HONESTY AND MORAL BEHAVIOR IN ECONOMIC GAMES

EDVITED BY: Steffen Huck, Agne Kajackaite and Nora Szech
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-751-4
DOI 10.3389/978-2-88971-751-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

HONESTY AND MORAL BEHAVIOR IN ECONOMIC GAMES

Topic Editors:
Steffen Huck, Social Science Research Center Berlin, Germany
Agne Kajackaite, Social Science Research Center Berlin, Germany
Nora Szech, Karlsruhe Institute of Technology (KIT), Germany

Citation: Huck, S., Kajackaite, A., Szech, N., eds. (2021). Honesty and Moral Behavior in Economic Games. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-751-4

Table of Contents

frontiers in Psychology

EDITORIAL
published: 04 October 2021
doi: 10.3389/fpsyg.2021.769856

Editorial: Honesty and Moral Behavior in Economic Games

Steffen Huck[1,2], *Agne Kajackaite*[1]* *and Nora Szech*[3]

[1] Social Science Research Center Berlin, Berlin, Germany, [2] University College London, London, United Kingdom, [3] Chair of Political Economy, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Keywords: morality, lying, honesty, ethics, economic games, behavioral economics, experiments

Editorial on the Research Topic

Honesty and Moral Behavior in Economic Games

RESEARCH ON MORAL BEHAVIOR IN BEHAVIORAL ECONOMICS

It is in crisis times that we can see the ailments of society under a magnifying glass and all three major crises we have recently been facing and continue to face, the financial crisis of 2007 and its aftermath, the current Corona crisis, and the climate crisis that will only ever become worse, have put a spotlight on greedy and dishonest behavior which needs to be tackled if societies want to escape such ordeals half-way unscathed. Yet politicians can simply ignore key problems in their campaigns to get more votes; decision makers can get involved in corrupt behaviors for monetary benefits; and ordinary citizens can simply close their eyes trying to justify selfish acts –fueling crises further. It is, thus, not surprising that immoral behaviors and their root causes have received increasing attention in the last decade of the social science literature. In this collection we present 11 exciting new studies exploring the morality of behavior from the vantage point of (behavioral) economics.

From a standard economic perspective, the decision to behave immorally for a monetary benefit is affected by only two factors—the probability of being caught and the penalty resulting from it (see Becker, 1968). However, the fast-growing literature in behavioral economics shows that many people would forego an immoral action, such as lying, even if there is no possibility of being caught and being punished (see, for instance, Gneezy, 2005; Mazar et al., 2008; Shalvi et al., 2011; Fischbacher and Föllmi-Heusi, 2013; Abeler et al., 2014, 2019; Gächter and Schulz, 2016; Kajackaite and Gneezy, 2017; Gneezy et al., 2018). These studies show that some people lie only partially or do not lie at all, because they have an intrinsic cost of lying (a self-image cost) and/or because they do not want to be perceived as liars by others or themselves (image concerns). Another stream of research shows that moral behavior can be eroded in market interactions and voting (see, for instance, Falk and Szech, 2013; Bartling et al., 2015, Falk et al., 2020; Ziegler et al., 2020), and that the psychological cost of immoral behavior can be reduced by choosing to be ignorant about the consequences of one's, actions on others (see Dana et al., 2007; Exley, 2015; Grossman and van der Weele, 2017; Serra-Garcia and Szech, 2021).

These papers are just the tip of the iceberg—indeed morality has become one of the most popular topics in behavioral economics, and we are learning many valuable lessons about the forces influencing people's choices to behave in a more or less moral or honest way. With this Research

OPEN ACCESS

Edited and reviewed by:
Valerio Capraro,
Middlesex University, United Kingdom

***Correspondence:**
Agne Kajackaite
agne.kajackaite@wzb.eu

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *02 September 2021*
Accepted: *10 September 2021*
Published: *04 October 2021*

Citation:
Huck S, Kajackaite A and Szech N (2021) Editorial: Honesty and Moral Behavior in Economic Games. Front. Psychol. 12:769856. doi: 10.3389/fpsyg.2021.769856

Topic, we contribute to the literature by shedding more light on mechanisms that drive morally relevant behaviors.

THIS RESEARCH TOPIC

This Research Topic consists of 11 research papers, with each of them using lab or field experiments to answer their research questions. The content of the contributions, forming this special issue, ranges from contributions on lying behavior (contributions by Behnk and Reuben; Dunaiev and Khadjavi; Jacquemet et al.; Vorsatz et al.; Waeber), bribing (contributions by Balafoutas et al.; Wang and Chen), pro-sociality (contributions by Regner; Regner and Matthey), and discrimination (contribution by Feess et al.) up to an experiment aiming to reduce meat consumption (contribution by Haile et al.). Taking a wholistic viewpoint as in Bandura (2016), morally relevant behavior may include caring about nature, the environment, and animals as well. A reduction in meat consumption may help us tackle the climate crisis. More broadly, as in the current Corona crisis, fostering morally relevant behaviors will hopefully contribute to dealing with the fallout from major crises, ideally helping to overcome them successfully.

FUTURE DIRECTIONS

As demonstrated by this special issue, behavioral economics of morality is a fruitful field of research. While the topic is slowly maturing, the scope for future studies remains large with many important understudied applications. One of these is science itself including, as we were learning while writing this introduction, the very subfield this issue deals with. While it is tempting to dwell on the irony it is probably more interesting and clearly more important to understand the mechanisms that enable fraudulent behavior in the sciences. The case is complicated partially because of academia's self-government. When immoral behavior can only be verified by a select few, the question easily becomes who observes the observer? And what are the observer's interests? Of course, universities do not like scandals. But sweeping things under the carpet may ultimately be the more dangerous strategy. There is a lot to be worked on.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Abeler, J., Becker, A., and Falk, A. (2014). Representative evidence on lying costs. *J. Public Econ*. 113, 96–104. doi: 10.1016/j.jpubeco.2014.01.005

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica* 87, 1115–1153. doi: 10.3982/ECTA14673

Bandura, A. (2016). *Moral Disengagement: How People Do Harm and Live With Themselves*. New York, NY: Macmillan, 544.

Bartling, B., Weber, R., and Yao, L. (2015). Do markets erode social responsibility? *Q. J. Econ*. 130, 219–266. doi: 10.1093/qje/qju031

Becker, G. S. (1968). Crime and punishment: an economic approach. *J. Polit. Econ*. 76, 169–217. doi: 10.1086/259394

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econ. Theory* 33, 67–80. doi: 10.1007/s00199-006-0153-z

Exley, C. (2015). Excusing selfishness in charitable giving: the role of risk. *Rev. Econ. Stud*. 83, 587–628. doi: 10.1093/restud/rdv051

Falk, A., Neuber, T., and Szech, N. (2020). Diffusion of being pivotal and immoral outcomes. *Rev. Econ. Stud*. 87, 2205–2229. doi: 10.1093/restud/rdz064

Falk, A., and Szech, N. (2013). Morals and markets. *Science* 340, 707–711. doi: 10.1126/science.1231566

Fischbacher, U., and Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *J. Eur. Econ. Assoc*. 11, 525–547. doi: 10.1111/jeea.12014

Gächter, S., and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature* 531, 496–499. doi: 10.1038/nature17160

Gneezy, U. (2005). Deception: the role of consequences. *Am. Econ. Rev*. 95, 384–394. doi: 10.1257/0002828053828662

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *Am. Econ. Rev*. 108, 419–453. doi: 10.1257/aer.20161553

Grossman, Z., van der Weele (2017). Self-image and willful ignorance in social decisions. *J. Eur. Econ. Assoc*. 15, 173–217. doi: 10.1093/jeea/jvw001

Kajackaite, A., and Gneezy, U. (2017). Incentives and cheating. *Games Econ. Behav*. 102, 433–444. doi: 10.1016/j.geb.2017.01.015

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res*. 45, 633–644. doi: 10.1509/jmkr.45.6.633

Serra-Garcia, M., and Szech, N. (2021). The (In)Elasticity of moral ignorance. *Manag. Sci*. Available online at: https://www.cesifo.org/en/publikationen/2019/working-paper/inelasticity-moral-ignorance

Shalvi, S., Dana, J., Handgraaf, M. J. J., and De Dreu, C. K. W. (2011). Justified ethicality: observing desired counterfactuals modifies ethical perceptions and behavior. *Organ. Behav. Hum. Decis. Process*. 115, 181–190. doi: 10.1016/j.obhdp.2011.02.001

Ziegler, A., Romagnoli, G., and Offerman, T. (2020). Morals in multi-unit markets. *Tinbergen Institute Discussion Paper 2020-072/I*. Amsterdam. Available online at: https://ideas.repec.org/p/tin/wpaper/20200072.html

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 31 May 2021
doi: 10.3389/fpsyg.2021.668674

Changing Hearts and Plates: The Effect of Animal-Advocacy Pamphlets on Meat Consumption

Menbere Haile[1], *Andrew Jalil*[2], *Joshua Tasoff*[1]* *and Arturo Vargas Bustamante*[3]

[1] Department of Economic Sciences, Claremont Graduate University, Claremont, CA, United States, [2] Department of Economics, Occidental College, Los Angeles, CA, United States, [3] Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, United States

Social movements have driven large shifts in public attitudes and values, from anti-slavery to marriage equality. A central component of these movements is moral persuasion. We conduct a randomized-controlled trial of pro-vegan animal-welfare pamphlets at a college campus. We observe the effect on meat consumption using an individual-level panel data set of approximately 200,000 meals. Our baseline regression results, spanning two academic years, indicate that the pamphlet had no statistically significant long-term aggregate effects. However, as we disaggregate by gender and time, we find small statistically significant effects within the semester of the intervention: a 2.4 percentage-point reduction in poultry and fish for men and a 1.6 percentage-point reduction in beef for women. The effects disappear after 2 months. We merge food purchase data with survey responses to examine mechanisms. Those participants who (i) self-identified as vegetarian, (ii) reported thinking more about the treatment of animals or (iii) expressed a willingness to make big lifestyle changes reduced meat consumption during the semester of the intervention. Though we find significant effects on some subsamples in the short term, we can reject all but small treatment effects in the aggregate.

Keywords: vegan, animal advocacy, randomized controlled trial, pamphlets, leaflets

OPEN ACCESS

Edited by:
Steffen Huck,
Social Science Research Center Berlin, Germany

Reviewed by:
Manja Gärtner,
German Institute for Economic Research (DIW), Germany
Stefan Penczynski,
University of East Anglia, United Kingdom
Iris Vermeir,
Ghent University, Belgium

***Correspondence:**
Joshua Tasoff
joshua.tasoff@cgu.edu

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *16 February 2021*
Accepted: *28 April 2021*
Published: *31 May 2021*

Citation:
Haile M, Jalil A, Tasoff J and Vargas Bustamante A (2021) Changing Hearts and Plates: The Effect of Animal-Advocacy Pamphlets on Meat Consumption. Front. Psychol. 12:668674. doi: 10.3389/fpsyg.2021.668674

1. INTRODUCTION

During the twentieth century, animal farming radically transformed from small family farms to large-scale concentrated animal feeding operations (CAFO), also referred to as factory farms (Norwood and Lusk, 2011). According to USDA data, 99% of US farmed animals are raised in CAFOs (Anthis, 2019). Gains in efficiency have come at the expense of the welfare of the animals. For example, pigs are confined for months in crates measuring only 14 square feet, prohibiting virtually all movement including walking and turning around (Norwood and Lusk, 2011). Egg-laying hens are placed in cramped cages with only 67 square inches allotted per bird, less than one 8.5 × 11 inch sheet of paper (93.5 square inches). To prevent aggression in cramped quarters, a half to a third of their highly sensitive beaks are severed, possibly leading to chronic pain (Duncan, 2001; Fraser et al., 2001; Cheng, 2006; Norwood and Lusk, 2011). These practices are standard industry protocol in the U.S.

While measuring the animals' wellbeing directly is not possible, the evidence indicates that farmed animals suffer under these conditions. Confined pigs show signs of extreme stress, such as bar-biting and other repetitive behaviors. They also become unresponsive, remaining passive when splashed with water, poked, or prodded—a likely sign of severe depression (Broom and Johnson, 1993; Broom et al., 1995; Vieuille-Thomas et al., 1995; Marchant and Broom, 1996). The tight confinement imposed on egg-laying hens prevents exercise which leads to osteoporosis and broken bones. As many as 30% of hens have broken bones before slaughter. During forced molting, hens show signs of severe distress, including aggression and stereotyped pacing (Gregory and Wilkins, 1989; Duncan, 2001). Given the vast amount of meat consumption, the scale of the suffering is likely immense. Based on data from the FAO of the United Nations, approximately 70 billion land animals are slaughtered for food every year, with 74% raised in CAFOs (Sanders, 2018).

This treatment of farmed animals violates the principles of many ethical theories, including utilitarian and deontological frameworks (Singer, 1979; Regan, 1983). Amongst ethicists who write on the issue, "there is widespread (though not perfect) consensus that it is generally morally better for the typical North American to eat less factory farmed meat" (Schwitzgebel et al., 2020). This view is consistent with mainstream American attitudes toward factory farms. In a recent survey, based on a representative sample of the U.S. population, a ban on factory farming, slaughterhouses and animal farming garnered substantial support: 49, 47, and 33%, respectively (Anthis, 2017). Other researchers replicated these results using a different sample (Norwood and Murray, 2018). However, support of the system through meat consumption continues, perhaps for a variety of reasons including but not limited to ignorance or neglect of conditions on farms, a lack of perceived individual agency to effect change, the invisibility of the victims, and the challenge of changing one's habits. In that same study, 58% of the sample agreed that "most farmed animals are treated well," despite the fact that 99% of farmed animals in the U.S. are raised in CAFOs under the aforementioned conditions.

In this paper, we ask whether moral persuasion through pamphlets can lead to changes in behavior, specifically meat consumption. Moral persuasion, or moral suasion, is the use of normative appeals and rhetoric to affect behavior. It has been a centerpiece of many social movements in history. Harriet Beecher Stowe's *Uncle Tom's Cabin* was one of the bestselling books of its day and is widely believed to have changed attitudes against slavery prior to the American Civil War. Martin Luther King Jr., during the 1960's civil rights movement, advocated for a vision of America in which people would not be judged by their appearance but by their character. More recently, advocates for marriage equality have made their case across numerous platforms, using a strategic legal campaign through litigation and legislative advocacy. Increased social acceptability and positive portrayal of LGBT individuals persuaded millions to change their minds on marriage equality (Kowal, 2015). In each case, moral persuasion led to dramatic changes in legal institutions and social norms.

The animal advocacy movement offers unique advantages for investigation. First, it is ongoing with aims that are far from the status quo. Great scope for change remains. Second, the animal advocacy movement persuades individuals to alter their behavior, specifically to avoid eating meat. Behavior change with individual-level panel data allows for high-powered tests. Third, a common method of persuasion in this social movement is the dissemination of pamphlets to pedestrians. This medium affords us the ability to randomly assign moral persuasion at the individual level.

Studying the effect of moral-advocacy pamphlets specifically is interesting in its own right. Pamphlets have been a historically important medium for advocacy. Martin Luther's *95 Theses* instigated the Protestant Reformation, Thomas Paine's pamphlet *Common Sense* popularized the argument for the American Revolution, and Martin Luther King Jr's *Letter from a Birmingham Jail* advanced the campaign for civil rights (Forman, 2017). Pamphlets are also inexpensive to produce and disseminate, and their physicality may capture more attention than some digital mediums.

We set up a table at a college campus and verbally solicited participation. Undergraduate student subjects were randomly assigned to an animal-advocacy or placebo pamphlet. The animal-advocacy pamphlet specifically requested that people refrain from eating meat to improve the wellbeing of farmed animals. The placebo pamphlet made no mention of diet. We estimated the effect of the treatment pamphlet on meal purchases in the college's main dining halls, with an individual-level data set of over 200,000 food purchases. We supplemented the purchase data with a follow-up survey 1 month after disseminating the pamphlet.

Several studies have attempted to measure the effect of animal-welfare pamphlets on meat consumption but all have used self-reports and have been under powered (Animal Charity Evaluators, 2013; Hennessy, 2016; Flens et al., 2018). See Animal Charity Evaluators (2017) and Peacock and Sethu (2017) for reviews. Our paper adds to the growing literature that uses actual consumption data to measure the effect of an intervention designed to reduce meat consumption. To our knowledge, there are only three previous randomized-controlled trials in this area with real meat consumption data in the field. One examines the effect of defaults (Hansen et al., 2019) and the other two examine the effect of education (Jalil et al., 2020; Schwitzgebel et al., 2020). Three other studies conduct field experiments with exogenous variation to estimate the effect of menu manipulation and product placement (Garnett et al., 2019, 2020; Vandenbroele et al., 2019).

Our pamphlet uses an animal welfare message to persuade individuals to reduce their meat consumption. The extent to which such messages are effective is unclear. Many studies conclude that most people do not want to harm sentient beings, but engage in cognitive dissonance. Rothgerber (2020) develops a psychological framework, termed "meat-related cognitive dissonance," for how individuals evade guilt when their food choices lead to animal harm (e.g., avoiding information, belittling "do-gooders," denigrating the animals, formulating pro-meat justifications, rejecting responsibility). Based on survey-level

evidence, Bastian et al. (2012) find that when told of the suffering animals experience in meat production, people ascribe lower mental attributes to those animals, likely to justify their continued participation in the system. Schröder and McEachern (2004) find that while people generally agree that cruelty toward animals is intolerable, they develop strategies to cope with the harm animals experience due to their demand for meat. By contrast, in a recent review of the literature, Bianchi et al. (2018) conclude that interventions focusing on animal welfare are associated with intentions to consume less meat. Sonoda et al. (2018) provide evidence that consumers care about animal welfare considerations in their food purchases. Schwitzgebel et al. (2020) find that college students, in response to a class on the ethics of eating meat that talks about animal suffering, reduce their actual meat consumption.

We find in the aggregate, looking at the data over a 2-year time period, no statistically significant effects of the treatment pamphlet. We can reject treatment effects of 1.9 percentage points or larger with 95% confidence. Likewise, when we look at treatment effects in the semester of the intervention and the subsequent semester we find no significant effects. It is only when we disaggregate the effect by time and gender, as we specified in our pre-analysis plan, we find statistically significant effects. Men and women change their diets during the semester of the intervention. Men reduce their consumption of poultry and fish by 2.4 percentage points (5.2%) and increase their consumption of vegetarian and vegan meals by 2.3 percentage points (10.6%), roughly the same magnitude. Women decrease their consumption of beef by 1.5 percentage points (13.3%), but weakly increase their consumption of poultry and fish (ns). Though we expected differences by gender, as we found nearly twice the treatment effect from women compared to men in our previous study (Jalil et al., 2020), we did not expect this pattern, nor do we have a good explanation for it. In the long run, the effects are no longer statistically significant.

Our survey data helps to provide additional insight regarding the mechanisms of the intervention. Those participants who (1) self-identified as vegetarian, (2) reported thinking more about the treatment of animals or (3) expressed a willingness to make big lifestyle changes reduced meat consumption during the semester of the intervention. Together, the evidence suggests that the treatment is more effective for those already predisposed toward meat reduction.

Overall, the effects are small, short-lasting, or non-existent. We test for treatment effects with many other subsamples and only find null results. This is presented in our online **Supplementary Information** document. Given (1) the unexpected and unexplained gender differences, (2) the overall weak treatment effects, and (3) multiple hypothesis testing with mostly null results, we do not have high confidence that a replication study would produce the same pattern of significant findings. We do think that the significant results are still informative when properly contextualized within the larger literature.

On the flip side, pamphleteering is an inexpensive intervention. Given the low costs, the evidence is also insufficient to claim that pamphleteering is cost ineffective. Even very small effect sizes may justify pamphleteering if the cost of disseminating a pamphlet is miniscule.

2. MATERIALS AND METHODS

2.1. Experimental Procedures

We recruited undergraduate students at a U.S. college campus. Experimenters positioned themselves at various locations on campus at times with heavy foot traffic, and asked students to participate in a scientific study that involved receiving a pamphlet, being contacted by email for an online survey, and having a chance to win a gift card. Experimenters gave students a short description of the study and a consent form. No mention of meat consumption or animal welfare occurred before consenting. After reading and signing the consent forms, the students received either the treatment or control pamphlet. The experimenters did not discuss the contents of the pamphlet or give further information. In the consent form, in addition to their name and signature, students provided an email address for future contact. In total, 685 students participated. The pamphleteering was conducted over 2-week segments about 1 month after the start of the spring and fall semesters of 2019.

Approximately 1 month after subjects received their pamphlet, we emailed subjects a link to an online Qualtrics survey that took 3–5 min to complete. We incentivized participation in the survey through a random drawing for a $50 Amazon gift card. The participation rate was 49% with 338 subjects completing the survey. People who eat less meat were more likely to select into the survey (please see **Supplementary Information** for details).

We individually randomized subjects into treatment and control groups based on their student ID. The treatment group received the animal-advocacy pamphlet *Compassionate Choices*, produced by an activist group, Vegan Outreach. The pamphlet discusses the impact of factory farming and the conditions under which farm animals are treated. The pamphlet also contains information on how to eat less meat, i.e., discussions about the health benefits of eating a plant-based diet, meal ideas that contain no animal products, and personal testimonies from people who have made the choice to adopt a vegetarian lifestyle. While this latter information could also influence behavior, the majority of the pamphlet—its salient message—focuses on animal welfare. The barriers to diet change likely include lack of knowledge about the welfare condition of animals in farms, lack of an emotional connection to the suffering of animals, lack of knowledge about health and plant-based diets, and lack of knowledge about easily available plant-based options. The pamphlet attempts to address all of these issues. However, there are likely other barriers to diet change that the pamphlet simply cannot address, such as a long-ingrained habit of meat consumption.

The control group received the pamphlet *The Cruelty Behind the Cuteness*, a pamphlet produced by the Humane Society of the United States. It discusses problems with "puppy mills." It does not mention diet.

2.2. Data

We collected three types of data: food-purchase data, post-intervention Qualtrics survey data, and administrative data on gender. Students swiped their ID card, via their meal plan, to purchase food at the dining facilities. Cashiers chose one of four buttons that register the main entree: beef, poultry, fish, and veg. Vegetarian and vegan ("veg") meals were always available at every food station, offering students a choice between a plant-based and meat-based dish. Meals were a la carte, allowing us to observe students' food choices. The prices for the meat and veg options were usually the same. We exclude snacks and purchases where cashiers did not differentiate between meat and non-meat options (i.e., Friday evenings, weekday mornings and weekends). We collected data for four consecutive semesters totaling roughly 200,000 meals.

The survey data came from our online Qualtrics survey conducted a month after the intervention for both treatment and control groups. The survey questions asked about participant's demographic information, self-identified current diet, memory recall on the pamphlets, views toward treatment of farm animals, impact of personal choice, attempted diet change, reasons for changing, willingness to make big lifestyle changes, etc. The full details are in the **Supplementary Information**. We collected administrative data on gender from the card office for study participants. We have registered a research protocol containing the pre-analysis plan for this experiment at the AEA RCT registry with ID AEARCTR-0003871. Our pre-analysis plan is publicly available at www.socialscienceregistry.org/trials/3871.

3. RESULTS

3.1. Main Treatment Effects

We set up a booth at a U.S. college campus and recruited passing students to participate in an experiment. We randomly assigned subjects to either a placebo or treatment pamphlet. The placebo pamphlet discussed pet adoption and problems with puppy mills. The treatment pamphlet described the conditions of animals at factory farms and made an explicit call to action to adopt a vegan diet. The data set used in our analysis contains approximately 200,000 meal purchases from 685 students. For all subjects, we observe a baseline period prior to the intervention and a post-intervention period, allowing us to estimate within-person changes in diet. Recruitment was ongoing, leading to exogenous variation in the timing of the intervention and helping to control for any seasonal or calendar effects.

