables

Table 3: Number of group observations with die-roll outcome $\neq 5$ per treatment

(a) All groups

	No. of females in the 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 in the 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 4: Robustness of main results (all coefficients)

	(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

Notes: The table note for Table 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 for the 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.

B 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 18 and 19, 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 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 6), where Figure 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 8). Afterwards, groups were formed, and group members met for the first time on the page "Group video and audio test" (see Figure 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 10), where subjects read the instructions for Task 1, were shown the die roll, and reports were made.

Figure 6: Screenshot of the page “Individual video and audio test”

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 7: Screenshot of external test page

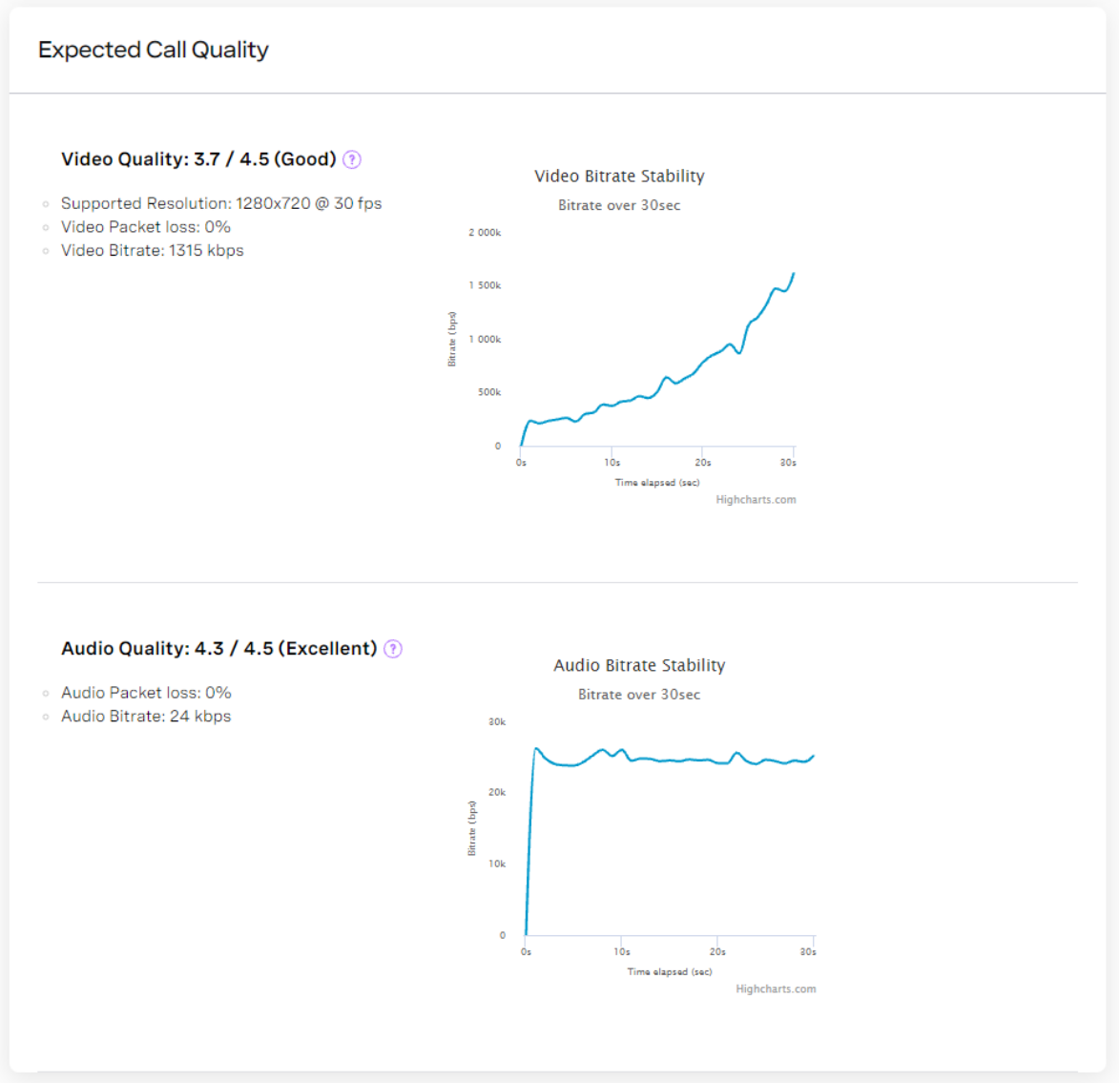


Figure 8: Screenshot of the page “Individual general overview”