Our empirical strategy is to regress food choice on a treated indicator. We categorize all the food items into "beef," "poultry/fish," and "vegetarian" (see Methods). The omitted category is "salad bar," in which students choose from primarily vegetable options, but meat options are also present. Cashiers do not distinguish between salads with and without meat. We omit this category because its contents are ambiguous, however we retain "salad bar" observations in our analysis. We define a fourth category "meat" as containing "beef" or "poultry/fish." To estimate the average treatment effects of the intervention, we use a difference-in-difference framework. Specifically, we estimate the following logit regression:

$$\log\left(\frac{F_{m,i,d,h}}{1-F_{m,i,d,h}}\right) = \alpha + \beta_0 T_m + \beta_1 A_m + \beta_2 T_m A_m + \rho_i + \gamma_d + \delta_h + \varepsilon_{m,i,d,h} \quad (1)$$

where F is one of four food indicator variables that equals one if the meal purchase, m, belongs to that category (beef, poultry/fish, meat, or veg) for individual i, on day d, at hour h. T_m is an indicator variable for meals purchased by an individual in the treatment group, and A_m is an indicator variable for meals purchased after receiving the pamphlet. The key variable of interest is the interaction term, $T_m * A_m$, which measures the change in the food outcome variable after receiving the pamphlet for the treatment group, relative to the control group. The interaction term estimates the effects of the pamphlet on participants' food choices. We control for individual (ρ_i), date (γ_d), and hour (δ_h) fixed effects, and cluster standard errors at the individual level. We display all results as average marginal effects.

Table 1 displays the aggregate treatment effect of the animal-advocacy pamphlet. Columns (1)–(4) display the treatment effect over the full sample period 20 August 2018 to 2 June 2020. None of the coefficients are statistically different from zero. In Columns (5)–(8), we show the treatment effect during the semester of the intervention and in Columns (9)–(12) we show the treatment effect in the semesters after the intervention (they include meals purchased before the intervention, and meals purchased after the semester of the intervention). There is no statistically significant effect of the treatment on food choice for any of our outcomes and in any of our time windows.

Past research has shown that men and women respond to interventions aimed at diet-change in different ways (Jalil et al., 2020). As we specified in our pre-analysis plan, we estimate the treatment effects disaggregated on men and women in **Table 2**. We interact the treated indicator with a gender indicator and display the treatment effect by gender. In Columns (1)–(4), during the semester of the intervention, men significantly decrease their consumption of poultry or fish by 2.4 percentage points (5.2%) and increase their consumption of vegetarian/vegan meals by roughly the same magnitude, 2.3 percentage points (10.6%), suggesting substitution from meat to vegetarian/vegan meals. Overall, meat consumption for men falls by the same magnitude as the decline in poultry/fish, 2.4 percentage points (3.6%). Women, in contrast, significantly reduce beef consumption by 1.5 percentage points (12.5%). Poultry and fish consumption increases, though insignificantly, which explains why overall meat consumption does not fall for women. This finding suggests substitution from red meat (beef) to poultry/fish for women. It also explains the lack of detectable effects in **Table 1**, which does not disaggregate by gender. In Cols (5)–(8) of **Table 2**, in the semesters following the intervention, none of the effects remain statistically significant.

We find statistically significant effects by gender within the semester of the intervention, but not afterwards. **Table 3** examines the time path of this effect more closely by breaking apart the treatment effect into three time-windows: the month after the intervention, the second month after the intervention,

TABLE 1 | Main effect of pamphlet on food consumption.

	All observations				Semester of intervention				Semester after intervention			
	(1) Beef	(2) Poultry/fish	(3) Veg	(4) Meat	(5) Beef	(6) Poultry/fish	(7) Veg	(8) Meat	(9) Beef	(10) Poultry/fish	(11) Veg	(12) Meat
Treated	−0.002	−0.001	0.003	−0.003	−0.006	−0.004	0.007	−0.010	−0.001	0.002	0.000	0.001
	(0.005)	(0.008)	(0.007)	(0.008)	(0.005)	(0.008)	(0.008)	(0.008)	(0.006)	(0.010)	(0.010)	(0.011)
Mean of DV	0.161	0.394	0.279	0.556	0.158	0.397	0.279	0.555	0.164	0.393	0.276	0.557
PseudoR2	0.116	0.105	0.134	0.151	0.12	0.11	0.137	0.155	0.115	0.101	0.133	0.15
Clusters	681	685	686	685	676	685	686	685	677	684	685	685
N	199,716	199,756	199,963	199,835	126,780	127,209	127,394	127,292	156,727	156,962	157,087	157,008

The coefficents measure logit average marginal treatment effects with standard errors clustered at the individual level for all observations, the semester of intervention and after. Mean DV: mean of the dependent variable. All columns control for individual, date, and hour fixed effects.

and later. All columns use the same pre-intervention period, i.e., all meals purchased before the pamphleteering, but restrict the post-intervention period to different windows: the first month after the intervention in Cols (1)–(4), the second month after the intervention in Cols (5)–(8), and afterwards in Cols (9)–(12). The results show no significant effects in the first month after the intervention. Instead, the reductions in beef and poultry/fish for men and women, respectively, are statistically significant in the second month after the intervention. While those coefficients are negative in the first month, they are larger in magnitude and only become significant in the second month.

3.2. Heterogeneous Treatment Effects Using Survey Data

Our survey questions help to uncover the mechanisms behind the intervention. We first explore whether the intervention had heterogenous effects as a function of diet. Self-identified vegetarians may have already wrestled with the ethical issues of meat consumption in the past and been more receptive to accept the message. We find that self-reported vegetarians actually purchase a non-negligible fraction of their meals as meat: approximately 17.5% for untreated observations (the control group and pre-intervention treatment group). This finding suggests that while self-identified vegetarians may strive to reduce their meat consumption, they may not be successful at eliminating it from their diet.

We test whether the pamphlet had a significant effect on the fraction of individuals who identify as vegetarian. One month after receiving the pamphlet, the survey asked participants to self-report their diet. We find no significant difference between the control (17.8%) and treatment (12.1%) conditions in the fraction of survey takers who report being vegetarian (see **Supplementary Information** for details). Because the randomization of individuals into control and treatment groups should have led to roughly equal percentages of vegetarians in both groups pre-pamphleteering, this finding of similar percentages post-pamphleteering suggests that the pamphlet did not cause a significant increase in self-identified vegetarianism in the treatment group. However, in **Table 4**, we interact an indicator variable for self-reported vegetarians with the treated indicator to estimate heterogenous treatment effects. **Table 4** shows that self-reported vegetarians strongly reduce their poultry/fish consumption in the first month by 13.1 percentage points—effectively reducing their consumption of poultry/fish nearly to zero. Their overall meat consumption also decreases by 9.9 percentage points (56.5%), though it is not statistically significant. Over longer time windows the treatment effect on poultry/fish becomes non-significant.

Our survey provides additional variables about the mechanism of action. Interacting the treatment with these variables can help reveal the role of various mechanisms. The full analysis is in the **Supplementary Information**. Here we report the significant findings. We find one mechanism variable that predicts lower treatment effects. A survey question asked, "Reading the leaflet(s) taught me about (choose all the reasons that apply)." *Taughtme* takes the value one if the subject clicked

TABLE 2 | Effect of pamphlets by gender.

	Semester of intervention				Semester after intervention			
	(1) Beef	(2) Poultry/fish	(3) Veg	(4) Meat	(5) Beef	(6) Poultry/fish	(7) Veg	(8) Meat
Treated	0.003	–0.024**	0.023*	–0.024*	–0.003	–0.015	0.020	–0.020
	(0.008)	(0.011)	(0.012)	(0.013)	(0.009)	(0.015)	(0.016)	(0.018)
Treated × female	–0.019*	0.037**	–0.025*	0.023	0.004	0.029	–0.029	0.034
	(0.010)	(0.016)	(0.014)	(0.016)	(0.013)	(0.020)	(0.019)	(0.021)
P-value female	0.040	0.235	0.864	0.959	0.932	0.262	0.432	0.282
Mean of untreated DV female = 0	0.208	0.462	0.217	0.67	0.218	0.456	0.214	0.674
Mean of untreated DV female = 1	0.12	0.349	0.325	0.47	0.125	0.347	0.321	0.472
PseudoR2	0.12	0.111	0.137	0.155	0.115	0.102	0.133	0.15
Clusters	676	685	686	685	677	684	685	685
N	126,780	127,209	127,394	127,292	156,727	156,962	157,087	157,008

*The coefficents measure logit average marginal treatment effects with standard errors clustered at the individual level for the semester of intervention and after the semester of intervention. Mean DV: mean of the dependent variable. All columns control for individual, date, and hour fixed effects. *p < 0.1; **p < 0.05; ***p < 0.01.*

the radio button, "The treatment of animals in farms." In **Table 5**, we interact this indicator with the treated indicator. Those who stated that the pamphlet taught them about the mistreatment of farm animals exhibited no change in meat consumption while those who stated that they were not taught by the pamphlet significantly decreased their meat consumption by 3.6 percentage points (7.2%) in the first month. The effect is no longer significant after the first month.

We find two variables that are significantly associated with the treatment effect: (1) thought more and (2) willing to make big lifestyle changes. A survey question stated, "After reading the leaflet I thought more about (choose all the reasons that apply)," and if the subject checked, "The treatment of animals in farms," then we code the indicator *thoughtmore* as one, otherwise it is zero. This variable is designed to identify those individuals who read the pamphlet and reported thinking more about the treatment of farm animals. Another survey question asked, "How willing are you to make lifestyle changes to help reduce mistreatment of farm animals?" The options were, "Not willing to make any lifestyle changes," "Willing to make small lifestyle changes," "Willing to make moderate lifestyle changes," and "Willing to make big lifestyle changes." This question identifies the degree to which individuals report a willingness to change their behavior.

Table 5 shows that subjects who indicated thinking more about the treatment of farm animals significantly decreased their consumption of meat. In the second month, the decrease in meat consumption is 4.3 percentage points (8.8%) for this group and significant at the $p = 0.034$. Turning to the "willingness for change question," **Table 5** shows that those who state they are willing to make a big change significantly decreased their meat consumption in the second month by 11.2 percentage points (70.9%) ($p = 0.027$). Again the effect occurs not immediately, but rather in the second month.

These results should be interpreted with caution. In the **Supplementary Tables A.5–A.17**, we test 17 mechanism variables, including the variables in **Table 5**, for heterogeneous treatment effects each over 3 time windows. Limiting ourselves to only the meat outcome yields 51 tests. By luck, some of these tests are expected to be significant. To correct for multiple hypothesis tests, we compute sharpened False Discovery Rate (FDR) *q*-values. These can be interpreted as *p*-values corrected for multiple-hypothesis testing. We use the method of Benjamini et al. (2006) as presented in Anderson (2008). None of the *q*-values are below 0.1, suggesting that the mechanism results are unlikely to replicate.

4. DISCUSSION

The results show that the animal-advocacy pamphlets had no detectable aggregate effects in the short or long term. We are able to reject treatment effects of reducing meat in the first semester by 2.6 percentage points or larger ($CI = [-0.026, 0.006]$), in the second semester by 2.1 percentage points or larger ($CI = [-0.021, 0.023]$), and over both semesters by 1.9 percentage points or larger ($CI = [-0.019, 0.013]$), with 95% confidence. Moreover, the method of distributing the pamphlet (i.e., asking participants to sign a consent form and then weeks later, complete a survey) may have led to greater engagement with the pamphlet than what would have occurred outside of the setting of a study. As such, the effects we observe may be larger than the true effects in a real-world context. We can reject all but small treatment effects in the aggregate.

Disaggregating by time and gender, we find that men significantly reduce their poultry and fish consumption and women significantly reduce their beef consumption but only during the semester of the intervention. Whereas, men reduced overall meat consumption by switching from poultry and fish toward vegetarian and vegan meals, women appear to have switched from beef toward poultry and fish, suggesting a

TABLE 3 | Effect of pamphlets by gender–first month, second month, after.

	First month				Second month				After 2 months			
	(1) beef	(2) Poultry/fish	(3) Veg	(4) Meat	(5) Beef	(6) Poultry/fish	(7) Veg	(8) Meat	(9) Beef	(10) Poultry/fish	(11) Veg	(12) Meat
Treated	–0.004	–0.018	0.024	–0.025	0.014	–0.026**	0.018	–0.014	–0.003	–0.015	0.018	–0.021
	(0.009)	(0.014)	(0.015)	(0.016)	(0.009)	(0.013)	(0.014)	(0.014)	(0.008)	(0.014)	(0.015)	(0.017)
Treated × female	–0.005	0.023	-0.022	0.024	–0.032***	0.045**	-0.022	0.014	0.001	0.032*	-0.028	0.035*
	(0.013)	(0.018)	(0.017)	(0.019)	(0.012)	(0.019)	(0.017)	(0.019)	(0.012)	(0.019)	(0.018)	(0.020)
P-value female	0.322	0.662	0.881	0.895	0.045	0.180	0.739	0.984	0.827	0.168	0.406	0.249
Mean of untreated DV female = 0	0.212	0.459	0.214	0.671	0.21	0.461	0.216	0.671	0.217	0.457	0.215	0.675
Mean of untreated DV female = 1	0.124	0.346	0.326	0.47	0.119	0.352	0.321	0.471	0.125	0.347	0.321	0.472
PseudoR2	0.121	0.109	0.138	0.154	0.12	0.108	0.135	0.154	0.115	0.102	0.134	0.15
Clusters	675	685	685	685	671	685	685	684	677	685	685	685
N	106,331	106,847	106,970	106,892	104,071	104,629	104,757	104,677	159,585	159,856	159,947	159,869

*The coefficients measure logit average marginal treatment effects with standard errors clustered at the individual level. The estimation is split in to three periods after the intervention; first month, second month, and after 2 months. Mean DV: mean of the dependent variable. All columns control for individual, date, and hour fixed effects. *p < 0.1; **p < 0.05; ***p < 0.01.*

TABLE 4 | Heterogeneous effects by veg identification.

	First month				Second month				After 2 months			
	(1) Beef	(2) Poultry/fish	(3) Veg	(4) Meat	(5) Beef	(6) Poultry/fish	(7) Veg	(8) Meat	(9) Beef	(10) Poultry/fish	(11) Veg	(12) Meat
Treated	–0.007	0.004	0.013	–0.001	–0.007	0.009	–0.007	0.003	–0.008	–0.004	0.000	–0.012
	(0.010)	(0.013)	(0.014)	(0.014)	(0.009)	(0.014)	(0.014)	(0.015)	(0.008)	(0.014)	(0.015)	(0.016)
Treated × vegn	0.006	–0.135**	0.049	–0.098	0.001	–0.002	0.022	–0.001	0.025	0.088	–0.004	0.082
	(0.040)	(0.068)	(0.040)	(0.061)	(0.048)	(0.079)	(0.038)	(0.061)	(0.035)	(0.087)	(0.041)	(0.061)
P-value vegn	0.984	0.051	0.095	0.095	0.893	0.928	0.673	0.977	0.603	0.331	0.913	0.238
Mean of untreated DV vegn = 0	0.168	0.418	0.252	0.586	0.165	0.423	0.249	0.587	0.172	0.418	0.254	0.59
Mean of untreated DV vegn = 1	0.0515	0.124	0.585	0.175	0.0505	0.125	0.584	0.175	0.0491	0.119	0.591	0.168
PseudoR2	0.13	0.124	0.154	0.167	0.131	0.122	0.154	0.167	0.126	0.117	0.154	0.167
Clusters	336	342	342	343	335	342	342	342	339	342	342	343
N	51,830	52,174	52,203	52,237	51,233	51,540	51,581	51,595	77,377	77,633	77,579	77,661

*The coefficents measure logit average marginal treatment effects with standard errors clustered at the individual level. Vegn refers to self - reported vegetarians. The estimation is split in to three periods after the intervention; first month, second month, and after 2 months. Mean DV: mean of the dependent variable. All columns control for individual, date, and hour fixed effects. *p < 0.1; **p < 0.05; ***p < 0.01.*

TABLE 5 | Heterogenous effects by survey measures: "taught me," "thought more," "willingness to make a change."

	First month	Second month	After 2 months
	(1) Meat	(2) Meat	(3) Meat
Treated	–0.036**	0.010	–0.001
	(0.017)	(0.021)	(0.024)
Treated × taughtme	0.051	–0.026	–0.026
	(0.033)	(0.031)	(0.041)
P-value taughtme	0.602	0.471	0.441
Mean of untreated DV taughtme = 0	0.499	0.502	0.497
Mean of untreated DV taughtme = 1	0.524	0.527	0.53
PseudoR2	0.196	0.195	0.188
Clusters	209	208	209
N	31,922	31,841	47,386
Treated	–0.015	0.042*	0.012
	(0.018)	(0.022)	(0.026)
Treated × thoughtmore	–0.013	–0.084***	–0.061
	(0.030)	(0.030)	(0.042)
P-value thoughtmore	0.250	0.034	0.135
Mean of untreated DV thoughtmore = 0	0.53	0.531	0.522
Mean of untreated DV thoughtmore = 1	0.486	0.489	0.498
PseudoR2	0.192	0.191	0.185
Clusters	214	213	214
N	32,860	32,743	48,456
Treated	–0.053	0.120	0.075*
	(0.120)	(0.089)	(0.039)
Treated × smallchange	0.066	–0.124	–0.094**
	(0.116)	(0.099)	(0.047)
Treated × moderatechange	0.035	–0.113	-0.058
	(0.119)	(0.098)	(0.045)
Treated × bigchange	–0.075	–0.232**	–0.155
	(0.167)	(0.099)	(0.113)
P-value smallchange	0.437	0.865	0.452
P-value moderatechange	0.359	0.739	0.434
P-value bigchange	0.273	0.027	0.462
Mean of untreated DV nochange	0.584	0.583	0.587
Mean of untreated DV smallchange	0.638	0.638	0.633
Mean of untreated DV moderatechange	0.488	0.487	0.494
Mean of untreated DV bigchange	0.148	0.158	0.163
PseudoR2	0.169	0.168	0.166
Clusters	334	333	334
N	50,694	50,121	75,505

*The coefficents measure logit average marginal treatment effects with standard errors clustered at the individual level. The variable meat = beef + poultry + fish purchases. Mechanisms: Leaflet taught me about treatment of animals in farms, I thought more about treatment of animals in farms and willigness for change. The estimation is split in to three periods after the intervention; first month, second month, and after 2 months. Mean DV: mean of the dependent variable. All columns control for individual, date, and hour fixed effects. *p < 0.1; **p < 0.05; ***p < 0.01.*

substitution effect in terms of the types of meat purchased[1]. Past research shows that interventions designed to reduce meat consumption may have different effects on men and women. Jalil et al. (2020) find that a classroom intervention on the climate-change impact of food choices reduced meat consumption nearly twice as much for women as for men. We again find that men and women react to the intervention in different ways. Given that men and women have dramatically different baseline diets, it is natural that treatment effects differ between the two groups. In our study, before the intervention, men consume approximately 75% more beef, 30% more poultry/fish, 35% less vegetarian meals, and 40% more total meat than women. Ex ante, we expected to again see larger treatment effects for women than for men. The pattern that we instead observe is unexpected and not easily explained by past results or differences in initial consumption patterns. Furthermore, these treatment effects are temporary, manifesting not immediately, but rather in the second month after the intervention. The unexplained pattern and delayed onset, combined with the marginal significance (*p*-values for the estimated treatment effects never go below 0.01), and short-lasting effects cast doubt that these results would replicate.

If the treatment effects are real, why were they short-lived? Reducers may have exhausted their motivation, found vegetarian/vegan options to be inconvenient or less desirable, or forgot their intention as old habits took over. Habit change is hard and many interventions designed to change habits fail (see for example Carden and Wood, 2018). Alternatively, the winter and summer vacations may have played a disruptive role. In between semesters, students usually go back home. Sustaining a new diet while living away from campus could be challenging. The vacations may be culpable, coinciding with the attenuation of the treatment effects by gender.

Self-identified vegetarians experience very large changes in diet, though the standard errors are also large. Self-identified vegetarians consume about 17.5% of their meals with meat, with more than two-thirds of those meat-based meals containing poultry/fish. Why are "vegetarians" eating meat in the first place? The vegetarian diet may be aspirational, new or temporary, vegetarians may slip, or vegetarians may purchase some meals for their friends. In any case, the treatment reduces poultry/fish consumption by 13.1 percentage points, effectively eliminating it in the first month. Overall meat consumption reduces by 9.9 percentage points (57%), though it is not significant at conventional levels ($p = 0.095$). Whereas, those who eat a large fraction of their meals with meat have much greater capacity to reduce their consumption, self-identified vegetarians, who are already eating much smaller amounts of meat, reduce the most in response to the pamphlets. This finding suggests that, to the extent that the treatment had any effect, it may

[1]Substituting beef for poultry and fish may have perverse ethical consequences. Because cows are large, whereas chickens and most fish are small, this substitution results in more animals slaughtered. Furthermore, the conditions under which chickens are raised are arguably worse than the conditions for cows. Both factors bring into question whether a shift from beef to poultry and fish is an ethical improvement. Due to this concern, the developers of our treatment pamphlet included three times more images of chickens and fish compared to cows (12 vs. 4).

have primarily affected people who were already aligned with the message. We present more evidence in support of this interpretation below.

We found two mechanism variables that predicted larger treatment effects, and one mechanism variable that paradoxically predicted weaker treatment effects. We caution that these significant effects do not survive sharpened FDR correction. It is quite likely that these correlations reflect sampling error and will not replicate. However, we offer some interpretation on the chance that they reflect real changes. Those who "thought more" about the issue after reading the pamphlet and those who reported a willingness to make "big change" exhibited significant reductions in meat consumption in the second month. These findings suggest that those who engaged more with the pamphlet, i.e., thought more about the ethical issues, were more likely to change their diet. Expression of willingness to make a big change was an effective leading indicator of that change. These two variables may have identified individuals with greater intrinsic motivation to change. Interestingly, those who said that the pamphlet "taught" them about the treatment of animals in farms exhibited no treatment effect, but those who did not click this response did exhibit a treatment effect. We interpret this finding as evidence that the pamphlet affected those who were already aware of the issue, but not those who were previously ignorant—and for whom the pamphlet taught them new information. Jalil et al. (2020) found the same result in their study. The pamphlet appears to have been more effective with people who were already aware of the issue. These findings suggest that those who know the least about an issue may be the least likely to respond to this type of policy intervention.

The "stages of change" model from the field of psychology may explain this finding (Prochaska and DiClemente, 1982, 1983). This model posits an order of stages that a person moves through on the path to behavior change: precontemplation, contemplation, action, and maintenance. The pamphlets may shift some from the precontemplation to the contemplation stage, and others from the contemplation to action stages. Only the latter shift results in behavior change. This theory may help explain why the treatment was effective for some groups—those already at the contemplation stage, i.e., self-identified vegetarians and individuals for whom the information was not new.

Though we are able to reject all but small treatment effects, this does not imply that pamphlets are cost-ineffective. Our treatment pamphlet costs $0.07. In highly trafficked corridors, a volunteer can hand out 100 or more pamphlets in an hour. As an example, consider an opportunity cost of $15/h for a volunteer who can hand out 100 pamphlets in an hour. An effect of a 1 percentage-point decrease in meat consumption for 1 month would be equivalent to turning two average meat eaters (who eat about half of their meals with meat) into vegetarians for a month, for a total cost of $22 ($15 for the hour of pamphleteering plus $7 for the cost of 100 pamphlets). If we consider only lunch and dinner (120 meals over 1 month for two individuals), converting half of those meals from meat to vegetarian/vegan would come at cost of $0.37 per meal ($22/60 meals = $0.37). Depending on the estimated ethical (and environmental) externalities, the pamphlet could be cost effective. In contrast, a $0.37 subsidy may not be as effective at inducing a person to switch their meal from meat to vegetarian, though this an open empirical question.

We can compare the effectiveness of pamphleteering to other interventions. Two other recent studies, using real purchase data, have examined the effects of information-based interventions to reduce meat consumption. Both occur on college campuses. Jalil et al. (2020) find that students who listen to a 50-min class lecture on climate change and health reduce their meat consumption by 6.1 percentage points in the semester of the intervention, with a 95% confidence interval of [−0.094, −0.027]. Schwitzgebel et al. (2020) find that students in a philosophy class assigned to think about the ethics of eating meat reduce their meat consumption by 6.3 percentage points for several weeks, with a 95% confidence interval of [−0.102, −0.026]. Because these confidence intervals do not overlap with those from our study, we can conclude that these other interventions have larger effects.

Why are the effects of pamphleteering smaller than the effects from the classroom interventions? The classroom interventions involved nearly an hour of lecture or discussion time in the aforementioned studies, along with required readings in Schwitzgebel et al. (2020). By contrast, reading the pamphlet only takes 5–15 min, a fraction of the time of the classroom interventions, and does not involve multimedia (e.g., videos). Some students only skimmed the pamphlet and others did not read it. Another possibility is that pamphlets are less effective at challenging prior beliefs than active learning. The pamphlet addresses the same ethical issues as those in Schwitzgebel et al. (2020), but the mode of engagement, i.e., asking students to ponder and critically discuss the ethics of eating meat in a class setting, may have caused students to more directly question their preexisting notions.

In conclusion, we provide the first evidence of the effect of animal-advocacy pamphlets on meat consumption using real consumption data. Given that treatment effects are likely small, future work should focus on casting a wider net, via research designs capable of recruiting orders of magnitude more subjects. The welfare of animals on factory farms will continue to be an important issue as global demand for meat grows.

DATA AVAILABILITY STATEMENT

The datasets generated for this article are not readily available because it is owned by an anonymous college and the authors do not have permission to share it publicly. Individual participant data that underlie the results reported in this article will be made available to researchers who provide a methodologically sound proposal. Proposals should be directed to corresponding author at joshua.tasoff@cgu.edu. To gain access, data requestors will need to sign a data access agreement. Additional materials, i.e. Stata code to replicate statistical analysis, intervention materials (leaflets, survey, etc.), are publicly available.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of the host college.

The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MH: implementation, software, and analysis. AJ: conceptualization, experiment design, analysis, and Writing. JT: funding acquisition, conceptualization, experiment design, experiment implementation, analysis, and writing. AV: conceptualization and analysis. All authors contributed to the article and approved the submitted version.

FUNDING

We are extremely grateful to Open Philanthropy Project for funding.

ACKNOWLEDGMENTS

We thank The Humane League Labs, Vegan Outreach, and the Humane Society of the United States for donating pamphlets. The paper has greatly benefitted from conversations with Lewis Bollard and Jacob Peacock, as well as seminar participants from various conferences. We are indebted to the staff of the host college; without their support, hard work and help, the project would not have been possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.668674/full#supplementary-material

REFERENCES

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: a reevaluation of the abecedarian, perry preschool, and early training projects. *J. Am. Stat. Assoc.* 103, 1481–1495. doi: 10.1198/016214508000000841

Animal Charity Evaluators (2013). *Leafleting Study*. Available online at: https://animalcharityevaluators.org/advocacy-interventions/interventions/leafleting/2014-march/2013-leafleting-study/ (accessed November 6, 2020).

Animal Charity Evaluators (2017). *Leafleting*. Available online at: https://animalcharityevaluators.org/advocacy-interventions/interventions/leafleting/ (accessed November 5, 2020).

Anthis, J. R. (2017). *Survey of US Attitudes Towards Animal Farming and Animal-Free Food*. Available online at: https://sentienceinstitute.org/animal-farming-attitudes-survey-2017. Sentience Institute (accessed November 5, 2020).

Anthis, J. R. (2019). *US Factory Farming Estimates*. Available online at: https://sentienceinstitute.org/us-factory-farming-estimates (accessed November 5, 2020).

Bastian, B., Loughnan, S., Haslam, N., and Radke, H. R. (2012). Don't mind meat? the denial of mind to animals used for human consumption. *Pers. Soc. Psychol. Bull.* 38, 247–256. doi: 10.1177/0146167211424291

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93, 491–507. doi: 10.1093/biomet/93.3.491

Bianchi, F., Dorsel, C., Garnett, E., Aveyard, P., and Jebb, S. A. (2018). Interventions targeting conscious determinants of human behaviour to reduce the demand for meat: a systematic review with qualitative comparative analysis. *Int. J. Behav. Nutr. Phys. Act.* 15, 1–25. doi: 10.1186/s12966-018-0729-6

Broom, D. M., and Johnson, K. G. (1993). *Stress and Animal Welfare*, Vol. 993. Dordrecht: Springer.

Broom, D. M., Mendl, M. T., and Zanella, A. J. (1995). A comparison of the welfare of sows in different housing conditions. *Ani. Sci.* 61, 369–385. doi: 10.1017/S1357729800013928

Carden, L., and Wood, W. (2018). Habit formation and change. *Curr. Opin. Behav. Sci.* 20, 117–122. doi: 10.1016/j.cobeha.2017.12.009

Cheng, H. (2006). Morphopathological changes and pain in beak trimmed laying hens. *Worlds Poult. Sci. J.* 62, 41–52. doi: 10.1079/WPS200583

Duncan, I. J. H. (2001). Animal welfare issues in the poultry industry: is there a lesson to be learned? *J. Appl. Ani. Welfare Sci.* 4, 207–221. doi: 10.1207/S15327604JAWS0403_04

Flens, G., Moleman, P., and de Rooy, L. (2018). *The Effectiveness of Leafletting on Reducing the Consumption of Animal Products in Dutch Students*. Working Paper.

Forman, J. J. (2017). *Locking Up Our Own: Crime and Punishment in Black America*. New York, NY: Farrar, Straus and Giroux.

Fraser, D., Mench, J., and Millman, S. (2001). *Farm Animals and Their Welfare in 2000*. (Washington, DC: Humane Society Press), 87–99.

Garnett, E. E., Balmford, A., Sandbrook, C., Pilling, M. A., and Marteau, T. M. (2019). Impact of increasing vegetarian availability on meal selection and sales in cafeterias. *Proc. Natl. Acad. Sci. U.S.A.* 116, 20923–20929. doi: 10.1073/pnas.1907207116

Garnett, E. E., Marteau, T. M., Sandbrook, C., Pilling, M. A., and Balmford, A. (2020). Order of meals at the counter and distance between options affect student cafeteria vegetarian sales. *Nat. Food* 1, 485–488. doi: 10.1038/s43016-020-0132-8

Gregory, N. G., and Wilkins, L. J. (1989). Broken bones in domestic fowl: handling and processing damage in end-of-lay battery hens. *Br. Poult. Sci.* 30, 555–562. doi: 10.1080/00071668908417179

Hansen, P. G., Schilling, M., and Malthesen, M. S. (2019). Nudging healthy and sustainable food choices: three randomized controlled field experiments using a vegetarian lunch-default as a normative signal. *J. Public Health*. fdz154. doi: 10.1093/pubmed/fdz154

Hennessy, S. R. (2016). *The Impact of Information on Animal Product Consumption*. Master's thesis, University of Illinois at Urbana-Champaign.

Jalil, A. J., Tasoff, J., and Vargas Bustamante, A. (2020). Eating to save the planet: Evidence from a randomized controlled trial using individual-level food purchase data. *Food Policy* 95:101950. doi: 10.1016/j.foodpol.2020.101950

Kowal, J. F. (2015). *The Improbable Victory of Marriage Equality*. New York, NY: Brennan Center for Justice. (accessed November 5, 2020).

Marchant, J., and Broom, D. (1996). Effects of dry sow housing conditions on muscle weight and bone strength. *Animal Sci.* 62, 105–114. doi: 10.1017/S1357729800014387

Norwood, B., and Murray, S. (2018). *Food Demand Survey*. Oklahoma State University, 5.

Norwood, F. B., and Lusk, J. L. (2011). *Compassion, by the Pound: The Economics of Farm Animal Welfare*. Oxford: Oxford University Press.

Peacock, J., and Sethu, H. (2017). *Which Request Creates the Most Diet Change?: A Reanalysis*. Technical report. Open Science Framework.

Prochaska, J. O., and DiClemente, C. C. (1982). Transtheoretical therapy: toward a more integrative model of change. *Psychotherapy* 19, 276–288. doi: 10.1037/h0088437

Prochaska, J. O., and DiClemente, C. C. (1983). Stages and processes of self-change of smoking: toward an integrative model of change. *J. Consult. Clin. Psychol.* 51, 390–395. doi: 10.1037/0022-006X.51.3.390

Regan, T. (1983). "Animal rights, human wrongs," in *Ethics and Animals*, (London: Springer), 19–43.

Rothgerber, H. (2020). Meat-related cognitive dissonance: a conceptual framework for understanding how meat eaters reduce negative arousal from eating animals. *Appetite* 146:104511. doi: 10.1016/j.appet.2019.104511

Sanders, B. (2018). *Global Animal Slaughter Statistics And Charts* Olympia, WA. Available online at: https://faunalytics.org/global-animal-slaughter-statistics-and-charts/ (accessed November 5, 2020).

Schröder, M. J., and McEachern, M. G. (2004). Consumer value conflicts surrounding ethical food purchase decisions: a focus on animal welfare. *Int. J. Cons. Stud.* 28, 168–177. doi: 10.1111/j.1470-6431.2003.00357.x

Schwitzgebel, E., Cokelet, B., and Singer, P. (2020). Do ethics classes influence student behavior? Case study: teaching the ethics of eating meat. *Cognition* 203:104397. doi: 10.1016/j.cognition.2020.104397

Singer, P. (1979). Killing humans and Killing animals. *Inquiry* 22, 145–156. doi: 10.1080/00201747908601869

Sonoda, Y., Oishi, K., Chomei, Y., and Hirooka, H. (2018). How do human values influence the beef preferences of consumer segments regarding animal welfare and environmentally friendly production? *Meat Sci.* 146, 75–86. doi: 10.1016/j.meatsci.2018.07.030

Vandenbroele, J., Slabbinck, H., Van Kerckhove, A., and Vermeir, I. (2019). Mock meat in the butchery: nudging consumers toward meat substitutes. *Organ. Behav. Hum. Decis. Proc.* 163, 105–116. doi: 10.1016/j.obhdp.2019.09.004

Vieuille-Thomas, C., Pape, G. L., and Signoret, J. P. (1995). Stereotypies in pregnant sows: indications of influence of the housing system on the patterns expressed by the animals. *Appl. Ani. Behav. Sci.* 44, 19–27. doi: 10.1016/0168-1591(95)00574-C

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 June 2021
doi: 10.3389/fpsyg.2021.614575

What's Behind Image? Toward a Better Understanding of Image-Driven Behavior

*Tobias Regner**

Department of Economics, University of Jena, Jena, Germany

Our experimental design systematically varies image concerns in a dictator/trust game. In comparison to the baseline, we either decrease the role of self-image concerns (by providing an excuse for selfish behavior) or increase the role of social-image concerns (by conveying the transfer choice to a third person). In this set up, we analyze the underlying processes that motivate subjects to give less/more. Controlling for distributional preferences and expectations, our results indicate that moral emotions (guilt and shame) are a significant determinant of pro-social behavior. The disposition to guilt explains giving in the baseline, while it does not when an excuse for selfish behavior exists. Subjects' disposition to shame is correlated to giving when their choice is public and they can be identified.

JEL Classifications: C72, C91, D03, D80

Keywords: social preferences, pro-social behavior, experiment, guilt, shame, reciprocity, self-image concerns, social-image concerns

Edited by:
Agne Kajackaite,
Social Science Research Center Berlin, Germany

Reviewed by:
Emilia I. Barakova,
Eindhoven University of Technology, Netherlands
Valerio Capraro,
Middlesex University, United Kingdom

***Correspondence:**
Tobias Regner
tobias.regner@uni-jena.de

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *06 October 2020*
Accepted: *20 April 2021*
Published: *09 June 2021*

Citation:
Regner T (2021) What's Behind Image? Toward a Better Understanding of Image-Driven Behavior. Front. Psychol. 12:614575. doi: 10.3389/fpsyg.2021.614575

1. INTRODUCTION

What drives pro-social behavior, what motivates us to give more than we have to, even in non-repeated interactions? These questions led to a substantial body of academic work, the literature on social preferences[1]. Early models assumed distributional preferences as an explanation of other regarding behavior (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Andreoni and Miller, 2002), later on beliefs were incorporated into the utility function to take the effect of motives like reciprocity or emotions into account (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Battigalli and Dufwenberg, 2007). Most recently, models emerged that consider the role of image as a determinant of pro-sociality (Akerlof and Kranton, 2000; Bodner and Prelec, 2003; Bénabou and Tirole, 2006, 2011; Ellingsen and Johannesson, 2008; Mazar et al., 2008; Andreoni and Bernheim, 2009; Grossman and van der Weele, 2017).

The insight of such self-/social-image models is that pro-social behavior can depend on the context. Exposure of our choice to others increases the chance of pro-social behavior (e.g., Ariely et al., 2009). Likewise, weakening the connection between action and the self-inducing moral wiggle room tends to decrease it as we are able to attribute the selfish action to the context, instead of having to connect selfish behavior to the self-image (e.g., Dana et al., 2007).

Our experimental study, a modified dictator game, sets out to test what are underlying psychological processes of image-driven behavior. We vary the extent of image concerns that may affect the transfer choice (by decreasing the role of self- and increasing the role of social-imageof

[1]Note that our study focuses on (behavioral) economics and the emergence and evolution of social preferences within this field. While topics like altruism, cooperation and pro-sociality have been discussed before in related disciplines, our starting point is due to the focus on behavioral economics.

concerns). We also turn the dictator game into a trust one in order to study potential interaction effects between image and reciprocal concerns. This setup allows us to focus on determinants image-driven behavior, while we control for factors that are already known to motivate pro-social choices (distributional preferences and expectations).

Our results suggest that behavioral differences resulting from a variation of image concerns origin from moral emotions. The disposition to guilt determines transfers in the baseline but not when the connection between action and outcome is less clear. The disposition to shame is correlated with the transfer size only when another subject gets to know the transfer and potentially sees who made the transfer.

As our study offers a psychological foundation for image-driven behavior, we synthesize existing modeling approaches of social preferences. Belief-dependent models propose that psychological correlates like the disposition to guilt affect pro-sociality (in combination with expectations). Our results indicate that the role of moral emotions in explaining pro-social behavior goes beyond that. Moral emotions may be the responsible underlying process for behavior that has been attributed to image concerns.

A by-product of our design is an estimate of the relative explanatory power of the respective factors influencing pro-social behavior considered in our study. We find that the estimates of a one standard deviation change are very similar for the social value orientation, the second-order beliefs, the disposition to guilt, and the disposition to shame (between 1.06 and 1.17 with a mean transfer of 5.70).

The paper is organized as follows: In section 2, we describe the experiment and present behavioral predictions. Results are reported and discussed in section 3. Section 4 provides the conclusion.

2. STUDY

Our study consists of a lab experiment, which is preceded by an online survey that was administered through an Internet platform 1 week before the experiment.

2.1. Experimental Design

An allocation decision is at the core of the game played in the experiment. Three players are matched together. Player Y chooses how to divide 20 euros between himself and player X with every integer $0 \leq t \leq 20$ possible as the transfer. A third player, Z, is passive and is not affected by the choice. Subjects know that the game is played just once. Our 4×2 between-subjects design varies the game along two dimensions: image (MorEx; baseline; Obs; ObsID) and reciprocity [dictator game (DG) vs. trust game (TG)].

In the *image* dimension, we change the extent of image concerns triggered by player Y's choice. In the baseline and all other conditions, subjects know that player X only learns the received transfer at the very end of the experiment. In Obs, they are informed that player Z observes the transfer. In ObsID, subjects know that Z is informed about the transfer but also about the cubicle number of Y. Moreover, the instructions remind them that at the end of the experiment subjects are called to the front of the lab to receive their earnings. Thus, their cubicle number is announced before they walk to the front. While this is standard procedure in the lab, we also informed them that the order of paying the subjects is varied. It can be from cubicle 1 to 30, or decreasing from 30 to 1, or from 15/16 going down-/upward. Hence, they are made aware that irrespective of their actual cubicle number, a player Y can be seen by his/her Z. Finally, in MorEx, subjects are informed that there is a 50% chance of nature overwriting Y's choice. In that case, the computer picks any possible transfer with equal chance. While Y is told whether his decision is implemented or not, X does not find out whether the received transfer is Y's choice or overwritten by the computer. This setup offers a moral/situational excuse for behaving selfish by introducing uncertainty. See Exley (2015) or Regner and Matthey (2017) for similar designs.

TABLE 1 | Overview of image treatments and their features.

Image treatment	MorEx	Baseline	Obs	ObsID
50% chance of overwriting	Yes	No	No	No
Z observes transfer	No	No	Yes	No
Z informed about Y's cubicle number	No	No	No	Yes

In *reciprocity*, the game is either played as it is (dictator) or with a preceding stage (trust) in which player X has a binary choice between entering the game (and letting Y decide) and an outside option that results in a payoff of 5 euros for both X and Y. We ask trustees for their decisions independent of the trustor's choice, that is, we use the strategy method (Selten, 1967).

To summarize, our experimental design systematically varies image concerns by decreasing self-image concerns in MorEx (with social-image concerns kept constant) and by increasing social-image concerns in Obs and ObsID (while self-image concerns remain constant). See **Table 1** for an overview of the image treatments and their features. We also compare a dictator to a trust game setting in order to study potential interaction effects between image and reciprocal concerns.

After game choices were made, we asked subjects for their probabilistic (or distributional) first- and second-order *beliefs*. For subjects Y, this is the belief with respect to X's choice to enter the game (first-order belief), and the belief about the expectation of subject X with respect to subject Y's transfer (second-order belief). For subjects X, this is the belief about subject Y's transfer (first-order belief), and the belief about the expectation of subject Y with respect to subject X's choice to enter the game (second-order belief). Subject Z was asked for two first-order beliefs (with respect to X's choice to enter the game and about Y's transfer) and two second-order beliefs (about the expectation of X with respect to Y's transfer and about the expectation of Y with respect to X's choice to enter the game).

The probabilistic beliefs were collected as vectors for a series of intervals. Regarding the choice to enter the game, subjects could assign their first-order belief to the options 0 (no) and 1 (yes) and the second-order belief to the intervals $[0, 10), [10, 20), [20, 30), ..., [90, 100]$ percent. They could distribute their belief regarding the transfer to the following

intervals: [0, 2), [2, 4), [4, 6), ..., [18, 20] euro. The software made sure that the numbers a subject is assigned sum up to 100%. Figure 2 in **Appendix A** shows a screenshot of the decision interface for the first-order belief of a player X (or Z). Beliefs were elicited using a quadratic scoring rule. In contrast to a linear scoring rule, a quadratic one is incentive compatible which tends to result in more accurate predictions (see Palfrey and Wang, 2009; Armantier and Treich, 2013; Schotter and Trevino, 2014). In contrast to point (or non-probabilistic) forecasts, probabilistic forecasts allow participants to express uncertainty about their belief. See Manski and Neri (2013) for a comprehensive account of probabilistic and non-probabilistic elicitation of second-order beliefs.

The instructions informed subjects that after stage 1 of the experiment, consisting of the game as described, they will play two more stages for which they receive instructions in due course. In stage 2, the same game was played but roles were changed: subjects were rotated, that is, player Y of stage 1 now played as X, Z as Y, and X as Y. In stage 3, subjects were rotated once more so that each subject played in each role. Resulting payoffs of all stages were only announced at the end of the experiment (after stage 3). Subjects knew that one of the three stages was randomly chosen as payoff-relevant.

2.2. Measures From the Online Survey

A week before the actual lab experiment, subjects participated in an online survey administered through an Internet platform. The aim of the survey was to assess subjects' *social value orientation* and their *dispositions with respect to guilt and shame* in advance of the actual experiment.

The *social value orientation* (SVO) slider measure (Murphy et al., 2011) consists of six primary items and nine optional ones. In each item, subjects face a resource allocation choice over a well-defined continuum of payoffs for themselves and someone else: for instance, one item features a trade-off between the perfectly altruistic choice of (50, 100) and the perfectly individualistic choice of (100, 50). In between these extreme values, there are always seven allocations that allow for intermediate choices. From choices in the six primary items, the SVO angle is computed, a continuous measure that we employ as a proxy for the subjects' concern for the payoff of others. The SVO angle reflects individualistic (maximizing own payoffs), competitive (maximizing the difference between own and other's payoff), inequality averse (minimizing the difference between own and other's payoff), and efficiency (joint payoff maximizing) motives.

The Guilt And Shame Proneness scale (GASP) by Cohen et al. (2011) is an innovative scale to measure individuals' *dispositions with respect to guilt and shame*. It assumes that private transgressions trigger feelings of guilt, while public transgressions trigger feelings of shame. Hence, their guilt scenarios are all set in the private domain, and the shame scenarios are always public situations. It also incorporates the self-behavior conceptualizations of shame and guilt and additionally distinguishes evaluative responses from action orientations. In total, the GASP consists of 16 real-life scenarios. Subjects are asked to imagine they were in that situation and indicate the likelihood that they would react in the way described at the end of the scenario[2]. While the ability to evaluate own behavior (captured by the NBE sub-scale) should be most indicative for pro-social guilt-driven behavior, the evaluative sub-scale for shame (NSE) should be indicative for an ability to anticipate feeling ashamed.

2.3. Behavioral Predictions

The literature of social preferences started off with outcome-based models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Andreoni and Miller, 2002) using distributional preferences in order to explain pro-social behavior. Subsequently, with the development of belief-dependent models (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Battigalli and Dufwenberg, 2007), the role of expectations as a determinant of pro-social behavior gained attention.

In line with this literature, we generally expect that two factors motivate the choice of the transfer in our experiment: the individual's distributional preferences and expectations about what the recipient expects to get. Thus, we expect that the size of the transfer is positively correlated with our proxy for the level of distributional preferences, the SVO angle, and the second-order beliefs[3] of the player who sends the transfer.

More recently, image concerns have been incorporated in the economic modeling of pro-social behavior. *Self-image* concerns (Murnighan et al., 2001; Bodner and Prelec, 2003; Mazar et al., 2008; Bénabou and Tirole, 2011) explain pro-social behavior as a consequence of desiring a self-image (alternatively, a self-concept or behavioral standard) of not being selfish[4]. As deviating from the pro-social self-image is psychologically costly, selfish choices result, only if the monetary gain of a selfish action outweighs that cost. Supporting the relevance of self-image concerns, a series of studies, started by Dana et al. (2007), finds that pro-social choices are significantly reduced when moral excuses for selfish behavior are available. Evidence of such "moral wiggle room" indicates/suggests that the effect of self-image concerns is toned down, if the connection between actions and the self is blurred. Once individuals are able to attribute their selfish action to the context, instead of having to connect selfish behavior to their self-image, they tend to behave more selfish.

Individuals can also have *social-image* concerns (Bénabou and Tirole, 2006; Ellingsen and Johannesson, 2008; Andreoni and Bernheim, 2009), if they desire not to appear selfish to others, especially their peers. Due to such concerns for their social reputation, individuals would be more likely to make a pro-social

[2]For guilt there are four scenarios with negative behavior-evaluations (NBE) and four scenarios with repair responses (REP). For shame, there are four scenarios with negative self-evaluations (NSE) and four scenarios for withdrawal responses (WIT). See **Appendix B** for details of the GASP questionnaire.

[3]A stream of research tests the robustness of beliefs' causal effect on behavior (e.g., Vanberg, 2008; Ellingsen et al., 2010; Bellemare et al., 2011; Costa-Gomes et al., 2014; Kawagoe and Narita, 2014; Khalmetski, 2016; Ederer and Stremitzer, 2017).