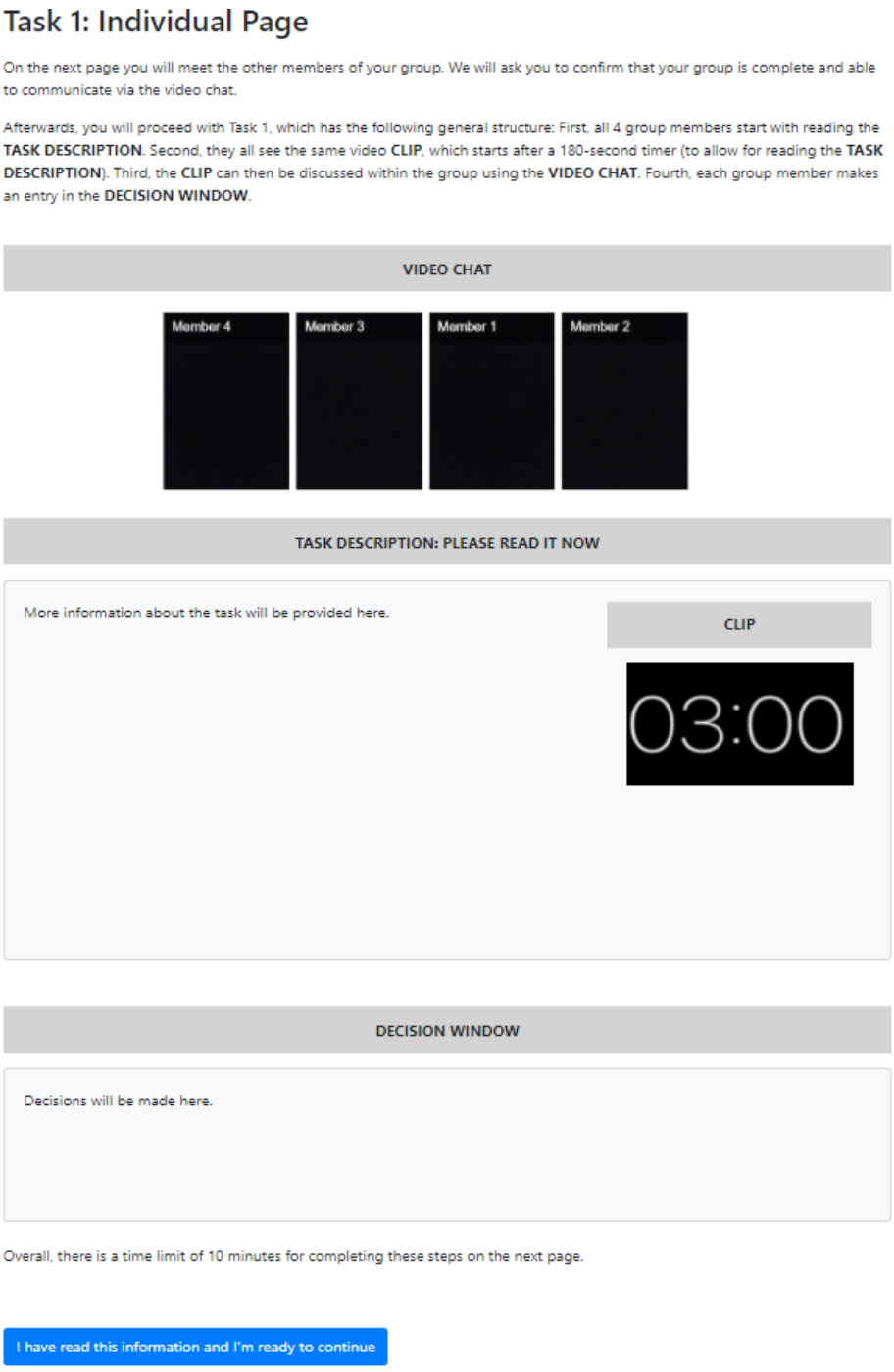


Figure 9: Screenshot of the page “Group video and audio test”