[4]A related stream of literature proposes preferences for following (personal/social) norms as an explanation for pro-sociality in anony mous one-shot situations, see Capraro and Perc (2021) for a review. Moral preference models (e.g., Bicchieri and Chavez, 2010; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016) find empirical support in various experiments (e.g., Capraro and Rand, 2018; Tappin and Capraro, 2018; Capraro et al., 2019).

choice, if an audience they care about is able to observe their decision. Plenty of empirical evidence from the lab (e.g., Kurzban et al., 2007; Ariely et al., 2009; Henry and Sonntag, 2019) and the field (e.g., Lacetera and Macis, 2010; Bursztyn and Jensen, 2017) highlights the importance of the social image component when it comes to pro-social behavior.

Our experimental design systematically varies image concerns. In comparison to the baseline, the treatment MorEx decreases the role of self-image concerns as it provides a moral excuse for selfish behavior. As subjects know that the transfer may be overwritten, players *Y* may send a low amount, and *X* cannot distinguish whether it was *Y*'s choice or forced by the computer. Moreover, they know that their choice of a low transfer may not actually matter as it could be replaced by the computer[5]. Thus, we expect that transfers tend to be smaller.

Hypothesis 1. *When moral excuses are available (MorEx), transfers are, on average, lower than in the baseline.*

In comparison to the baseline, treatments Obs and ObsID increase the role of social-image concerns as the transfer choice is conveyed to a third person. Thus, selfish behavior potentially bears a reputational cost. We consider treatment Obs as a weak manipulation of social image (public exposure) though, since due to the anonymity of the experiment the choice cannot be traced back to a specific subject. This anonymity is lifted in treatment ObsID. Subjects know that their observer (player *Z* when they played as *Y*) is not only informed about the transfer but also might well be able to identify him-/herself at the end of the experiment. Note that other dimensions of social-image concerns are constant across conditions. Players *X* always know what they receive but never find out who sent it. The experimenter sees the subjects when handing over their payoffs but does not know the stage and role of the subject.

Hypothesis 2. *When the choice can be observed (Obs, ObsID), transfers are, on average, higher than in the baseline.*

Our next hypothesis addresses the interplay between image concerns and reciprocity. Some studies already investigated self-image concerns in the context of reciprocity. Regner and Matthey (2017) and Regner (2018a) find that the effect of moral wiggle room prevails in the context of reciprocity, while van der Weele et al. (2014) do not[6]. Ellingsen and Johannesson (2008) propose that social-image concerns depend on the audience; people care relatively more about the approval of peers, of people who are like them. Trustors' behavior (entering the game vs. outside option) can be seen as a signal about their pro-sociality which would tend to affect the concern trustees have for them. In Obs and ObsID, having been trusted in the first place may increase the level of approval toward the trustor and potentially amplifies the positive effect of social-image concerns on behavior. Hence—assuming reciprocal behavior in the baseline—we test whether reciprocity is more pronounced when an audience exists.

Hypothesis 3. *The difference between transfers in trust and dictator is, on average, higher in a public context (Obs and ObsID) than in the baseline.*

Given that image concerns affect behavior across treatments—on top of the effect of beliefs and SVO—we are also interested in the processes behind this relationship. An aversion to experience guilt, as proposed by Battigalli and Dufwenberg (2007), can be a determinant of the allocation decision: the more I believe you were disappointed due to my choice, the more guilt I would anticipate to feel. According to Tangney (1995) individuals differ in the degree to which they are prone to feel guilt. Thus, expectations as well as the sensitivity of a person to experience guilt influence the choice, and a series of empirical evidence supports these relationships (e.g., Charness and Dufwenberg, 2006; Pelligra, 2011; Bracht and Regner, 2013; Khalmetski, 2016; Cartwright, 2019). In our baseline, the relationship between the choice of *Y* and what *X* receives is transparent. Therefore, we expect the subjects' disposition to guilt to be positively correlated with the size of the transfer. However, the introduced uncertainty in MorEx means that subjects do not have to link the outcome of the recipient to their choice. They can tell themselves that even though *X* might be disappointed by the chosen transfer, there is still a 50% chance that their choice does not count[7]. Hence, we expect a breakdown of the relationship between the subjects' disposition to guilt and the size of the transfer in MorEx.

Hypothesis 4. *The GASP NBE (disposition to guilt) is a positive determinant of the transfer in the baseline (and Obs) while it is not in MorEx.*

Next, we look at social-image concerns in more detail. What are the potential underlying processes behind increased pro-social behavior in a public context? Based on insights from social psychology (e.g., Combs et al., 2010; Wolf et al., 2010), transgressions of morally accepted behavior in the public trigger feelings of shame (while transgressions that remain within the self lead to feelings of guilt).

In the context of our experiment, the morally accepted behavior is arguably an even split, that is, a transfer of 10. The more a subject falls short of that amount, the higher the resulting transgression might be. Hence, we expect that in ObsID subjects'

[5]See Exley (2015) and Regner and Matthey (2017) for related studies that employ uncertainty in order to introduce an excuse for selfish behavior. In the model of Tirole et al. (2016) uncertainty is one of several ways to create moral wiggle room which, in turn, the self can use as a narrative to legitimate a selfish action. Note that uncertainty could also be interpreted to have a positive effect on transfers. In the self-signaling model of Grossman (2015), uncertainty cheapens the expected cost of sending a pro-social signal and more giving is predicted. However, his experimental tests do not provide conclusive supporting evidence.

[6]In a related setting, Malmendier et al. (2014) analyze subjects' behavior when an "exit option" is introduced into a double dictator game. They find substantial sorting out in a positive reciprocity condition (about 30%) but less than in a neutral condition (50%).

[7]Uncertainty is also a feature in the design employed by, for instance, Charness and Dufwenberg (2006) or Bracht and Regner (2013). Their chance move means that even though the agent behaves pro-social still a bad outcome for the principal can occur. Instead, an opportunistic choice for sure results in a bad outcome for the principal and implies that the agent knows this leads to disappointment. Note that the effect of uncertainty in our manipulation is broader/stronger, since an opportunistic choice (sending a low amount) does not necessarily mean that the principal receives a low amount.

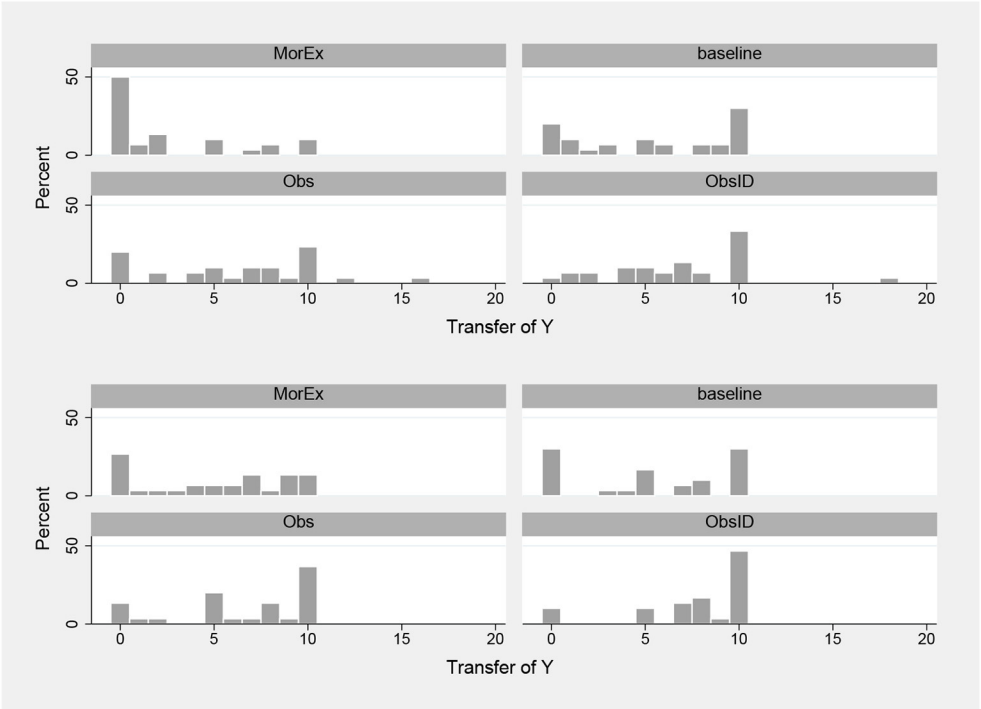


FIGURE 1 | Histograms of transfers by treatment (top: dictator; bottom: trust).

disposition to shame is, on average, positively correlated with the transfer as they anticipate that a low transfer might result in a shameful experience.

Hypothesis 5. *The GASP NSE (disposition to shame) is a positive determinant of the transfer in ObsID.*

2.4. Participants and Procedures

We recruited 240 subjects from various disciplines at the local university using ORSEE (Greiner, 2004). In each session, gender composition was approximately balanced and subjects took part only in one session. Subjects who already participated in similar experiments were excluded from the recruitment pool. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007) and took, on average, 60 min. The average earnings in the experiment have been €12.69 (plus a €2.50 show-up fee and €3 for completing the online survey). Only subjects who completed the online survey were allowed to participate in the experiment. However, one subject slipped through the controls, and survey data are not available.

Upon arrival at the laboratory, subjects were randomly assigned to one of the computers. Each computer is in a cubicle that does not allow communication or visual interaction. After subjects finished reading the instructions, they were asked to answer a set of control questions in order to ensure understanding. After all subjects had answered the questions correctly, the experiment started. At the end of the experiment subjects were paid in cash according to their performance. Privacy was guaranteed during the payment phase.

3. RESULTS

We start with some descriptives of the data and proceed then to regression analyses in order to test our hypotheses.

Figure 1 shows histograms of the transfer for each treatment. Transfers increase along the image dimension of our design [means are 3.77 (MorEx), 5.37 (baseline), 6.37 (Obs), 7.32 (ObsID)] and the histograms give an indication why. In MorEx, the distributions peak at zero. The ones in the baseline are bi-modal, featuring a spike at zero and one at the equally splitting transfer of ten. This is also the case in Obs for the dictator condition, while in the trust condition and in ObsID, the spike at zero disappears.

Table 2 presents the results of OLS regressions with robust standard errors. The dependent variable is the transfer Y sends to X. The specification in the first column includes dummies for the treatments (TG represents the trust condition) and a control for the stage as some subjects played as Y in stage 1, some in 2, and some in 3. The dummy for MorEx is negative and significant at the 5%-level, the dummy for Obs is not significantly different from zero, and the dummy for ObsID is positive and significant at the 1%-level. The specification in column 2 adds the SVO angle and second-order beliefs as further control variables. Their coefficients are positive and highly significant, while the significance levels of MorEx and ObsID drop. The specification in column 3 adds an interaction term between TG and the second-order beliefs. The dummy for TG is positive and significant at the 5%-level, while the interaction term between TG and the second-order beliefs is negative and significant at the 5%-level.

TABLE 2 | Treatment comparison.

	(1)	(2)	(3)	(4)
MorEx	−1.60** (0.71)	−1.07* (0.62)	−0.99 (0.63)	−1.98** (0.86)
Obs	1.00 (0.73)	1.00 (0.64)	1.00 (0.63)	0.91 (0.83)
ObsID	1.95*** (0.68)	1.49** (0.58)	1.49*** (0.57)	0.74 (0.76)
TG	0.88* (0.49)	0.46 (0.45)	2.44** (0.97)	1.53 (1.24)
Stage	−0.66** (0.31)	−0.46* (0.27)	−0.46* (0.27)	−0.46* (0.27)
SVO angle		0.10*** (0.021)	0.10*** (0.021)	0.10*** (0.021)
2nd order beliefs		0.30*** (0.075)	0.44*** (0.10)	0.43*** (0.11)
TG × 2nd order beliefs			−0.29** (0.13)	−0.29** (0.14)
MorEx × TG				2.01 (1.25)
Obs × TG				0.17 (1.24)
ObsID × TG				1.53 (1.14)
Constant	6.24*** (0.83)	1.62** (0.78)	0.84 (0.83)	1.30 (0.92)
Observations	240	239	239	239
R^2	0.136	0.348	0.362	0.373

*OLS regressions; robust standard errors in parentheses; *p < 0:10, **p < 0:05, ***p < 0:01; one subject managed to participate in the experiment without having completed the online survey; variance inflation factors are all <2, indicating no concern for multicollinearity.*

The dummy for ObsID is significant at the 5%-level, and the dummy for MorEx is still negative but not at a significant level. Finally, specification 4 adds interaction terms between the trust condition and the image dummies in order to test for the effect of reciprocity. Neither the main effects nor the interaction terms are significant. Dropping the controls, SVO angle and second-order beliefs deliver the same qualitative results.

Our results are consistent with our expectation that the SVO angle and second-order beliefs are significant determinants of the transfer. Furthermore, they support our hypothesis that transfers are higher when social-image concerns matter, albeit only in the treatment with potential public identification (ObsID). Also, our hypothesis regarding MorEx—transfers are lower when self-image concerns are inhibited by an available excuse—finds support although the effect is weakened when adding controls. While we find general reciprocal behavior (positive and significant coefficient of TG in specification 3), the effect in the baseline is not strong enough to be significant and neither are the interactions between the TG and other treatment dummies (specification 4). The apparent lack of baseline reciprocity complicates the testing of the respective hypothesis (increased reciprocal behavior in treatments Obs and ObsID), and we will get back to this later. Finally, the negative interaction effect between TG and second-order beliefs (in combination with their positive main effects) indicates that either the mere fact of being in the trust condition or second-order beliefs increase transfers in TG but not both factors jointly.

Result 1. *In MorEx, transfers are, on average, lower than in the baseline.*

Result 2. *In ObsID, transfers are, on average, higher than in the baseline.*

TABLE 3 | Processes within each image condition.

	MorEx	Baseline	Obs	ObsID
TG	2.29 (1.79)	2.86 (1.91)	4.88** (2.22)	1.86 (2.30)
SVO angle	0.13*** (0.040)	0.081** (0.034)	0.091** (0.041)	0.13*** (0.028)
2nd order beliefs	0.24 (0.19)	0.67*** (0.17)	0.65** (0.25)	0.31* (0.16)
TG × 2nd order beliefs	−0.0041 (0.26)	−0.48* (0.25)	−0.77** (0.31)	−0.095 (0.27)
GASP_NBE	−0.46 (0.48)	0.95** (0.40)	−0.22 (0.45)	−0.76* (0.38)
GASP_NSE	0.80 (0.62)	−0.64 (0.46)	0.74 (0.53)	1.13*** (0.41)
Stage	−0.052 (0.55)	−0.33 (0.51)	−0.33 (0.59)	−0.78* (0.44)
Constant	−3.38 (3.19)	−1.54 (2.66)	−2.05 (3.90)	0.55 (2.73)
Observations	59	60	60	60
R^2	0.332	0.477	0.251	0.449

*OLS regressions; robust standard errors in parentheses; *p < 0:10, **p < 0:05, ***p < 0:01; variance inflation factors are all <5, indicating no concern for multicollinearity.*

We proceed to a more detailed analysis of image-driven behavior. For this purpose, **Table 3** presents one OLS regression for each image condition. The dependent variable is again *Y*'s transfer. Explanatory variables are dummies for the TG condition, the SVO angle, second-order beliefs, and the two sub-scales from the GASP that proxy the disposition to guilt (NBE) and shame (NSE).

In MorEx, only the SVO angle is significant (at the 1%-level). In the baseline, the SVO angle and second-order beliefs are significant. In addition, the NBE sub-scale is significant at the 5%-level. Results in Obs resemble the overall results presented in **Table 2**: SVO angle, second-order beliefs, TG, and the interaction term between the last two are significant. In ObsID, the SVO angle and the NSE sub-scale are significant at the 1%-level.

Results in MorEx and the baseline support the respective hypotheses. With full transparency between actions and outcomes, the moral compass of subjects seems to be intact. As subjects know their transfer potentially disappoints *X*, anticipated guilt seems to keep them from sending low amounts, in line with the results of Bracht and Regner (2013). In contrast, when a low transfer choice does not necessarily mean a small received amount, the beliefs/guilt/pro-sociality system appears to break down. Only a base level of pro-sociality remains in the data.

Result 3. *The disposition to guilt is positively correlated to the transfer in the baseline but not in MorEx.*

Also, results in ObsID are consistent with our corresponding hypothesis. The disposition of the subjects to shame is a significant determinant of their transfer, when the setting is public and they could be recognized by the person who is informed about their transfer. The shame effect appears to crowd out the effect of second-order beliefs and of the TG treatment. In an additional specification, we included an interaction term between the second-order beliefs and the disposition to shame. The interaction is not significant. The effect of shame seems to stand on its own. This seems to suggest that the effect of

shame (and anticipating it) is not about someone else and their expectations but the self.

Result 4. *The disposition to shame is positively correlated to the transfer in ObsID but not in baseline.*

Based on the treatment-specific coefficients shown in **Table 3**, the following changes of the estimated transfer result from a one standard deviation change of our main explanatory variables: SVO angle (1.06–1.7), second-order beliefs (1.06–2.29), GASP NBE (1.17), and the GASP NSE (1.16). Ranges express minimum/maximum values when the coefficient is significant.[8] Given that the mean transfer is 5.70, a one standard deviation change results in roughly a 20% variation of the transfer (using lower bound estimates), independent of which factor changes. Hence, our statistically significant results also seem economically relevant.

Our results so far show that behavioral differences resulting from a variation of image concerns appear to have a sound psychological foundation in moral emotions. The disposition to guilt—in combination with expectations—determines transfers in the baseline but not in MorEx when the connection between action and outcome is less clear. The disposition to shame is correlated with the transfer size in ObsID when another subject gets to know the transfer and potentially sees who made the transfer. Increased pro-social behavior under public exposure is in line with results in Tadelis (2011). In his experiment, trustees cooperate significantly more often when their choice is announced to the entire lab than in the baseline. He does not elicit subjects' disposition to shame, though. Our results are less clear with respect to the interplay between image concerns and reciprocity (hypothesis 3). We do find overall reciprocal behavior (after controlling for the SVO angle, second-order beliefs, and their interaction with the trust dummy), but the effect is not significant in the baseline alone. Moreover, there is no evidence of increased reciprocity in the treatments with a public context (Obs and ObsID). A possible explanation is that our manipulation can be regarded as relatively weak. We use the strategy method for trustors' choices and, therefore, trustees do not know for sure whether they are trusted or not[9]. The actual effect of the trust condition on the transfer may, however, be affected by the beliefs of the subjects.

We turn to our beliefs data in an attempt to shed more light on this. Figure 3 in **Appendix E** shows the distribution of second-order beliefs (see Table 6 in **Appendix D** for summary statistics). Recall that we elicited probabilistic beliefs, not just point beliefs. That is, each subject told us the distribution of their beliefs, allocating probability weights to 10 intervals. Thus, Figure 3 in **Appendix E** illustrates the average weights, across subjects, for all intervals. Generally, in MorEx-dictator, subjects express the most pessimistic second-order beliefs. Most strikingly, about 33% of the probability mass is, on average, assigned to the interval including a first-order belief of a transfer of zero. In contrast, this is the case for only about 13% in MorEx-trust (ranksum test, $p = 0.006$). In baseline, this pattern is similar but less pronounced (about 26% in baseline-dictator vs. 13% in baseline-trust, ranksum test, $p = 0.06$). Beliefs in Obs and ObsID tend to be more optimistic in trust than in dictator, although beliefs of a zero transfer are practically equal. Overall, it seems that the trust condition has a positive effect on the subjects' second-order beliefs, which are positively correlated with the transfer. Thus, testing for the true effect of the trust condition in public settings would require to take second-order beliefs into account. Indeed, results of a mediation analysis indicate that the effect of the trust condition on the transfer is partly mediated by second-order beliefs[10]. Therefore, our overall results suggest that besides the direct effect of the trust condition on the transfer, there exists an indirect effect *via* higher second-order beliefs.

4. CONCLUSIONS

Our experiment systematically varies the role of image concerns in order to study the underlying processes that determine pro-social behavior. In comparison to our baseline, our design reduces the role of self-image concerns by providing a moral excuse for selfish behavior in the MorEx condition, and it allows for social-image concerns by introducing an audience in conditions Obs and ObsID.

We find that behavior across the conditions is in line with image concerns: Transfers are lower in MorEx and higher in ObsID. Our further analysis provides a psychological basis for image-driven behavior. We show that the disposition to guilt, a known determinant of pro-social behavior in previous research and also significant in our baseline, does not guide subjects when a moral excuse exists. Under public exposure of the transfer and potential facial identification of the subject who made the transfer, the disposition to shame is a significant determinant of the transfer choice.

Thus, our results suggest that moral emotions, like guilt and shame, are an important driver behind context-dependent pro-sociality[11]. Does that mean our pro-social choices are "emotional," rooted in system 1? Two recent meta-studies analyze the role of intuition and deliberation in cooperation. While Fromell et al. (2020) find no significant difference when the intuitive system 1 was promoted at the expense of the deliberative system 2, Rand (2016) reports a 17% increase of "pure" cooperation when intuition was promoted over deliberation. It is the anticipation of guilt/shame that is behind pro-social choices in the belief-dependent models. Such an active avoidance of a

[8] See Table 4 in the **Appendix C** for summary statistics of the respective variables.

[9] While it is generally acknowledged that there are no systematic behavioral differences between strategy method and direct response (Brandts and Charness, 2011), some evidence exists that trustees behave less trustworthy when using the strategy method (Casari and Cason, 2009). See also García-Pola et al. (2020) for behavioral differences in a related setting.

[10] Following Baron and Kenny (1986), we, first, establish that there is a correlation between the trust dummy and the transfer (controlling for the treatment dummies and the SVO angle) and that the trust dummy and second-order beliefs are correlated. Moreover, second-order beliefs as well as the trust dummy are correlated with the transfer. See **Appendix D** for details.

[11] A reviewer pointed out that guilt and shame are culture-specific characteristics. In the vein of WEIRD (see Henrich et al., 2010), it is important to note that our results are based on a predominantly German-speaking student sample. They may not necessarily extend to non-WEIRD contexts.

potentially unpleasant situation arguably requires deliberation while emotions are involved at the same time. Thus, it is not necessarily an intuitive action, yet one based on emotions.

Furthermore, we find that our proxy for distributional preferences—the SVO angle—is a consistent determinant of the transfer size across all treatments. Also second-order beliefs—the key parameter of belief-dependent models—are a significant explanatory factor of the transfer. Interestingly, all four factors seem to have a similar impact on the size of the transfer (a one standard deviation change results in roughly a 20% variation).

It seems that distributional preferences, expressed by the SVO angle in our setting, provide a base level of pro-sociality that is unaffected by our treatment manipulations. Beliefs about others' expectations appear to play a major role in determining pro-sociality in treatments without (successful) manipulation. If the connection between choice and outcome is manipulated to be less transparent, the positive influence of second-order beliefs (and the disposition to guilt) on the size of the transfer erodes. Likewise, second-order beliefs seem to play only a marginal role when our treatment manipulation allows for public identification of the transfer and who sent it. The positive effect of the disposition to shame, a self-focused construct, appears to crowd out the impact of beliefs about others.

To conclude, we discuss the limitations and the possible future expansions of our study. One potential concern about our results is that the sample size per treatment cell (30 subjects) is not big, thus statistical power might be an issue. The sample size in preceding related studies is, however, similarly small (e.g., Dana et al., 2007; Andreoni and Bernheim, 2009). Hence, the effect sizes seem big enough for such samples.

Our experiment identifies shame as a channel that is behind increased giving in public situations. Committing a moral transgression by not giving as much as is expected would result in experiencing shame. Anticipating these psychological consequences of selfish behavior results in a transfer that is deemed compliant with morally/socially accepted behavior. A similar channel one could think of is pride. By giving more than expected one would experience pride or prestige. The psychological scale we used, the GASP, does not include a measure of pride, and therefore, we cannot test this potential channel further. It appears plausible that pride has a positive effect on giving, especially in situations where individuals can stick out from the crowd, in a positive sense, instead of avoiding a potentially shameful experience. However, the results of Samek and Sheremeta (2014) do not indicate a "prestige" effect in a related setting, a public goods game.

Our implementation of uncertainty is just one way to reduce the role of self-image concerns. Other ways to introduce moral wiggle room exist (e.g., plausible deniability, delegation, and strategic ignorance), and self awareness can also be manipulated directly. It remains to be seen, to what extent reduced giving following other interventions is also explained by the erosion of the beliefs/guilt system.

Finally, our experimental design considers two treatments with an exposure to an audience, Obs and ObsID. Both result in higher average transfers than in the baseline but only the difference in ObsID is significant, and this seems to be rooted in the disposition to shame. Although a third party is informed about the transfer, it seems that it is the public identification that kicks off the processes that lead to significantly increased giving. Nevertheless, the results in Obs differ from those in the baseline. Hence, a distinct process—not based on the disposition to guilt—might have been triggered. Either way, for the exposure effect in social-signaling models, public identification, like in ObsID, appears to be necessary.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Max Planck Institute of Economics ethics committee. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I would like to thank the audiences at the IMEBESS in Barcelona and the ESA world congress in Berlin. Maximilian Wechsung provided excellent research assistance. Financial support by the Max Planck Institute of Economics, Jena, Germany and by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 628902—was gratefully acknowledged. A working paper version of this study is published as Regner (2018b).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.614575/full#supplementary-material

REFERENCES

Akerlof, G. A., and Kranton, R. E. (2000). Economics and identity. *Q. J. Econ.* 115, 715–753. doi: 10.1162/003355300554881

Andreoni, J., and Bernheim, B. D. (2009). Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. *Econometrica* 77, 1607–1636. doi: 10.3982/ECTA7384

Andreoni, J., and Miller, J. (2002). Giving according to garp: an experimental test of the consistency of preferences for altruism. *Econometrica* 70, 737–753. doi: 10.1111/1468-0262.00302

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Am. Econ. Rev.* 99, 544–555. doi: 10.1257/aer.99.1.544

Armantier, O., and Treich, N. (2013). Eliciting beliefs: proper scoring rules, incentives, stakes and hedging. *Eur. Econ. Rev.* 62, 17–40. doi: 10.1016/j.euroecorev.2013.03.008

Baron, R. M., and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51:1173. doi: 10.1037/0022-3514.51.6.1173

Battigalli, P., and Dufwenberg, M. (2007). Guilt in games. *Am. Econ. Rev.* 97, 170–176. doi: 10.1257/aer.97.2.170

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *J. Appl. Econometr.* 26, 437–453. doi: 10.1002/jae.1227

Bénabou, R., and Tirole, J. (2006). Incentives and prosocial behavior. *Am. Econ. Rev.* 96, 1652–1678. doi: 10.1257/aer.96.5.1652

Bénabou, R., and Tirole, J. (2011). Identity, morals, and taboos: beliefs as assets. *Q. J. Econ.* 126, 805–855. doi: 10.1093/qje/qjr002

Bicchieri, C., and Chavez, A. (2010). Behaving as expected: public information and fairness norms. *J. Behav. Decis. Mak.* 23, 161–178. doi: 10.1002/bdm.648

Bodner, R., and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *Psychol. Econ. Decis.* 1, 105–126.

Bolton, G. E., and Ockenfels, A. (2000). ERC: a theory of equity, reciprocity, and competition. *Am. Econ. Rev.* 90, 166–193. doi: 10.1257/aer.90.1.166

Bracht, J., and Regner, T. (2013). Moral emotions and partnership. *J. Econ. Psychol.* 39, 313–326. doi: 10.1016/j.joep.2013.09.007

Brandts, J., and Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Exp. Econ.* 14, 375–398. doi: 10.1007/s10683-011-9272-x

Bursztyn, L., and Jensen, R. (2017). Social image and economic behavior in the field: identifying, understanding, and shaping social pressure. *Annu. Rev. Econ.* 9, 131–153. doi: 10.1146/annurev-economics-063016-103625

Capraro, V., Jagfeld, G., Klein, R., Mul, M., and Van De Pol, I. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Sci. Rep.* 9:11880. doi: 10.1038/s41598-019-48094-4

Capraro, V., and Perc, M. (2021). Mathematical foundations of moral preferences. *J. R. Soc. Interface* 18:20200880. doi: 10.1098/rsif.2020.0880

Capraro, V., and Rand, D. G. (2018). Do the right thing: experimental evidence that preferences for moral behavior, rather than equity or efficiency *per se*, drive human prosociality. *Judgm. Decis. Mak.* 13, 19–111. doi: 10.2139/ssrn.2965067

Cartwright, E. (2019). A survey of belief-based guilt aversion in trust and dictator games. *J. Econ. Behav. Organ.* 167, 430–444. doi: 10.1016/j.jebo.2018.04.019

Casari, M., and Cason, T. N. (2009). The strategy method lowers measured trustworthy behavior. *Econ. Lett.* 103, 157–159. doi: 10.1016/j.econlet.2009.03.012

Charness, G., and Dufwenberg, M. (2006). Promises and partnership. *Econometrica* 74, 1579–1601. doi: 10.1111/j.1468-0262.2006.00719.x

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the gasp scale: a new measure of guilt and shame proneness. *J. Pers. Soc. Psychol.* 100:947. doi: 10.1037/a0022641

Combs, D. J., Campbell, G., Jackson, M., and Smith, R. H. (2010). Exploring the consequences of humiliating a moral transgressor. *Basic Appl. Soc. Psychol.* 32, 128–143. doi: 10.1080/01973531003738379

Costa-Gomes, M. A., Huck, S., and Weizsäcker, G. (2014). Beliefs and actions in the trust game: creating instrumental variables to estimate the causal effect. *Games Econ. Behav.* 88, 298–309. doi: 10.1016/j.geb.2014.10.006

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econ. Theory* 33, 67–80. doi: 10.1007/s00199-006-0153-z

Dufwenberg, M., and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games Econ. Behav.* 47, 268–298. doi: 10.1016/j.geb.2003.06.003

Ederer, F., and Stremitzer, A. (2017). Promises and expectations. *Games Econ. Behav.* 106, 161–178. doi: 10.1016/j.geb.2017.09.012

Ellingsen, T., and Johannesson, M. (2008). Pride and prejudice: the human side of incentive theory. *Am. Econ. Rev.* 98, 990–1008. doi: 10.1257/aer.98.3.990

Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games Econ. Behav.* 68, 95–107. doi: 10.1016/j.geb.2009.04.021

Exley, C. L. (2015). Excusing selfishness in charitable giving: the role of risk. *Rev. Econ. Stud.* 82, 587–628. doi: 10.1093/restud/rdv051

Falk, A., and Fischbacher, U. (2006). A theory of reciprocity. *Games Econ. Behav.* 54, 293–315. doi: 10.1016/j.geb.2005.03.001

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868. doi: 10.1162/003355399556151

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10, 171–178. doi: 10.1007/s10683-006-9159-4

Fromell, H., Nosenzo, D., and Owens, T. (2020). Altruism, fast and slow? Evidence from a meta-analysis and a new experiment. *Exp. Econ.* 23, 979–1001. doi: 10.1007/s10683-020-09645-z

García-Pola, B., Iriberri, N., and Kovářík, J. (2020). Hot versus cold behavior in centipede games. *J. Econ. Sci. Assoc.* 6, 226–238. doi: 10.1007/s40881-020-00096-z

Greiner, B. (2004). *The Online Recruitment System ORSEE 2.0–A Guide for the Organization of Experiments in Economics*. Cologne: Mimeo; Department of Economics, University of Cologne.

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *J. Econ. Behav. Organ.* 117, 26–39. doi: 10.1016/j.jebo.2015.05.008

Grossman, Z., and van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *J. Eur. Econ. Assoc.* 15, 173–217. doi: 10.1093/jeea/jvw001

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not weird. *Nature* 466, 29–29. doi: 10.1038/466029a

Henry, E., and Sonntag, J. (2019). Measuring image concern. *J. Econ. Behav. Organ.* 160, 19–39. doi: 10.1016/j.jebo.2019.02.018

Kawagoe, T., and Narita, Y. (2014). Guilt aversion revisited: an experimental test of a new model. *J. Econ. Behav. Organ.* 102, 1–9. doi: 10.1016/j.jebo.2014.02.020

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games Econ. Behav.* 97, 110–119. doi: 10.1016/j.geb.2016.04.003

Kimbrough, E. O., and Vostroknutov, A. (2016). Norms make preferences social. *J. Eur. Econ. Assoc.* 14, 608–638. doi: 10.1111/jeea.12152

Krupka, E. L., and Weber, R. A. (2013). Identifying social norms using coordination games: why does dictator game sharing vary? *J. Eur. Econ. Assoc.* 11, 495–524. doi: 10.1111/jeea.12006

Kurzban, R., DeScioli, P., and O'Brien, E. (2007). Audience effects on moralistic punishment. *Evol. Hum. Behav.* 28, 75–84. doi: 10.1016/j.evolhumbehav.2006.06.001

Lacetera, N., and Macis, M. (2010). Social image concerns and prosocial behavior: field evidence from a nonlinear incentive scheme. *J. Econ. Behav. Organ.* 76, 225–237. doi: 10.1016/j.jebo.2010.08.007

Malmendier, U., te Velde, V. L., and Weber, R. A. (2014). Rethinking reciprocity. *Annu. Rev. Econ.* 6, 849–874. doi: 10.1146/annurev-economics-080213-041312

Manski, C. F., and Neri, C. (2013). First-and second-order subjective expectations in strategic decision-making: experimental evidence. *Games Econ. Behav.* 81, 232–254. doi: 10.1016/j.geb.2013.06.001

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *J. Market. Res.* 45, 633–644. doi: 10.1509/jmkr.45.6.633

Murnighan, J. K., Oesch, J. M., and Pillutla, M. (2001). Player types and self-impression management in dictatorship games: two experiments. *Games Econ. Behav.* 37, 388–414. doi: 10.1006/game.2001.0847

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. *Judgm. Decis. Mak.* 6, 771–781. doi: 10.2139/ssrn.1804189

Palfrey, T. R., and Wang, S. W. (2009). On eliciting beliefs in strategic games. *J. Econ. Behav. Organ.* 71, 98–109. doi: 10.1016/j.jebo.2009.03.025

Pelligra, V. (2011). Empathy, guilt-aversion, and patterns of reciprocity. *J. Neurosci. Psychol. Econ.* 4:161. doi: 10.1037/a0024688

Rand, D. G. (2016). Cooperation, fast and slow: meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychol. Sci.* 27, 1192–1206. doi: 10.1177/0956797616654455

Regner, T. (2018a). Reciprocity under moral wiggle room: is it a preference or a constraint? *Exp. Econ.* 21, 779–792. doi: 10.1007/s10683-017-9551-2

Regner, T. (2018b). *What's Behind Image? Towards a Better Understanding of Image-Driven Behavior*. Technical report, Jena Economics Research Papers, University of Jena.

Regner, T., and Matthey, A. (2017). *Actions and the Self: I Give, Therefore I Am*? Mimeo.

Samek, A. S., and Sheremeta, R. M. (2014). Recognizing contributors: an experiment on public goods. *Exp. Econ.* 17, 673–690. doi: 10.1007/s10683-013-9389-1

Schotter, A., and Trevino, I. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.* 6, 103–128. doi: 10.1146/annurev-economics-080213-040927

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. *Beiträge Exp. Wirtschaftsforsch.* 1, 136–168.

Tadelis, S. (2011). *The Power of Shame and the Rationality of Trust*. Berkeley, CA: Haas School of Business working paper.

Tangney, J. P. (1995). Recent advances in the empirical study of shame and guilt. *Am. Behav. Sci.* 38, 1132–1145. doi: 10.1177/0002764295038008008

Tappin, B. M., and Capraro, V. (2018). Doing good vs. avoiding bad in prosocial choice: a refined test and extension of the morality preference hypothesis. *J. Exp. Soc. Psychol.* 79, 64–70. doi: 10.1016/j.jesp.2018.06.005

Tirole, J., Falk, A., and Bénabou, R. (2016). *Narratives, Imperatives and Moral Reasoning*. Technical report, Mimeo.

van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *Am. Econ. J. Microecon.* 6, 256–264. doi: 10.1257/mic.6.3.256

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations 1. *Econometrica* 76, 1467–1480. doi: 10.3982/ECTA7673

Wolf, S. T., Cohen, T. R., Panter, A., and Insko, C. A. (2010). Shame proneness and guilt proneness: toward the further understanding of reactions to public and private transgressions. *Self Identity* 9, 337–362. doi: 10.1080/15298860903106843

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 July 2021
doi: 10.3389/fpsyg.2021.687913

On Lies and Hard Truths

Sascha Behnk[1,2†] *and Ernesto Reuben*[3,4*†]

[1] Department of Banking and Finance, University of Zurich, Zurich, Switzerland, [2] IU International University of Applied Sciences, Erfurt, Germany, [3] Social Science Division, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates, [4] Center for Behavioral Institutional Design, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

We run an experimental study using sender-receiver games to evaluate how senders' willingness to lie to others compares to their willingness to tell hard truths, i.e., promote an outcome that the sender knows is unfair to the receiver without explicitly lying. Unlike in previous work on lying when it has consequences, we do not find that antisocial behavior is less frequent when it involves lying than when it does not. In fact, we find the opposite result in the setting where there is social contact between senders and receivers, and receivers have enough information to judge whether they have been treated unfairly. In this setting, we find that senders prefer to hide behind a lie and implement the antisocial outcome by being dishonest rather than by telling the truth. These results are consistent with social image costs depending on the social proximity between senders and receivers, especially when receivers can judge the kindness of the senders' actions.

Keywords: lying, hard truth, sender-receiver games, social image, antisocial behavior

OPEN ACCESS

Edited by:
Agne Kajackaite, Social Science Research Center Berlin, Germany

Reviewed by:
Julien Benistant, UMR5229 Institut des Sciences Cognitives Marc Jeannerod, France
Andrea Albertazzi, University of Essex, United Kingdom
Tilman Fries, Social Science Research Center Berlin, Germany

***Correspondence:**
Ernesto Reuben
ereuben@nyu.edu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *30 March 2021*
Accepted: *07 June 2021*
Published: *07 July 2021*

Citation:
Behnk S and Reuben E (2021) On Lies and Hard Truths. Front. Psychol. 12:687913. doi: 10.3389/fpsyg.2021.687913

1. INTRODUCTION

An extensive body of literature has shown that individuals face psychological costs from lying to others and has identified various factors moderating these costs[1]. A crucial moderator for lying behavior are individuals' social image concerns (see, Bénabou and Tirole, 2006; Andreoni and Bernheim, 2009). For example, Khalmetski and Sliwka (2019) developed a model that predicts partial lying due to image costs in the Fischbacher and Föllmi-Heusi (2013) die-rolling paradigm. Their findings indicate that individuals with a strong reputation sensitivity cover their lies by not always lying maximally and, thus, reducing their social image costs. Other studies substantiate these findings in different versions of the die-rolling paradigm by showing that social image costs mediate lying costs (Gneezy et al., 2018; Bašic and Quercia, 2020). While this research shows that image costs provide a strong motivation not to lie, the literature has not thoroughly investigated the impact of social image costs in settings where lying has negative consequences for others, but the alternative to lying is to be honestly antisocial. In this study, we explore circumstances under which implementing an antisocial outcome through a lie can be preferred to implementing it without lying.

The seminal experimental study on the interplay of lying behavior and its consequences is Gneezy (2005). This study shows that individuals show a lower willingness to act antisocially toward another person when an action involves lying compared to when it does not[2]. To establish this

[1] See, for example, Lundquist et al. (2009), Kartik (2009), Erat and Gneezy (2012), Cappelen et al. (2013), Gibson et al. (2013), Gneezy et al. (2013), López-Pérez and Spiegelman (2013), Abeler et al. (2014), Kajackaite and Gneezy (2017), Gneezy et al. (2018), and Alempaki et al. (2019). Moreover, see Tang et al. (2018) and Abeler et al. (2019) for meta-analyses of the literature and Sobel (2020) for a discussion on the distinction between lying and deception in games.

[2] This finding was later replicated by Hurkens and Kartik (2009) using the same design. Gneezy (2005) also shows that senders react to different monetary consequences of lying for the sender and the receiver. Hurkens and Kartik (2009) identify two types of individuals in this setting, those who never lie and those who lie whenever the monetary benefit from lying is preferred over being truthful.

result, Gneezy (2005) compares decisions in a sender-receiver game with those in a dictator game. In the sender-receiver game, players face two options: one pays more to the sender while the other pays more to the receiver. Receivers pick the option that determines both players' earnings, but they have no information about the payoff structure. Their only information stems from a message sent by the sender. Unbeknownst to the receiver, senders can send either (i) a dishonest message that tricks the receiver into believing that the option that favors the sender is their best choice or (ii) a truthful message that reveals the option that favors the receiver. In the dictator game, players face the same payoff structure and information as in the sender-receiver game. However, dictators simply choose the option to determine the earnings of both players. Gneezy (2005) finds that senders send the dishonest message less frequently than dictators choose the option that favors them.

Although dictators can implement the same outcomes as senders in the sender-receiver game, these games vary in meaningful ways. First, in the dictator game, receivers are not actively involved in the decision-making. Hence, in contrast to senders, dictators are not intentionally influencing their counterpart's payoff-relevant behavior. Second, the framing of the action changes. Dictators are making a choice that directly determines payoffs, while senders are simply transmitting information. In the latter case, there is more moral wiggle room since senders can convince themselves that receivers chose to listen to them and are therefore responsible for the outcome[3]. To wit, the receiver in the sender-receiver game is arguably more salient than the receiver in the dictator game, which can imply that social image costs play a more prominent role in the former than the latter. These dissimilarities make it hard to attribute the difference between the dictators' and senders' choices solely to the fact that the senders' choice involves lying.

Instead of a dictator game, we use a modified sender-receiver game as the no-lying baseline. More specifically, in this *Hard Truth* sender-receiver game, receivers are not passive since their choice determines both players' payment. The difference is that senders can only send messages that truthfully reveal the earnings of the receiver. In other words, we allow for a similar interaction between players (information transmission) as well as active decision-making by the receiver and only vary the type of messages available to the sender. This design allows us to make a more direct evaluation of the effect of lying in otherwise identical settings.

We further study the difference in the senders' willingness to tell a lie vs. a hard truth by varying the prominence of social image costs. More specifically, in addition to the anonymous (computerized) message transmission in our *Baseline* treatment, we run a *Face to Face* treatment where senders personally deliver the message to the receiver. Although senders' identity is not revealed, social contact with the receiver presumably increases the senders' social image costs[4]. Finally, we run a *Face to Face & Information* treatment where, in addition to personal delivery of the message, receivers are fully informed of the game's payoff distribution[5]. This information introduces an interesting dimension to the game. In this treatment, there is no ambiguity of the sender's intentions as receivers know how much money they earn if the sender reveals the prosocial option or the antisocial option[6]. Therefore, the difference between a dishonest message and a hard truth is that in the latter, receivers learn whether the sender treated them unfairly the moment they receive the message. By contrast, if the message is dishonest, receivers learn whether the sender treated them unfairly (and the fact that the sender lied) later when they are told their earnings. In other words, a dishonest message allows senders to mask their actions at the moment of personal contact. If personal contact heightens the importance of social image costs, this treatment allows us to study a setting where lying might actually imply smaller image costs than telling a hard truth.

2. MATERIALS AND METHODS

2.1. Experimental Design

In the experiment, participants are randomly matched into pairs to play a sender-receiver game. In each pair, one participant is assigned the role of the *sender* and the other the role of the *receiver*.

The receiver determines both participants' earnings by choosing one of ten options. There is one prosocial option that pays €10 to each participant, one antisocial option that pays the sender €17 minus an amount $x \in [€0, €6.5]$ and €3 to the receiver, and eight Pareto-dominated options that pay €4 to the sender and €0 to the receiver. At the beginning of the game, the computer randomly labels the ten options with a unique letter ranging from A to J. Only the sender knows how each option is labeled. **Table 1** is an example of a letter assignment and how this information is presented to the sender.

The task of the sender is to transmit a message to the receiver. There are two available messages. In the *Lying* condition, the first message, Message I, accurately reveals the label of the prosocial option and reads, "Option [letter paying the receiver €10] will earn you *more money than the other options*, 10 euros." The second message, Message II, is dishonest in that it reveals the label of the antisocial option but claims it is the best option for the receiver: "Option [letter paying the receiver €3] will earn you *more money than the other options*, 3 euros." In the *Hard Truth*

[3]See Bartling and Fischbacher (2012) for evidence that delegated decisions reduce the responsibility of a decision-maker for an antisocial outcome, even when the player to which the decision is delegated has strong incentives to act as the original decision-maker intended.

[4]Conrads and Lotz (2015) find that individuals lie less in face-to-face settings than in more anonymous settings.

[5]We designed the sender-receiver games so that almost all receivers implement the option mentioned in the sender's message irrespective of the message's content and whether they are informed of the payoff structure or not. In other words, we ensure that there are no strategic reasons for senders to send a message that does not correspond to the outcome they would like to see implemented. See section 3 for details.

[6]According to Sobel (2020), while *Hard Truth* messages do not involve lying, they might be deceptive depending on the receivers' beliefs and available information. One could argue that this is the case in the *Baseline* and *Face to Face* treatments, where receivers might be lead to think that the sender is acting in their best interest, but not in the *Face to Face & Information* treatment.

TABLE 1 | Example payoff table in the sender-receiver games (amounts in euros).

Option	A	B	C	D	E	F	G	H	I	J
Sender	4	4	10	4	$17 - x$	4	4	4	4	4
Receiver	0	0	10	0	3	0	0	0	0	0

condition, Message I and Message II simply indicate the amount the receiver will earn. Namely, Message I reads "Option [letter paying the receiver €10] will earn you 10 euros," while Message II reads "Option [letter paying the receiver €3] will earn you 3 euros[7]."

Our aim with these sender-receiver games is for us to be able to inform receivers of the payoff structure while maintaining the senders' incentive to reveal their preferences (in contrast to Gneezy, 2005; see Sutter, 2009). In other words, we selected the payoffs and number of Pareto-dominated options to ensure that enough receivers follow the message for senders to have an overriding incentive to choose the message corresponding to their preferred outcome in both the *Lying* and *Hard Truth* conditions[8].

We use a 2×3 experimental design with two conditions (*Lying* and *Hard Truth*) and three treatments. In the *Baseline* treatment, receivers do not know the payoffs associated with the prosocial and antisocial options, and senders transmit their message anonymously via the computer. This treatment has a similar information structure to the sender-receiver games based on the design of Gneezy (2005). The other treatments are designed to increase the senders' image costs.

In the *Face to Face* treatment, senders deliver the message to the receiver in person. Specifically, senders were asked to write down the message they chose on a blank sheet of paper and wait for an experimenter to come to their desk. The experimenter double-checked that the written message corresponded to the chosen message and then guided the sender to the receiver's desk. The sender handed the sheet over to the receiver and returned to his/her seat. During the delivery process, the experimenter ensured that there was no other communication between senders and receivers.

In the *Face to Face & Information* treatment, in addition to the personal message delivery, the receiver is informed in the instructions of the payoffs available in the 10 options (but stays blind regarding how the computer labels each option)[9]. Note that, since receivers know the payoff structure, we cannot use the same messages as in other treatments because a message stating that an option "will earn you *more money than the other options*, 3 euros" can be immediately identified as a lie during the message delivery. For this reason, we slightly change the wording of the messages of the *Lying* condition. Specifically, Message I reads "Option [letter paying the receiver €10] will earn you 10 euros," while Message II reads "Option [letter paying the receiver €3] will earn you 10 euros[10]."

We use the strategy method to measure precisely the senders' willingness to send an antisocial message. Specifically, senders choose between Message I and Message II in each of the 14 rows in **Table 2**. After that, the computer randomly selects one row to determine which message is sent. When receivers see the message, they are not informed of which row was selected by the computer. While Message I always pays €10, the payoff from Message II equals €17 min the amount x, which we systematically vary from €0 to €6.5 in steps of €0.5. Based on the value of x at which a sender switches from Message II to Message I, we can calculate the minimum monetary compensation senders must receive to send the antisocial message instead of the prosocial message. In other words, the monetary equivalent of the psychological cost borne by a sender for acting antisocially. Accordingly, we call this minimum compensation the senders' *antisocial cost*. More specifically, senders who choose Message I for all $x > c$ are classified as having an antisocial cost equal to $€6.75 - c$ (i.e., the midpoint of the interval $[€7 - c, €6.5 - c]$)[11].