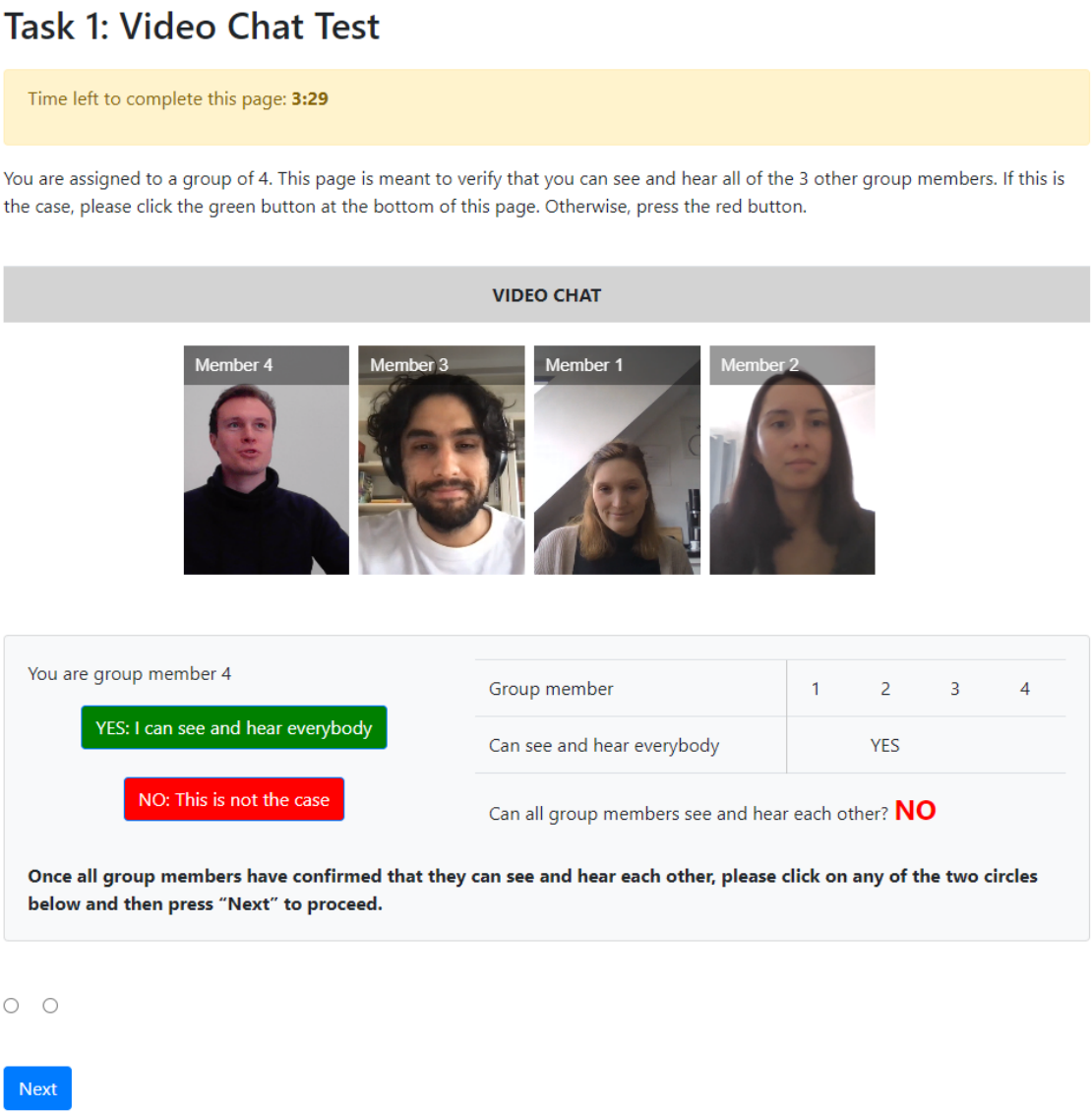
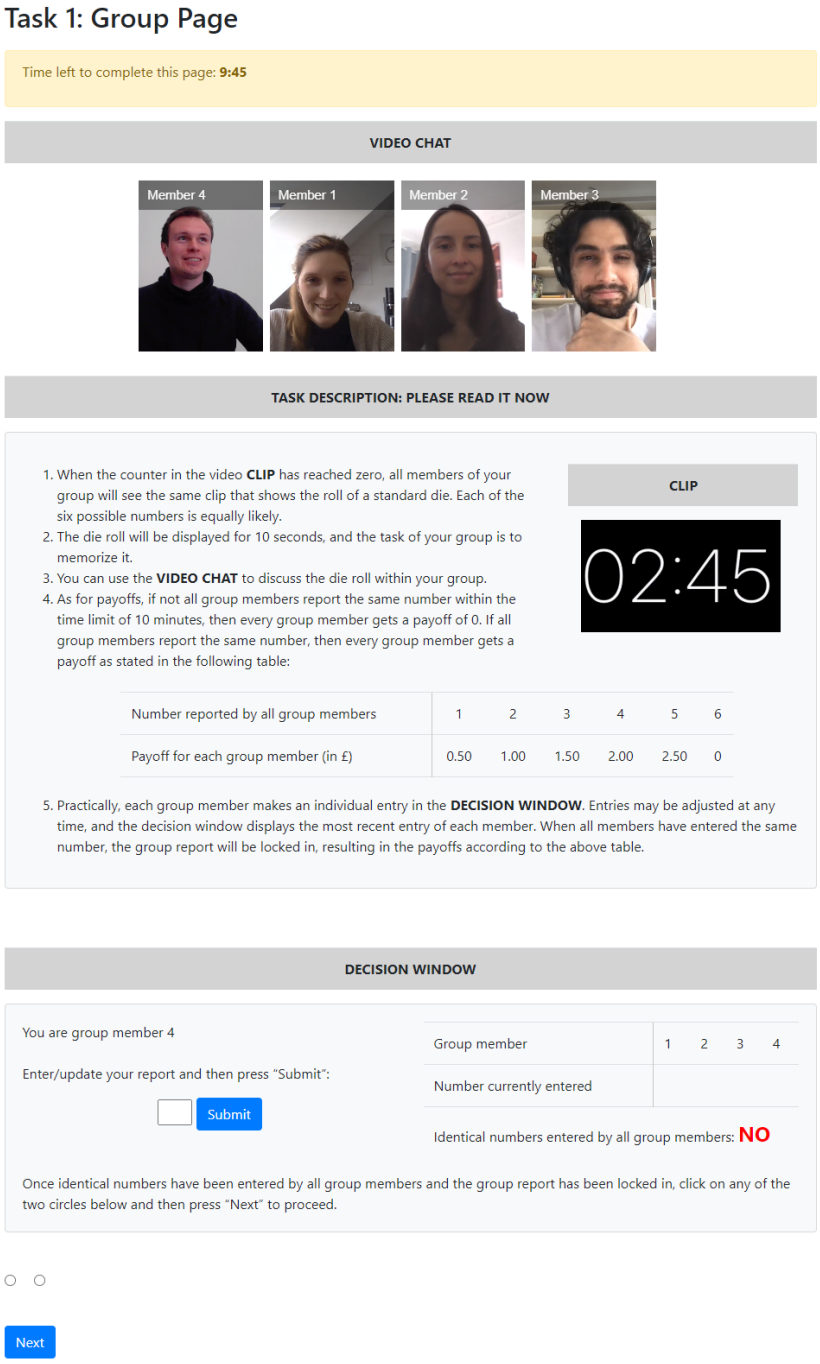


Figure 10: Screenshot of the page “Instructions and group decision”



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.

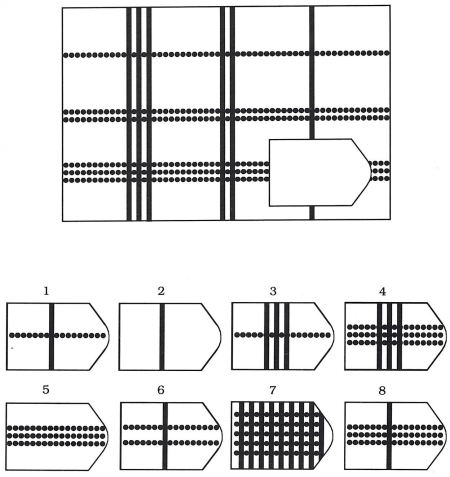
This task will take at most seven minutes, and you will receive a fixed payment of £1.50 for working on it.

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:



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.[37]

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:

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

- 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!