2.2. Procedures

We ran the experiment between February and June 2015 at the Laboratory of Experimental Economics (LEE) at University Jaume I in Castellon, Spain, with 240 undergraduate students comprising 121 men and 119 women from different faculties.

[7]These messages are based on those used by Gneezy (2005). In that paper, the lying message was "Option B will earn you more money than option A," when in fact, A paid the receiver more than B. Since we had more than two options, we used "other options" instead of "option A." Moreover, we added the amount in euros to make the message in the *Lying* condition comparable to that in the *Hard Truth* condition.

[8]Specifically, we chose monetary payoffs so that senders have a strict incentive to send the message corresponding to their preferred outcome as long as they expect more than 10% of the receivers to follow their message. To see this, denote the sender's utility if the antisocial option is implemented as $U(A)$, her utility if the prosocial option is implemented as $U(P)$, and her utility if a dominated option is implemented as $U(D)$. Furthermore, let $b \in [0, 1]$ be the sender's belief that the receiver follows the message. In this case, the sender's expected utility of sending Message I is $bU(P) + (1 - b)(1/9)U(A) + (1 - b)(8/9)U(D)$, and that of sending Message II is $bU(A) + (1 - b)(1/9)U(P) + (1 - b)(8/9)U(D)$. It follows that as long as $b > 1/10$, senders will choose Message I if $U(P) > U(A)$ and Message II if $U(P) < U(A)$. Note that this condition holds for 97.4% of the senders in our dataset.

[9]In all treatments, it is common knowledge that a message always reveals the label of either the prosocial or the antisocial option and never the label of one of the Pareto-dominated options.

[10]This difference implies a slight change in the nature of the lie between treatments. While in *Baseline* and *Face to Face* the sender lies about an option paying the receiver "more money than the other options," in *Face to Face & Information*, the sender lies about the stated amount "10 euros," which the receiver knows would pay more than other options. An alternative experimental design would be to use the messages from *Face to Face & Information* in all treatments. However, that would make those treatments less comparable to Gneezy (2005), which is why we opted for our current design.

[11]At the extremes, senders who always choose Message I are classified as having an antisocial cost equal to €7.25 and senders who always choose Message II as having equal to €0.25. Senders who switched more than once or switched from Message I to Message II.

TABLE 2 | Senders' choice lists (amounts in euros).

Row	Payoff of Message I	*x*	Payoff of Message II
1	10.00	0.00	17.00
2	10.00	0.50	16.50
3	10.00	1.00	16.00
4	10.00	1.50	15.50
5	10.00	2.00	15.00
6	10.00	2.50	14.50
7	10.00	3.00	14.00
8	10.00	3.50	13.50
9	10.00	4.00	13.00
10	10.00	4.50	12.50
11	10.00	5.00	12.50
12	10.00	5.50	11.00
13	10.00	6.00	11.00
14	10.00	6.50	10.50

Participants were recruited using ORSEE (Greiner, 2015). We conducted 12 sessions, each lasting around 1.5 h[12].

Upon arrival, participants were randomly assigned to computers. After that, the instructions for the experiment were read aloud by the experimenter, and participants were asked to answer a series of control questions (a sample of the instructions is available in the **Supplementary Material**). Participants could ask questions at any point. The experiment was conducted using z-Tree (Fischbacher, 2007).

Once senders chose a message for each of the 14 values of *x* (see **Table 2**), the computer randomly selected one of these values and displayed the text of the chosen message on the senders' screen. In the *Face to Face* and *Face to Face & Information* treatments, senders wrote down the message on a sheet of paper and walked over with an experimenter to hand the message over to the receiver. All participants were informed about the delivery process and knew that communication with other participants was forbidden. Once all senders returned to their desks, receivers were asked to type into the computer screen the message they received and choose one of the 10 options.

In addition, we elicited the senders' belief concerning the likelihood that receivers implement the message they receive. Specifically, after the senders delivered their chosen message but before they learned the final outcome, we asked them to indicate "out of 10 Players 2 [the receivers], how many will follow the message they received?" Senders were paid €0.25 for a correct guess[13].

After the experiment ended, participants were paid in cash. Average earnings were around €15, including belief elicitation and a €5 show-up fee.

2.3. Expected Behavior

In line with the literature, we expect to find similar results to Gneezy (2005) in the *Baseline* treatment. Namely, a lower willingness to choose the antisocial message when the message is dishonest than when it is truthful, implying that there are costs to lying. In other words, we expect that the senders' mean antisocial cost is higher in *Lying* than in *Hard Truth*.

The remaining two treatments allow us to test the effects of increasing social image costs on lying and transmitting hard truths. We first introduce social image costs due to the personal delivery of the message in the *Face to Face* treatment, where senders of antisocial messages have to face the receiver in person. In the *Face to Face & Information* treatment, we further increase social image costs because receivers are fully aware of the message's nature and, thus, of the sender's intentions when the message is personally delivered.

The literature shows that social image costs affect behavior in situations with lying (e.g., Gneezy et al., 2018; Bašic and Quercia, 2020) as well as without lying (for social image effects in dictator games see, e.g., Andreoni and Bernheim, 2009; Rigdon et al., 2009; Ockenfels and Werner, 2012)[14]. However, previous work is silent on whether these image costs are greater with or without lying. If the appearance of being dishonest produces larger image costs than that of being willing to transmit a hard truth, then the gap between the *Lying* condition and the *Hard Truth* condition would grow as we move from *Baseline* to *Face to Face*, where the mere physical contact with the receiver might trigger social image concerns, and then to *Face to Face & Information*, where the receiver can also evaluate the actions of the sender. Conversely, if the image costs are stronger in the *Hard Truth* condition than the *Lying* condition, then we would see the treatment differences narrow.

3. RESULTS

Our sample consists of 120 receivers and 114 senders: 57 senders in the *Hard Truth* condition (19 senders in each of the three treatments) and 57 senders in the *Lying* condition (19 senders in *Baseline*, 18 in *Face to Face*, and 20 in *Face to Face & Information*)[15]. Descriptive statistics of the main variables per treatment and condition are shown in **Supplementary Table 1**. We estimate the sample average treatment effects using OLS regressions with robust standard errors. The dependent variable is senders' antisocial cost in section 3.1 and the senders'

[12]Data from the *Face to Face* treatments are also used in Behnk et al. (2019). Data from the *Baseline* and *Face to Face & Information* treatments are exclusively used in this study.

[13]We also elicited participants' normative views and their beliefs about normative views of others. Furthermore, we elicited the receivers' expected fraction of antisocial messages. A rigorous analysis of these variables in the *Face to Face* treatment is reported in Behnk et al. (2019).

[14]Importantly, this research has demonstrated that social image costs can be triggered even in anonymous settings where, logically, their social image should not be a concern (Gneezy et al., 2018). Another interpretation of these results is that people also care about self-image. In other words, they want to signal to themselves through their actions that they are a prosocial individual (Bénabou and Tirole, 2006).

[15]Of the 120 senders, six senders switched more than once between Message I and Message II in the choice list. Since it is not clear what these participants' antisocial cost is, we dropped them from the statistical analysis.

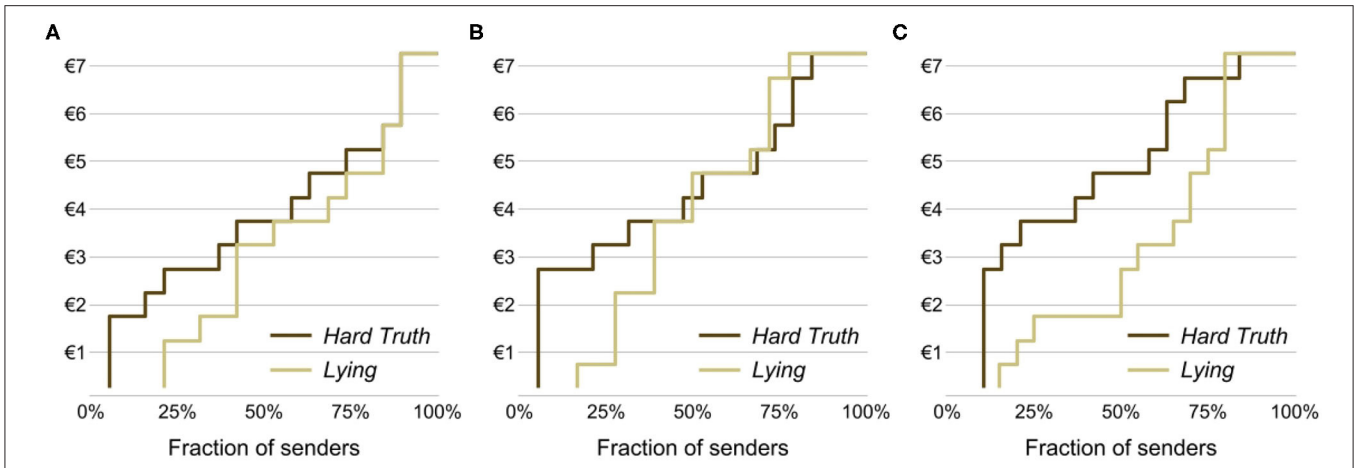


FIGURE 1 | Cumulative distributions of senders' antisocial cost depending on the condition and treatment. **(A)** Baseline. **(B)** Face to Face. **(C)** Face to Face & Information.

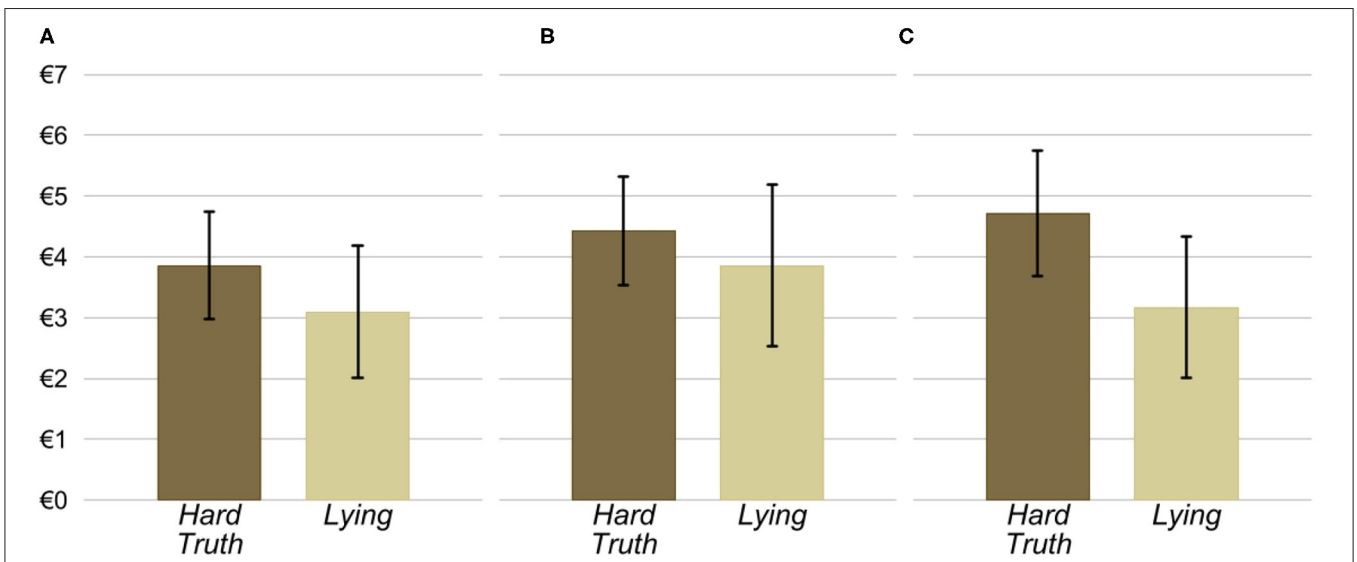


FIGURE 2 | Senders' average antisocial cost and 95% confidence intervals depending on the condition and treatment. **(A)** Baseline. **(B)** Face to Face. **(C)** Face to Face & Information.

beliefs about the likelihood that receivers follow the message in section 3.2. The independent variables correspond to treatment and condition dummy variables. The regressions are found in **Supplementary Table 2**. In addition, we report the results of non-parametric tests. All reported *p*-values are based on two-sided tests.

3.1. Senders' Antisocial Cost

Figure 1 depicts the cumulative distributions of the senders' antisocial cost in the *Lying* and *Hard Truth* conditions across the three treatments. **Figure 2** shows the senders' average antisocial cost in the two conditions by treatment. These figures suggest that senders are more willing to lie to the receiver than to transmit a hard truth. In fact, pooling observations across the three treatments, we find that the average antisocial cost in the *Lying* condition, €3.36, is significantly lower than the average antisocial cost in the *Hard Truth* condition, €4.34 ($p = 0.021$). The mean difference between conditions is substantial as it corresponds to 0.43 standard deviations[16].

Next, we look at each treatment separately. In the *Baseline* treatment, we find that, contrary to our expectations, the average antisocial cost is lower in *Lying* than in *Hard Truth* by €0.77 or 0.37 standard deviations. Albeit this difference is not statistically significant ($p = 0.257$). In other words, we do not find evidence that lying induces an additional cost over the cost of acting truthfully but antisocially.

[16]The *p*-values of comparing the distribution of antisocial costs across the two conditions using Wilcoxon-Mann-Whitney *U*-tests are as follows: $p = 0.026$ pooling across treatments; $p = 0.259$ in *Baseline*; $p = 0.612$ in *Face to Face*; $p = 0.058$ in *Face to Face & Information*.

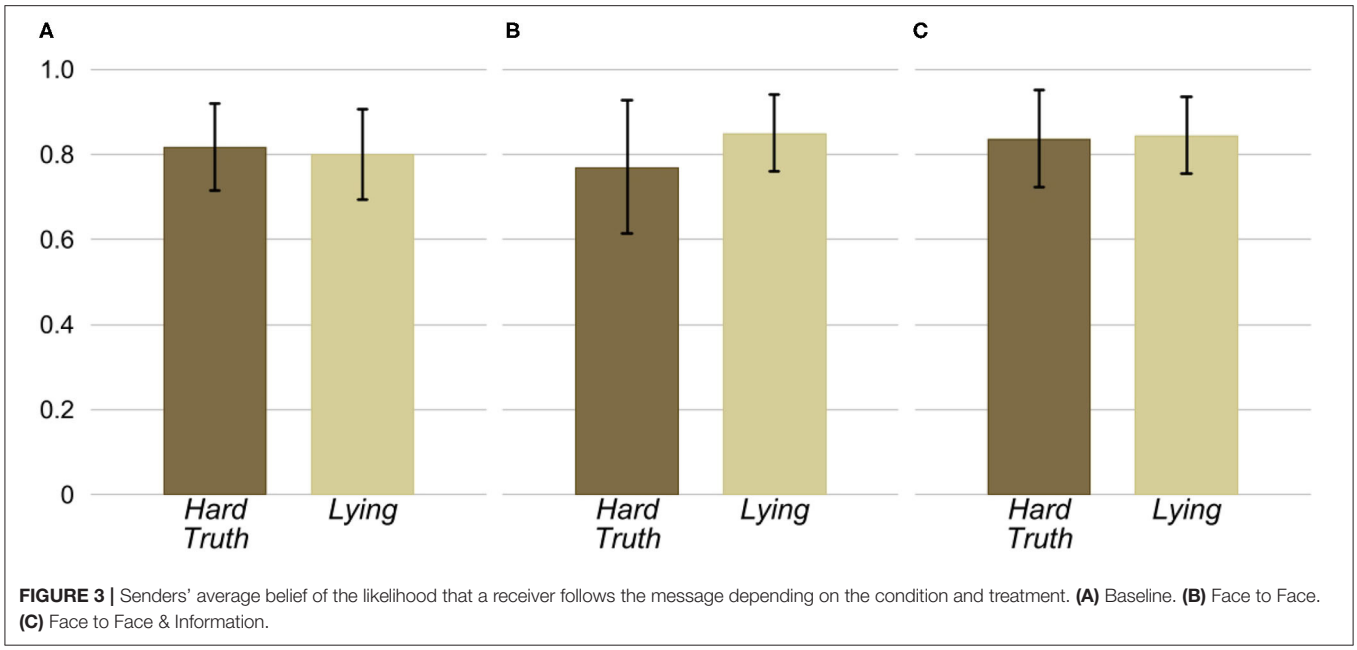


FIGURE 3 | Senders' average belief of the likelihood that a receiver follows the message depending on the condition and treatment. **(A)** Baseline. **(B)** Face to Face. **(C)** Face to Face & Information.

We find a similar result in the *Face to Face* treatment. Namely, a lower average antisocial cost in *Lying* compared to *Hard Truth*. As above, the difference between the two conditions, €0.57 or 0.25 standard deviations, is not statistically significant ($p = 0.453$).

Lastly, we look at the *Face to Face & Information* treatment, where social image costs are presumably highest. As in the other treatments, average antisocial costs are lower in *Lying* than in *Hard Truth*. Unlike the other treatments, at €1.55 or 0.64 standard deviations, this difference is noticeably bigger and statistically significant ($p = 0.040$).

3.2. Senders' Beliefs

One explanation for the lower willingness to send hard truths than dishonest messages is that senders expect a considerably lower fraction of receivers will follow the message they receive in the *Hard Truth* condition compared to the *Lying* condition. To explore this explanation, we analyze the senders' beliefs about the likelihood that receivers follow the message they receive. The senders' average belief for each condition and treatment is depicted in **Figure 3**[17]. The figure shows that the average belief is not substantially different across conditions in any of the treatments. Consistent with this observation, we do not find statistically significant differences in the senders' beliefs between the *Hard Truth* and *Lying* conditions in any of the three treatments ($p > 0.353$)[18].

To further check whether the senders' beliefs explain the difference between conditions, we ran additional OLS regressions with the senders' antisocial cost as the dependent variable. As independent variables, we include a dummy variable equal to one if the sender is in the *Lying* condition (and zero otherwise) and the senders' belief (i.e., the fraction of receivers they expect will follow the message). **Table 3** contains the regression's estimated coefficients pooling the data from all treatments as well as for each treatment separately. Also, as an additional robustness check, the table includes regressions where we also control for the senders' demographic characteristics (i.e., their gender and age). Overall, the senders' beliefs do not explain the difference between *Lying* and *Hard Truth*[19].

4. CONCLUSIONS

We investigate under which circumstances an antisocial action that involves a lie could be preferred over an otherwise identical antisocial action that is truthful. We use a series of sender-receiver games in which senders implement a prosocial or an antisocial outcome by sending a prosocial or antisocial message to the receiver. In one condition, the antisocial message involves lying to the receiver, while in the other, the message is truthful. Furthermore, we systematically vary the conditions of the message delivery to vary the social image costs of the sender.

Overall, we do not find evidence in any treatment that lying entails psychological costs above those of acting antisocially. In fact, in the treatment with the highest social image costs, the *Face to Face & Information* treatment, we find the opposite. Senders prefer to implement the antisocial outcome by lying

[17]The actual fraction of receivers who follow the message they receive equals 0.98 in *Lying* (0.95 in *Baseline*, 1.00 in *Face to Face*, and 1.00 in *Face to Face & Information*) and 0.84 in *Hard Truth* (0.90 in *Baseline*, 0.84 in *Face to Face*, and 0.77 in *Face to Face & Information*). Hence, senders' are somewhat pessimistic about the receivers following rate.

[18]The p-values of comparing the distribution of the senders' beliefs across the two conditions using Wilcoxon-Mann-Whitney U-tests are as follows: $p = 0.919$ pooling across treatments; $p = 0.714$ in *Baseline*; $p = 0.895$ in *Face to Face*; $p = 0.999$ in *Face to Face & Information*.

[19]The senders' beliefs are not statistically significant in any of the regressions. This result is to be expected given that most beliefs are relatively high, and a very low belief is required for it to be relevant to the sender's choice (see Footnote 8).

TABLE 3 | Regressions of the senders' antisocial cost on the condition and the senders' belief.

	All treatments		Baseline		Face to Face		Face to Face & Information	
Deception condition	–0.98*	–1.02*	–0.80	–0.78	–0.74	–0.61	–1.55*	–1.76*
	(0.42)	(0.42)	(0.67)	(0.68)	(0.75)	(0.74)	(0.76)	(0.78)
Sender's belief	0.35	0.50	–1.90	–1.97	2.03	2.31	-0.01	-0.11
	(0.91)	(0.92)	(1.57)	(1.62)	(1.44)	(1.41)	(1.79)	(1.82)
Demographic controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	114	114	38	38	37	37	39	39

*OLS estimates and standard errors in parentheses. *Indicates statistical significance at 5%.*

rather than by telling the truth. However, we should note that a potential caveat to this last result is the statistical power of this comparison. An *ex-post* power analysis using the observed means and standard deviations shows that the average treatment effect across the *Lying* and *Hard Truth* conditions in the *Face to Face & Information* treatment has a power of 0.52 for a significance level of 5%. Therefore, it would be premature to conclude that the psychological costs of lying are *lower* than those of telling a hard truth. Future work ought to gather more evidence to substantiate this effect. Having said that, the fact that in all three treatments, the senders' antisocial costs of implementing the antisocial outcome by lying are never higher than those of implementing the same outcome with a truthful message shows more convincingly that the willingness to lie is sensitive to the image costs of the no-lying alternative.

We think that our experiment highlights the need to understand the impact of social image costs on different decisions. In settings where actions have consequences for others, social image costs are present irrespective of whether the antisocial action involves lying or not. Hence, the social image cost of being perceived as dishonest needs to be compared to the social image cost of being perceived as someone willing to deliver hard or uncomfortable truths. Our results suggest that the discomfort experienced when delivering an antisocial message in person when the recipient can immediately interpret the message's content is higher than that of eventually being perceived as dishonest.

Our setup suggests that it is important to consider the timing of social contact and the moment when others learn the nature of one's actions, which is when they can judge them as good or bad. The personal delivery of the message when receivers are fully informed implies that an antisocial truthful message can be judged as bad at the moment of social contact. This simultaneity could make social image costs more salient. By contrast, a dishonest antisocial message will not be judged immediately but later on when the receiver learns the implemented message's outcome. This separation in time allows the sender to "hide behind the lie" at the moment of social contact. Therefore, even if the sender knows that the message will eventually be revealed as a lie, the social image cost of appearing dishonest occurs at a point where social image costs are likely to be less salient. We think this last result merits further study. We find that the antisocial costs of lying are substantially lower than those of telling a hard truth in the *Face to Face & Information* treatment, which supports this interpretation. However, we also find a smaller difference in the same direction in the *Face to Face* treatment[20]. Given that in the *Face to Face* the hard truth message does not reveal one's intentions, there can be reasons other than "hiding" to prefer a lie over a hard truth.

DATA AVAILABILITY STATEMENT

The data for this study is available at https://doi.org/10.3886/E143161V1. Replication materials are included in the **Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by IRB Columbia University (Protocol IRB-AAAO9551). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

ER gratefully recognizes financial support by Tamkeen under the NYU Abu Dhabi Research Institute Award CG005.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.687913/full#supplementary-material

[19] A difference-in-difference test of the average effect of the condition results in an insignificant result ($p = 0.362$).

REFERENCES

Abeler, J., Becker, A., and Falk, A. (2014). Representative evidence on lying costs. *J. Public Economics* 113, 96–104. doi: 10.1016/j.jpubeco.2014.01.005

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth–Telling. *Econometrica* 87, 1115–1153. doi: 10.3982/ECTA14673

Alempaki, D., Doğan, G., and Saccardo, S. (2019). Deception and reciprocity. *Exp. Econ.* 22, 980–1001. doi: 10.1007/s10683-018-09599-3

Andreoni, J., and Bernheim, B. D. (2009). Social image and the 50-50 norm: a theoretical and experimental analysis of audience effects. *Econometrica* 77, 1607–1636. doi: 10.3982/ECTA7384

Bartling, B., and Fischbacher, U. (2012). Shifting the Blame: On Delegation and Responsibility. *Rev. Econ. Stud.* 79, 67–87. doi: 10.1093/restud/rdr023

Bašic, Z., and Quercia, S. (2020). *The Influence of Self and Social Image Concerns on Lying*. Discussion Paper 2020-18, Max Planck Institute for Research on Collective Goods.

Behnk, S., Hao, L., and Reuben, E. (2019). *Shifting Normative Views: On Why Groups Behave More Antisocially Than Individuals*. Working paper, New York Univeristy Abu Dhabi.

Bénabou, R., and Tirole, J. (2006). Incentives and prosocial behavior. *Am. Econ. Rev.* 96, 1652–1678. doi: 10.1257/aer.96.5.1652

Cappelen, A. W., Sørensen, E. T., and Tungodden, B. (2013). When do we lie? *J. Econ. Behav. Organ.* 93, 258–265. doi: 10.1016/j.jebo.2013.03.037

Conrads, J., and Lotz, S. (2015). The effect of communication channels on dishonest behavior. *J. Behav. Exp. Econ.* 58, 88–93. doi: 10.1016/j.socec.2015.06.006

Erat, S., and Gneezy, U. (2012). White lies. *Manag. Sci.* 58, 723–733. doi: 10.1287/mnsc.1110.1449

Fischbacher, U. (2007). z-Tree: zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10, 171–178. doi: 10.1007/s10683-006-9159-4

Fischbacher, U., and Föllmi-Heusi, F. (2013). Lies in disguise-an experimental study on cheating. *J. Eur. Econ. Assoc.* 11, 525–547. doi: 10.1111/jeea.12014

Gibson, R., Tanner, C., and Wagner, A. F. (2013). Preferences for truthfulness: heterogeneity among and within individuals. *Am. Econ. Rev.* 103, 532–548. doi: 10.1257/aer.103.1.532

Gneezy, U. (2005). Deception: the Role of Consequences. *Am. Econ. Rev.* 95, 384–394. doi: 10.1257/0002828053828662

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *Am. Econ. Rev.* 108, 419–453. doi: 10.1257/aer.20161553

Gneezy, U., Rockenbach, B., and Serra-Garcia, M. (2013). Measuring lying aversion. *J. Econ. Behav. Organ.* 93, 293–300. doi: 10.1016/j.jebo.2013.03.025

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *J. Econ. Sci. Assoc.* 1, 114–125. doi: 10.1007/s40881-015-0004-4

Hurkens, S., and Kartik, N. (2009). Would i lie to you? on social preferences and lying aversion. *Exp. Econ.* 12, 180–192. doi: 10.1007/s10683-008-9208-2

Kajackaite, A., and Gneezy, U. (2017). Incentives and cheating. *Games Econ. Behav.* 102, 433–444. doi: 10.1016/j.geb.2017.01.015

Kartik, N. (2009). Strategic communication with lying costs. *Rev. Econ. Stud.* 76, 1359–1395. doi: 10.1111/j.1467-937X.2009.00559.x

Khalmetski, K., and Sliwka, D. (2019). Disguising lies-image concerns and partial lying in cheating games. *Am. Econ. J.* 11, 79–110. doi: 10.1257/mic.20170193

López-Pérez, R. and Spiegelman, E. (2013). Why do people tell the truth? Experimental evidence for pure lie aversion. *Exp. Econ.* 16, 233–247. doi: 10.1007/s10683-012-9324-x

Lundquist, T., Ellingsen, T., Gribbe, E., and Johannesson, M. (2009). The aversion to lying. *J. Econ. Behav. Organ.* 70, 81–92. doi: 10.1016/j.jebo.2009.02.010

Ockenfels, A., and Werner, P. (2012). 'Hiding behind a small cake' in a newspaper dictator game. *J. Econ. Behav. Organ.* 82, 82–85. doi: 10.1016/j.jebo.2011.12.008

Rigdon, M., Ishii, K., Watabe, M., and Kitayama, S. (2009). Minimal social cues in the dictator game. *J. Econ. Psychol.* 30, 358–367. doi: 10.1016/j.joep.2009.02.002

Sobel, J. (2020). Lying and Deception in Games. *J. Pol. Econ.* 128, 907–947. doi: 10.1086/704754

Sutter, M. (2009). Deception through telling the truth?! experimental evidence from individuals and teams. *Econ. J.* 119, 47–60. doi: 10.1111/j.1468-0297.2008.02205.x

Tang, H., Wang, S., Liang, Z., Sinnott-Armstrong, W., Su, S., and Liu, C. (2018). Are proselfs more deceptive and hypocritical? social image concerns in appearing fair. *Front. Psychol.* 9:2268. doi: 10.3389/fpsyg.2018.02268

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 July 2021
doi: 10.3389/fpsyg.2021.684755

Collective Honesty? Experimental Evidence on the Effectiveness of Honesty Nudging for Teams

Yuri Dunaiev[1] *and Menusch Khadjavi*[2,3,4]*

[1] Independent Researcher, Frankfurt, Germany, [2] Department of Spatial Economics, School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [3] Tinbergen Institute, Amsterdam, Netherlands, [4] Kiel Institute for the World Economy, Kiel, Germany

A growing literature in economics studies ethical behavior and honesty, as it is imperative for functioning societies in a world of incomplete information and contracts. A majority of studies found more pronounced dishonesty among teams compared to individuals. Scholars identified certain nudges as effective and cost-neutral measures to curb individuals' dishonesty, yet little is known about the effectiveness of such nudges for teams. We replicate a seminal nudge treatment effect, signing on the top of a reporting form vs. no signature, with individuals and confirm the original nudge treatment effect. We further ran the same experiment with teams of two that have to make a joint reporting decision. Our results show the effectiveness of the nudge for teams, which provides further confidence in the applicability of the nudge.

Keywords: honesty, lying, nudge, team, experiment

OPEN ACCESS

Edited by:
Nora Szech, Karlsruhe Institute of Technology (KIT), Germany

Reviewed by:
Julian Leslie, Ulster University, United Kingdom
Julien Benistant, UMR5229 Institut des Sciences Cognitives Marc Jeannerod, France

***Correspondence:**
Menusch Khadjavi
m.khadjavipour@vu.nl

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *23 March 2021*
Accepted: *15 June 2021*
Published: *08 July 2021*

Citation:
Dunaiev Y and Khadjavi M (2021) Collective Honesty? Experimental Evidence on the Effectiveness of Honesty Nudging for Teams. Front. Psychol. 12:684755. doi: 10.3389/fpsyg.2021.684755

INTRODUCTION

The subject of dishonesty and deception is undergoing intense study and arouses high concerns in the society, attracting much attention of policymakers and researchers from the fields of behavioral economics and psychology (e.g., Rosenbaum et al., 2014; Abeler et al., 2019; Gerlach et al., 2019; Köbis et al., 2019). Beyond ethical considerations, the economic harm caused by dishonesty is tremendous. The Association of Certified Examiners estimates that the typical firm losses are about 5% of revenues to occupational fraud each year, which translates into a loss of $3.6 billion at the global level (ACFE, 2020). Recent examples show that practices such as manipulation of financial and audit reports and fraudulent accounting methods are a major problem. Among convicted companies are big names such as Enron, Lehman Brothers, Madoff Investment Securities, and Parmalat. Other famous fraudulent practices are spying (Hewlett-Packard), violations of safety regulations (Southwest Airlines), and concealing emission levels (Volkswagen). In all of these fraud cases it was not a single individual who made the decision and guarded misconduct from coming to light, but teams of individuals who deceived in a conspirative manner.

Since Thaler and Sunstein (2009) introduced the concept of nudging to a larger audience, a number of experiments from psychology and economics have shown that certain nudges can work to reduce *individual* dishonesty (e.g., Mazar et al., 2008; Shu et al., 2012; Fellner et al., 2013[1]). A related literature on individual vs. team (dis)honesty developed contemporaneously and suggests that teams are often more dishonest than individuals (e.g., Cohen et al., 2009; Sutter, 2009; Danilov et al., 2013; Mühlheußer et al., 2015; Weisel and Shalvi, 2015; Korbel, 2017; Wouda et al., 2017; Kocher et al., 2018; Dannenberg and Khachatryan, 2020)[2]. The mechanisms that cause teams to be more dishonest include greater sophistication regarding the consequences of lying (Cohen et al., 2009; Sutter, 2009) and diffusion of responsibility regarding the moral misconduct of lying (Kocher et al., 2018)[3].

As dishonesty levels and mechanisms differ between individuals and teams, we regard it as a natural question whether nudges that are able to curb individual dishonesty remain effective for teams. In this paper we answer this question by employing the well-established math puzzle task paradigm and honesty nudge of Shu et al. (2012)[4]. To this end, we test whether we are able to replicate one of the treatment effects of Shu et al. (2012)—asking decision makers to sign that they will report honestly at the top of a reporting form compared to a no-signature control treatment. We ran the experiment for individuals and for teams to test for the robustness of this nudge.

Our experiment indeed successfully replicates the treatment effect of Shu et al. (2012) for individuals, adding further evidence that signing on top of the form can decrease dishonesty (compared to the no signature condition). For teams we find the same treatment effect, which shows further robustness of this nudge. The nudge seems to be able to work against the team dishonesty drivers like the diffusion of responsibility. We regard our finding as good news for policy makers who seek to employ such nudges as a tool for low-cost and effective anti-fraud and anti-corruption measures.

This paper proceeds as follows. In second section we provide the details of the experimental design, hypotheses and procedures. Third section presents the results and fourth section concludes.

[1]Note that there is a replication discussion around Mazar et al. (2008): see also Amir et al. (2018) and Verschuere et al. (2018). Verschuere et al. (2018) report one of the results of Mazar et al. (2008) does not replicate based on a meta-analysis with more than 5,000 participants. Amir et al. (2018) reply to Verschuere et al. (2018) and discuss conceptual challenges with direct replication studies.

[2]There is also a broader literature that compares economic decisions of individuals and teams, e.g., Bornstein et al. (2004), Charness and Sutter (2012), and Kugler et al. (2012).

[3]Regarding the diffusion of responsibility and ethical behavior, see also Falk and Szech (2013) and Falk et al. (2020).

[4]There are several treatments in Shu et al. (2012): Note that Kristal et al. (2020) report that the top-vs.-bottom-signature treatment effect of Shu et al. (2012) does *not* replicate for individuals. This is not the treatment effect we aim to replicate in this paper—we concentrate on the top-signature versus no-signature comparison. In the task participants need to find two numbers in a 4×3 table that sum to a specific number. In Shu et al. (2012), Mazar et al. (2008) and in our experiment this number is 10.

EXPERIMENTAL DESIGN

In this section we explain the details of the math puzzle (or matrix) task and the treatments we employed. We subsequently relate our treatments to hypotheses that originate from the current literature on lying of individuals and teams and finally provide information about the procedures of the experiment.

The math puzzle (matrix) task comprised sheets of paper with math puzzles (matrices) where two numbers sum exactly to a specific target number that is defined beforehand. In the case of Shu et al. (2012) and our experiment, each puzzle consisted of 12 three-digit numbers (with two decimal digits) of which two numbers sum exactly to the number 10. The task was to identify these two numbers and circle them in order to "solve" the respective puzzle. Each correctly solved puzzle yielded a piece-rate income, in our experiment 0.50 EUR. In the treatments with individuals (teams) we provided one (two) sheets of paper, with 20 puzzles per sheet of paper. Hence, a maximum of 10 EUR could be earned per participant in this task. Teammates could choose to work on each sheet separately or together. The time limit was strictly set to 5 min and stopped with a stop-clock. We calibrated the time limit to ensure that the solved puzzles are well-distributed between 0 and 20. Participants were asked to sum the score at the bottom of the puzzle sheet. **Figure 1** shows a complete sheet as used in our experiment.

If the number of correctly solved puzzles (or matrix exercises), i.e., the true score, is common knowledge, then it is straightforward for the researcher who conducts the experiment to multiply this score with 0.50 EUR and pay out the individual or team accordingly. If the true score is private knowledge of the individual or team, then it becomes interesting to investigate under which circumstances there is correct or elevated reporting of the true score.

In order to create a scenario in which participants would feel comfortable to over-report their score, we closely followed the procedure of Shu et al. (2011)—a study by three of the five authors of Shu et al. (2012) whose treatment effect we aim to replicate. We asked participants to dispose of the matrix paper sheet by inserting it into a paper shredder. The shredder was prepared in a way that the sheet would be partly shredded at the sides, but remain intact to retrace the scores. This incomplete shredding was not visible to participants, as the sheets moved through the shredder into a non-transparent bin. Note that for this replication approach we followed procedures of Shu et al. (2011) closely, which falls into a gray area of omitted information as categorized by Charness et al. (2021). While the scenario is suggestive of sheets being destroyed, we neither commented on sheets being destroyed nor did we indicate that we would not have a look at sheets after the sessions. This gave us the chance to learn the true score of all individuals and teams after the sessions and link them to the reported scores.

For score reporting we used the participation receipt (see **Figure 2**). The receipt included reporting the score, guessing the average score of others in the session (not incentivized),

8.08	2.00	2.55	0.87	9.08	9.53	4.95	1.98	3.79	5.25	3.37	6.43
4.87	4.51	3.83	4.55	8.51	7.35	1.26	5.15	8.09	4.68	8.54	6.75
2.29	7.20	2.95	8.27	1.96	0.47	6.74	4.54	4.33	4.06	4.87	7.39
0.78	9.74	7.05	6.66	3.85	3.52	2.55	1.58	6.21	4.41	3.48	5.32

5.44	5.56	7.46	8.66	0.05	7.95	3.03	5.22	8.61	2.54	5.13	4.75
4.93	2.54	7.29	8.29	1.73	3.05	3.82	7.57	8.99	3.84	8.81	7.12
4.22	3.23	0.14	1.17	2.28	1.91	7.87	6.18	8.44	8.12	5.85	2.86
6.26	5.73	9.96	0.57	8.83	4.54	8.06	1.62	8.82	5.19	4.87	6.41

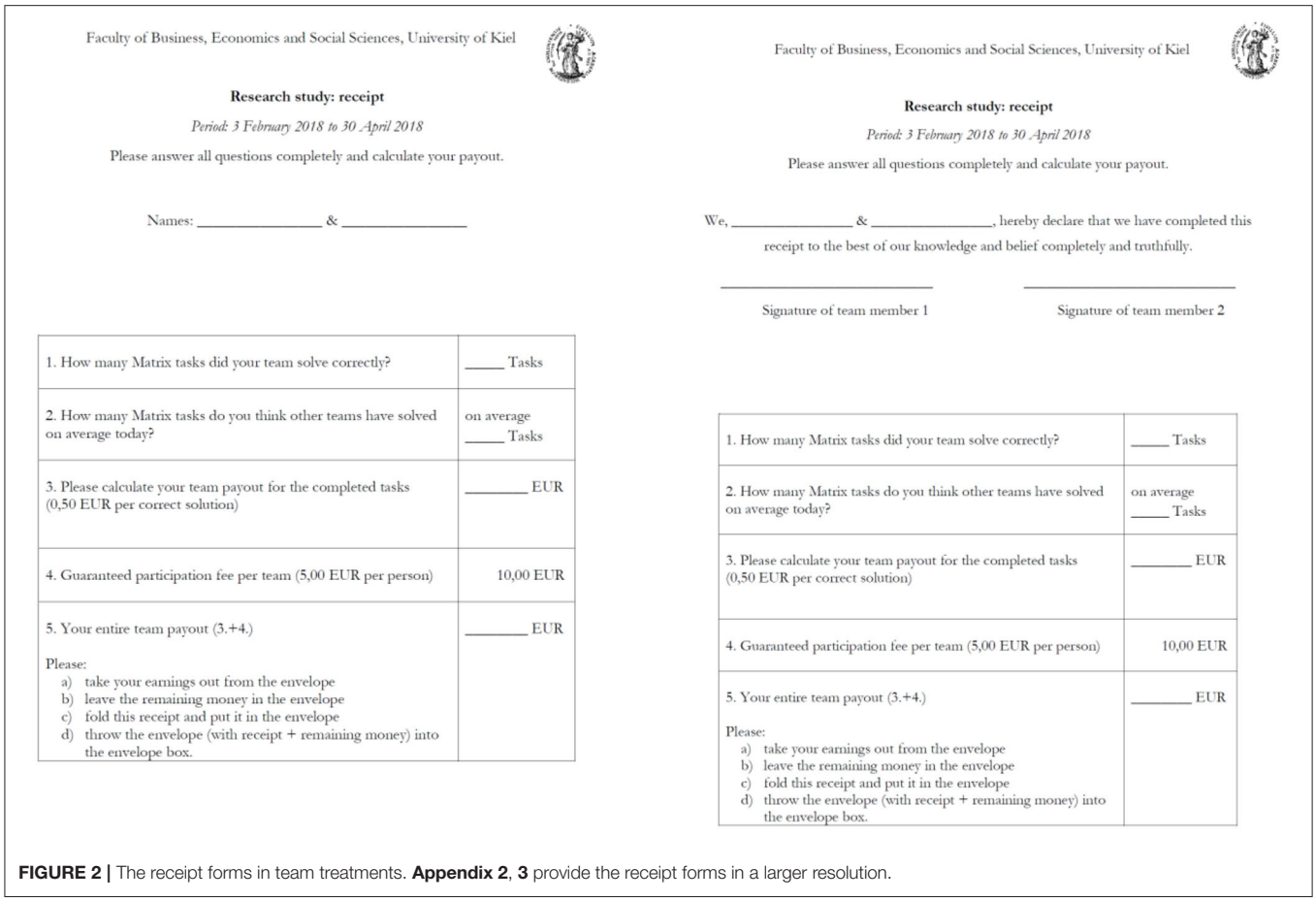
0.31	7.63	5.39	5.65	9.76	9.61	7.99	8.02	6.22	4.68	7.96	8.08
9.39	5.73	9.63	8.75	2.16	5.40	9.81	2.01	7.62	1.94	8.73	2.22
2.37	3.41	8.61	3.49	5.25	8.56	3.84	2.63	7.68	2.32	7.67	2.79
4.69	5.64	9.27	1.39	5.73	4.27	4.88	6.22	8.99	1.29	7.21	3.92

4.42	1.28	1.98	9.97	4.45	7.64	3.25	0.65	7.50	6.91	2.15	2.78
8.72	4.25	7.29	2.41	1.03	5.86	8.76	5.60	8.98	7.12	7.91	5.25
9.02	0.28	8.55	7.42	2.26	8.83	5.12	0.74	3.01	0.38	9.57	5.73
6.42	4.58	8.69	2.22	8.65	7.78	6.99	0.58	3.50	2.09	1.85	0.22

5.92	5.32	4.03	9.20	1.80	3.33	2.25	0.28	8.72	1.14	8.13	7.33
3.48	6.69	7.97	7.41	9.33	7.74	7.34	7.85	6.61	2.47	2.86	2.06
2.56	1.50	9.87	1.20	0.67	4.43	3.39	7.29	6.61	7.94	1.81	8.05
4.68	0.36	5.15	7.02	4.59	6.19	6.97	4.95	7.79	9.86	6.36	1.67

Number of correctly solved matrix-exercises: _____

FIGURE 1 | A complete math puzzle sheet (original is in A4 format).



Faculty of Business, Economics and Social Sciences, University of Kiel

Research study: receipt

Period: 3 February 2018 to 30 April 2018

Please answer all questions completely and calculate your payout.

Names: _____ & _____

1. How many Matrix tasks did your team solve correctly?	_____ Tasks
2. How many Matrix tasks do you think other teams have solved on average today?	on average _____ Tasks
3. Please calculate your team payout for the completed tasks (0,50 EUR per correct solution)	_____ EUR
4. Guaranteed participation fee per team (5,00 EUR per person)	10,00 EUR
5. Your entire team payout (3.+4.) Please: a) take your earnings out from the envelope b) leave the remaining money in the envelope c) fold this receipt and put it in the envelope d) throw the envelope (with receipt + remaining money) into the envelope box.	_____ EUR

Faculty of Business, Economics and Social Sciences, University of Kiel

Research study: receipt

Period: 3 February 2018 to 30 April 2018

Please answer all questions completely and calculate your payout.

We, _____ & _____, hereby declare that we have completed this receipt to the best of our knowledge and belief completely and truthfully.

Signature of team member 1 Signature of team member 2

1. How many Matrix tasks did your team solve correctly?	_____ Tasks
2. How many Matrix tasks do you think other teams have solved on average today?	on average _____ Tasks
3. Please calculate your team payout for the completed tasks (0,50 EUR per correct solution)	_____ EUR
4. Guaranteed participation fee per team (5,00 EUR per person)	10,00 EUR
5. Your entire team payout (3.+4.) Please: a) take your earnings out from the envelope b) leave the remaining money in the envelope c) fold this receipt and put it in the envelope d) throw the envelope (with receipt + remaining money) into the envelope box.	_____ EUR

FIGURE 2 | The receipt forms in team treatments. **Appendix 2**, **3** provide the receipt forms in a larger resolution.

multiplying the score with 0.50 EUR and adding a 5-EUR show-up fee per person. It is on this receipt that the individuals or teams could misreport their scores. Receipt forms for the respective treatments were handed to the participants after they had completed the matrix task. All individuals and teams had envelopes at their desk with 15 EUR (individuals) or 30 EUR (teams) in cash, so that any payment dividable by 0.50 EUR was possible. Subsequently, they took their payments out of the envelopes, folded and inserted the receipts into their envelopes, kept their cash payment and dropped the envelopes with the receipts, and unclaimed cash into a return box.

The receipt forms in all treatments included a line (lines) to provide the name of the individual (names of teammates). The difference between the no-signature and signature treatments consisted of the following additional statement at the top of the receipt form that the participants in the signature treatments: "We, [line(s) for name(s)], hereby declare that I (we) have completed this receipt to the best of my (our) knowledge and belief completely and truthfully." Participants in the signature treatments had to sign underneath the statement. Note that there were no instructions or information that suggested any form of detection or punishment related to the statement.

TABLE 1 | Treatment cells.

		Moral commitment	
		No signature	**Signature on top**
Decision maker composition	Individual	Ind_NOsig	Ind_sig
	Team	Team_NOsig	Team_sig

Shu et al. (2012) introduced an honesty nudge which is able to decrease dishonesty and fraud of individuals—signing on the top of a form compared to no signature. They suggested that this nudge helps to turn to an individual's morality and to promote honesty right before the deception may take place—in our experiment before potentially over-reporting the score.

Literature on the dishonesty of teams often points into the direction that teams are more prone to lying than individuals (Danilov et al., 2013; Mühlheußer et al., 2015; Weisel and Shalvi, 2015; Korbel, 2017; Wouda et al., 2017; Kocher et al., 2018; Dannenberg and Khachatryan, 2020). Teams tend to me more strategic about lying and deception (Cohen et al., 2009; Sutter, 2009) and diffusion of responsibility and moral disutility appear to be key drivers (Kocher et al.,

TABLE 2 | Descriptive statistics.

Metric	Treatments			
	Ind_NOsig	*Ind_sig*	*Team_NOsig*	*Team_sig*
Mean solved matrices, as checked by researchers	6.74	6.65	13.81	16.38
Mean solved matrices, as summarized on the matrix sheet	6.91	6.80	14.19	17.24
Mean matrices tried (marked with circles)	7.30	6.80	15.09	17.38
Mean reported matrices in the receipt form	8.74	7.05	17.38	17.09
Guess of mean solved matrices of others	8.35	7.15	14.95	16.71
Share willfully lying	39.1%	10.0%	33.3%	4.7%
Number of participants	23	20	42	42
Number of independent observations	23	20	21	21

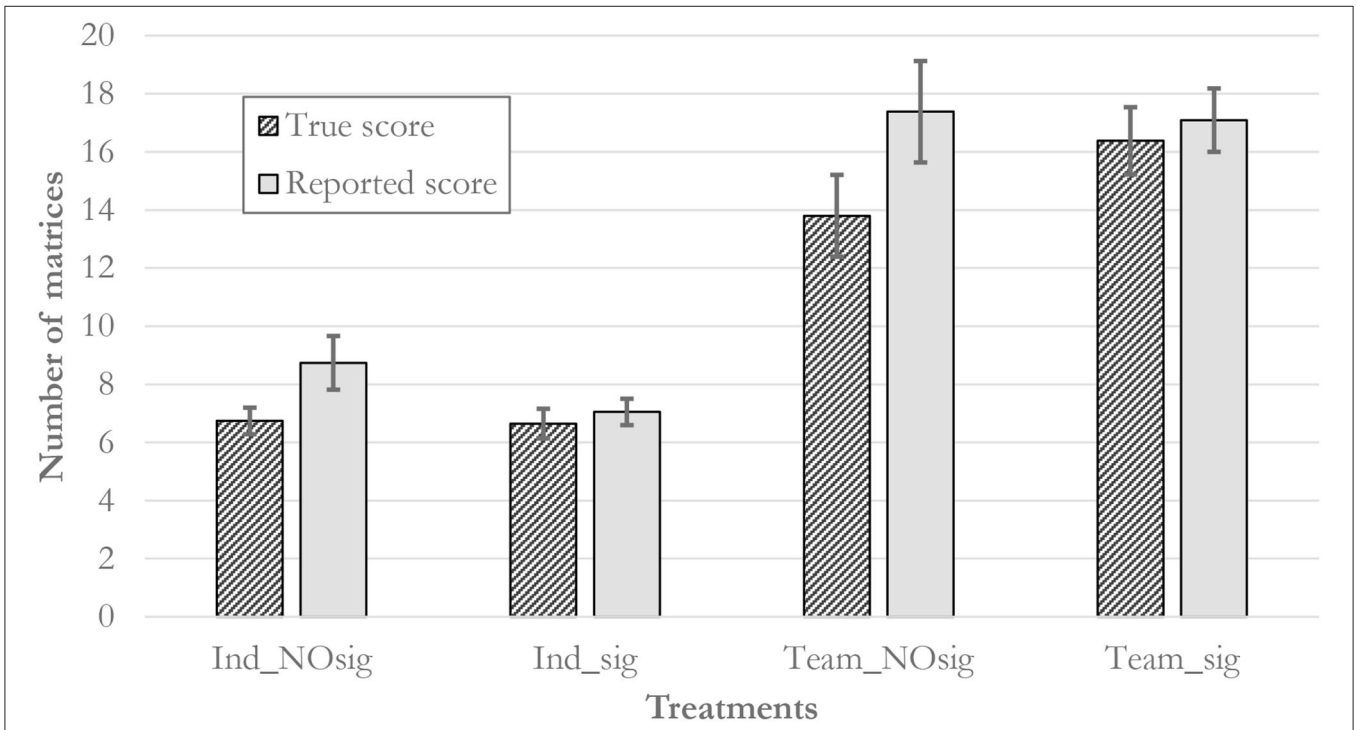


FIGURE 3 | Mean true and reported scores in the four treatments. The bars depict ±1 standard error.

2018). Given that these mechanisms appear to promote dishonesty of teams, it is questionable whether the signature honesty nudge remains effective for teams. If it does, it would be good news for practitioners who employ pledges with signatures to curb dishonesty—yet if the nudge treatment effect is limited to individuals, it would greatly reduce the usefulness of the nudge and potentially other similar nudges, as many fraudulent situations actually involve teams of decision makers. **Table 1** provides an overview of our treatments.

Based on the literature described above, we therefore formulate our key hypothesis that over-reporting of scores is lower in the *_sig* treatments compared to *_NOsig* treatments—both when comparing individuals' reporting decisions and teams' reporting decisions. Hence, we hypothesize that the nudge is effective for teams despite possible counteracting effects from diffusion of responsibility. In order to proceed with a testing our hypothesis, it was essential to replicate finding of Shu et al. (2012) for individual decision makers in our environment and conditions. A total of 127 students of the University of Kiel were recruited through the hroot platform (Bock et al., 2014) and participated in the experiment in the time period February to April 2018. There were 20 and 23 participants in *Ind_NOsig* and *Ind_sig* treatments, respectively. In the *Team_NOsig* and *Team_sig* treatments there were 42 participants per treatment, yielding 21 independent team observations per treatment[5]. The teams were formed randomly by participants of a session drawing numbers on balls from a non-transparent bag.

[5]See **Appendix 1** for instructions.

Following the literature on team dishonesty (e.g. Sutter, 2009), communication between team members may be important to let them get to know each other, develop intra-team trust, exchange thoughts on the task and on motivation to (mis)report the effort. For this reason, we implemented our experiment in a way that team members sat together in a large cubicle. Hence, face-to-face communication of team members was possible throughout the session.

In order to facilitate the team feeling even more, we implemented an additional stage using a creativity task before the actual matrix task and reporting[6]. This task was included in order to help teammates to get to know each other a bit better and "break the ice." Allowing communication when completing tasks together was supposed to mimic situations when teams are working and making decisions together in the real environment. In the creativity task individuals (in the *Ind_* treatments) and teams (in the *Team_* treatments) were given 10 min to create a picture of their choice by using a whiteboard and pins of different colors (see **Appendix 4** for an example). The instructions explicitly informed the participants that there were no incentives related to their creativity or performance and that they were free to do whatever they like. Note that all individuals and teams created a picture, even though an empty whiteboard would have been just as acceptable. In order to be consistent, participants in the *Ind_* treatments also performed this task, but alone. After this creativity task, we ran the matrix task describe above.

RESULTS

Table 2 provides summary statistics of our treatments and **Figure 3** provides an overview of mean reported as well as actually solved matrices. For the following analysis we compare the reported number of solved matrices with the number of solved matrices as noted down on the matrix sheet (see bottom of **Figure 1**) to detect willful dishonesty. We begin this section with an examination of the *Ind_* treatments in order to see whether our results confirm the treatment effect of Shu et al. (2012). In *Ind_sig* fewer individuals over-reported (10%, 2 out of 20) as compared to *Ind_NOsig* (39%, 9 out of 23), which is different based on a (two-sided) Fisher's exact test ($p = 0.039$). Employing Wilcoxon signed-rank tests for differences between score summaries and claimed scores in the receipt for each individual, we find that there is significant over-reporting in *Ind_NOsig* (8.74 reported matrices vs. 6.91 summarized matrices, $p = 0.0039$) and no detectable over-reporting in *Ind_sig* (7.05 vs. 6.80, $p = 0.500$). We therefore find strong support that including the signature nudge at the top of the receipt form reduces dishonesty significantly. Hence, we replicate Shu et al. (2012)'s result (signature on top vs. no signature) for individual decision makers.

We proceed with a similar analysis for the *Team_* treatments to detect whether the signature nudge remains effective in this scenario. Indeed, we find that there are 7 out of 21 teams (33.3%) that over-report their scores on the receipts in *Team_NOsig* compared to only 1 out of 21 (4.7%) in *Team_sig*. These propensities are, again, significantly different from each other (two-sided Fisher's exact test, $p = 0.045$). Wilcoxon signed-rank tests confirm that there is detectable over-reporting in *Team_NOsig* (17.38 matrices claimed vs. 14.19 matrices summarized as solved, $p = 0.0156$), there is no detectable different in *Team_sig* (17.09 vs. 17.24, $p = 0.9725$)[7]. We therefore find clear evidence that the signature nudge curbs dishonesty of teams effectively, alike the scenario for individuals. The result does not support a claim that teams' dishonesty is qualitatively different in a way that makes teams immune to this nudge.

CONCLUSION

This paper asked whether moral nudges that work to curb dishonesty of individuals also remain effective for teams—units that are ubiquitous in companies and have been shown to act more sophisticatedly and feel less responsible for their actions as the outcome of the team's decision rests on the shoulders of several team members (Falk and Szech, 2013; Kocher et al., 2018; Falk et al., 2020). We employ the seminal finding of Shu et al. (2012) who showed that asking for a signature to confirm honesty at the top of a form fosters honesty compared to no signature. The main argument is that this can help to turn to an individual's morality and promote honesty exactly before misreporting may take place.

After the successful replication of Shu et al. (2012)'s effect for individuals, we extended the finding by confirming that this nudge is equally effective for a team setting, resulting in an 86% decrease in the amount of cheating teams. In our eyes, the presented research makes an important contribution to a better understanding of team behavior and in developing instruments for preventing teams and individuals from deception and cheating.

To the best of our knowledge, this is the first study to investigate the effectiveness of moral nudges for teams and it should be considered as a starting point for avenue of future research. Future research may investigate the dimensions of familiarity of team members, which our creativity task aimed for, further. Likewise, our teams consisted of two members and future research could vary this dimension by examining behavior of larger teams. Field experimental methods could be used decrease scrutiny of laboratory experiments and similar studies with higher stakes could check for the robustness of our and Shu et al.'s findings. Such investigations seem promising to test the ecological validity of our results. We regard as highly policy-relevant to investigate team decision-making and develop

[6]See Kachelmeier et al. (2008), Erat and Gneezy (2016, 2017), Charness and Grieco (2019), Grözinger et al. (2020), and Kachelmeier and Williamson (2010) for economic experiments on creativity.

[7]In *Team_sig* there was even one team that reported a lower number than summarized on the matrix sheet, yet indeed the correct number when comparing the reported number of matrices with the correctly solved number checked by the research team.

cost-effective instruments like nudges that can be implemented in practice by organizations and policymakers to curb fraud and dishonesty of teams.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YD wrote his master thesis under the supervision of MK and this work is a concise product coming out of this collaboration. YD and MK developed the research question and the experiment material together. YD and MK ran the experiment together and YD analyzed the data. MK contributed the creativity task as a team-building exercise, wrote the paper, and financed the experiment. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.684755/full#supplementary-material

REFERENCES

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica* 87, 1115–1153. doi: 10.3982/ECTA14673

ACFE (2020). *Report to the Nations – 2020 Global Study on Occupational Fraud and Abuse*. Association of Certified Fraud Examiners. Available online at: https://acfepublic.s3-us-west-2.amazonaws.com/2020-Report-to-the-Nations.pdf (accessed March 20, 2021).

Amir, O., Mazar, N., and Ariely, D. (2018). Replicating the effect of the accessibility of moral standards on dishonesty: authors' response to the replication attempt. *Adv. Methods Pract. Psychol. Sci.* 1, 318–320. doi: 10.1177/2515245918769062

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *Eur. Econ. Rev.* 71, 117–120. doi: 10.1016/j.euroecorev.2014.07.003

Bornstein, G., Kugler, T., and Ziegelmeyer, A. (2004). Individual and team decisions in the centipede game: are teams more "rational" players?. *J. Exp. Soc. Psychol.* 40, 599–605. doi: 10.1016/j.jesp.2003.11.003

Charness, G., and Grieco, D. (2019). Creativity and incentives. *J. Eur. Econ. Assoc.* 17, 454–496. doi: 10.1093/jeea/jvx055

Charness, G., Samek, A., and van de Ven, J. (2021). *What is Considered Deception in Experimental Economics?* Working paper.

Charness, G., and Sutter, M. (2012). Teams make better self-interested decisions. *J. Econ. Perspect.* 26, 157–176. doi: 10.1257/jep.26.3.157

Cohen, T. R., Gunia, B. C., Kim-Jun, S. Y., and Murnighan, J. K. (2009). Do teams lie more than individuals? Honesty and deception as a function of strategic self-interest. *J. Exp. Soc. Psychol.* 45, 1321–1324. doi: 10.1016/j.jesp.2009.08.007

Danilov, A., Biemann, T., Kring, T., and Sliwka, D. (2013). The dark side of team incentives: experimental evidence on advice quality from financial service professionals. *J. Econ. Behav. Organ.* 93, 266–272. doi: 10.1016/j.jebo.2013.03.012

Dannenberg, A., and Khachatryan, E. (2020). A comparison of individual and team behavior in a competition with cheating opportunities. *J. Econ. Behav. Organ.* 177, 533–547. doi: 10.1016/j.jebo.2020.06.028

Erat, S., and Gneezy, U. (2016). Incentives for creativity. *Exp. Econ.* 19, 269–280. doi: 10.1007/s10683-015-9440-5

Erat, S., and Gneezy, U. (2017). Erratum to: Incentives for creativity. *Exp. Econ.* 20, 274–275. doi: 10.1007/s10683-016-9495-y

Falk, A., Neuber, T., and Szech, N. (2020). Diffusion of being pivotal and immoral outcomes. *Rev. Econ. Stud.* 87, 2205–2229. doi: 10.1093/restud/rdz064

Falk, A., and Szech, N. (2013). Morals and markets. *Science* 340, 707–711. doi: 10.1126/science.1231566

Fellner, G., Sausgruber, R., and Traxler, C. (2013). Testing enforcement strategies in the field: threat, moral appeal and social information. *J. Eur. Econ. Assoc.* 11, 634–660. doi: 10.1111/jeea.12013

Gerlach, P., Teodorescu, K., and Hertwig, R. (2019). The truth about lies: a meta-analysis on dishonest behavior. *Psychol. Bull.* 145, 1. doi: 10.1037/bul0000174

Grözinger, N., Irlenbusch, B., Laske, K., and Schröder, M. (2020). Innovation and communication media in virtual teams-an experimental study. *J. Econ. Behav. Organ.* 180, 201–218. doi: 10.1016/j.jebo.2020.09.009

Kachelmeier, S. J., Reichert, B. E., and Williamson, M. G. (2008). Measuring and motivating quantity, creativity, or both. *J. Account. Res.* 46, 341–373. doi: 10.1111/j.1475-679X.2008.00277.x

Kachelmeier, S. J., and Williamson, M. G. (2010). Attracting creativity: the initial and aggregate effects of contract section on creativity-weighted productivity. *Account. Rev.* 85, 1669–1691. doi: 10.2308/accr.2010.85.5.1669

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2019). Intuitive honesty versus dishonesty: meta-analytic evidence. *Perspect. Psychol. Sci.* 14, 778–796. doi: 10.1177/1745691619851778

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? We lie! why? Experimental evidence on a dishonesty shift in teams. *Manage. Sci.* 64, 3971–4470. doi: 10.1287/mnsc.2017.2800

Korbel, V. (2017). Do we lie in teams? An experimental evidence. *Appl. Econ. Lett.* 24, 1107–1111. doi: 10.1080/13504851.2016.1259734

Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., et al. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *Proc. Natl. Acad. Sci. U.S.A.* 117, 7103–7107. doi: 10.1073/pnas.1911695117

Kugler, T., Kausel, E. E., and Kocher, M. G. (2012). Are teams more rational than individuals? A review of interactive decision making in teams. Wiley Interdisciplinary Reviews: *Cogn. Sci.* 3, 471–482. doi: 10.1002/wcs.1184

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *J. Market. Res.* 45, 633–644. doi: 10.1509/jmkr.45.6.633

Mühlheußer, G., Roider, A., and Wallmeier, N. (2015). Gender differences in honesty: teams versus individuals. *Econ. Lett.* 128, 25–29. doi: 10.1016/j.econlet.2014.12.019

Rosenbaum, S. M., Billinger, S., and Stieglitz, N. (2014). Let's be honest: a review of experimental evidence of honesty and truth-telling. *J. Econ. Psychol.* 45, 181–196. doi: 10.1016/j.joep.2014.10.002

Shu, L. L., Gino, F., and Bazerman, M. H. (2011). Dishonest deed, clear conscience: when cheating leads to moral disengagement and motivated forgetting. *Pers. Soc. Psychol. Bulletin* 37, 330–349. doi: 10.1177/0146167211398138

Shu, L. L., Mazar, N., Gino, F., Ariely, D., and Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proc. Natl. Acad. Sci. U.S.A.* 109, 15197–15200. doi: 10.1073/pnas.1209746109

Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *Econ. J.* 119, 47–60. doi: 10.1111/j.1468-0297.2008.02205.x

Thaler, R. H., and Sunstein, C. R. (2009). *Nudge*. London: Penguin Books.

Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., et al. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Adv. Methods Pract. Psychol. Sci.* 1, 299–317. doi: 10.1177/2515245918781032

Weisel, O., and Shalvi, S. (2015). The collaborative roots of corruption. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10651–10656. doi: 10.1073/pnas.1423035112

Wouda, J., Bijlstra, G., Frankenhuis, W. E., Wigboldus, D. H., and Moore, D. (2017). The collaborative roots of corruption? A replication of Weisel and Shalvi (2015). *Collab. Psychol.* 3, 1–3. doi: 10.1525/collabra.97

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 July 2021
doi: 10.3389/fpsyg.2021.675319

When and Why Contexts Predict Unethical Behavior: Evidence From a Laboratory Bribery Game

Sining Wang[1] *and Tao Chen*[2]*

[1] Department of Economics, Case Western Reserve University, Cleveland, OH, United States, [2] Big Data Research Lab, Department of Economics, University of Waterloo, Waterloo, ON, Canada

In economic unethical decision-making experiments, one important methodological investigation is what types of contexts should be used to frame the instructions. Within the experimental economics community, using neutral-context instructions instead of loaded-context instructions is the mainstream practice. Because the loaded contexts may impact behavior in an unpredictable manner and therefore, put experimental control at risk. Nevertheless, using the loaded-context instructions could be advantageous in several ways. A properly framed context can help to facilitate learning and gain ecological validity. The challenge is whether we can identify when and why the loaded context may alter behavior. In this paper, we aim to test if being familiar with a loaded context can systematically influence unethical decisions in a bribery game. We conduct a laboratory bribery game experiment with three different treatments: the neutral-context treatment, the familiar-context treatment, and the unfamiliar-context treatment. Using the neutral-context treatment as a benchmark, we find that participants in the familiar-context treatment express stronger negative attitudes toward corruption. Attitudes toward unethical behavior are the same in the neutral-context treatment and the unfamiliar-context treatment. Behaviorally, the participants in the familiar-context treatment are much less likely to engage in corrupt activities. The neutral-context treatment and the unfamiliar-context treatment produce the same behavioral outcome.

Keywords: unethical decision, context effect, bribery game, corruption, experimental design

OPEN ACCESS

Edited by:
Nora Szech,
Karlsruhe Institute of Technology (KIT), Germany

Reviewed by:
Iveta Eimontaite,
Cranfield University, United Kingdom
Ion Juvina,
Wright State University, United States

***Correspondence:**
Tao Chen
t66chen@uwaterloo.ca

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *02 March 2021*
Accepted: *15 June 2021*
Published: *09 July 2021*

Citation:
Wang S and Chen T (2021) When and Why Contexts Predict Unethical Behavior: Evidence From a Laboratory Bribery Game. Front. Psychol. 12:675319. doi: 10.3389/fpsyg.2021.675319

INTRODUCTION

Over the past three decades, the study of unethical decision making has received increasing attention. In laboratory economic experiments, one commonly used technique to investigate the underlying motivation of unethical behavior is to put a decision maker in a position where he or she must decide whether to engage in economically rational but dishonest practices. In such experiments, an important methodological debate is whether one should frame the experimental instruction with neutral context or loaded context (Alekseev et al., 2017).

Within the experimental economics community, framing the instruction with neutral context is the mainstream practice. Smith (1976) proposed that people with varied backgrounds and preferences may interpret the value of ethics embedded in the context differently. The different interpretations are often unobservable, and therefore, will affect behavior in an unpredictable manner. To avoid uncontrollable data distortion, experimenters should use "neutralized"

instruction, and then induce the subjects' preferences with only monetary reward. However, this approach has been criticized because it focuses solely on the external incentives thus ignoring the importance of ethics and psychological costs (Bardhan, 2006). A large literature also suggests that using loaded context instruction could be advantageous—a meaningful context that is related with the research question can help the researcher better understand the participants' motives (Alm et al., 1992; Aronson, 1992, 1999; Andreoni, 1995; Andreoni and Miller, 2002; Abbink and Hennig-schmidt, 2006; Bardhan, 2006; Alatas et al., 2009; Barr and Serra, 2009; Armantier and Boly, 2014; Banerjee, 2016; Alekseev et al., 2017). Moreover, the loaded context can facilitate learning, making the experimental tasks more understandable to the participants (Wason and Shapiro, 1971; Griggs and Cox, 1982; Chou et al., 2009).

It is generally agreed that altering the experimental context could have profound effects on unethical decisions. The bone of contention is whether such effects are predictable. Many past studies have contributed to this heated and ongoing debate, yet little consensus has been reached. For instance, it is presumably that context plays a major role in determining people's decisions in bribery games—calling participants "Public officers" and "Firm owners" instead of "Player 1" and "Player 2" may lead to divergent behavioral outcomes. As a matter of fact, a considerable amount of evidence has been found to support this conjecture (Eckel and Grossman, 1996; Cooper et al., 1999; Carpenter et al., 2008; Laury and Taylor, 2008; Alatas et al., 2009; Barr and Serra, 2009). However, multiple studies show that the neutral context and the loaded context produce the same behaviors in bribery games (Cooper et al., 1999; Barr and Serneels, 2004; Abbink and Hennig-schmidt, 2006; Armantier and Boly, 2014).

The question we address in this paper is whether the effect of context is always unpredictable. In particular, we use a laboratory bribery game as an example to examine what kind of experimental context may influence unethical decisions in a systematic, predictable way.

NOT ALL CONTEXTS ARE CREATED EQUAL

Past studies in bribery games examine the distinctions between two types of contexts: either neutral context (framed with abstract language, no specific background story) or loaded context (framed with a specific background story). However, we consider such a dichotomous view insufficient: Not all the loaded contexts have the same impact on decisions. Extensive evidence suggests that emotional responses triggered by the context alter people's behavior. It is worthwhile to take account of how people's real-life experiences may influence their perceptions of the loaded contexts, which in turn, affect decision.

Alekseev et al. (2017) proposed to distinguish between three types of contexts. The first type, which is called the "abstract context" or "neutral context," uses neutral language such as "player A," "option B" and so on: The neutral context is not related to any specific background story. The second type, which is called the "meaningful context," presents the experimental tasks in specific scenarios. However, the artificial scenarios do not evoke emotions or connotations. The third type, which is called the "evocative context," presents the tasks in scenarios that are not only related to a real-life situation, but also evoke strong emotional responses. Inspired by this insight, we consider people's emotional responses might be the key to understand the mechanism through which contexts affect people's decision in unethical decision-making experiments.

From the psychology literature, Blanchette and Caparos (2013) showed that emotion plays a significant role in logical reasoning and decision making. In particular, they suggested that contexts that is relevant to individual's past experiences are more likely to evoke emotional responses. Consequently, people tend to devote more cognitive resources to such decision-making situations. To put it in another way, a decision-maker would be more "emotional" in contexts that is relevant to themselves. Goel and Vartanian (2011) compared people's reasoning process in neutral contexts and emotionally charge contexts. They found that under certain conditions, the emotional factors in the context can foster a more vigilant, systematic information-processing style. Greene et al. (2001) investigated the changes in brain activities when people respond to ethical dilemmas. The same ethical dilemma was presented in two contexts: personal context (where the participants are more engaged emotionally) and impersonal context (where the participants are less engaged emotionally). They found that responding to personal ethical dilemmas produces increased brain activity in areas associated with emotional processing. Besides, they also found people have to spend more cognitive resources to overcome their emotional responses in the personal context.

All the above studies lead to the point that emotion and context jointly determine behavior. In the realm of unethical decision-making, we argue that an evocative context may alter people's reasoning and behavior by increasing the emotional charge. For instance, in bribery games, unethical behaviors typically impose negative externalities to the society, which might bring the individual with considerable psychological costs. Adopting the evocative context may make the psychological costs more salient. When people are facing scenarios that evoke strong (negative) emotional responses, they are more likely to think about the negative consequences of their decision. Accordingly, their behaviors in the lab can better reflect what they may do in naturally occurring environments in their everyday life.

In the current study, we aim to test if being familiar with a loaded context can systematically influence unethical decisions in a bribery game. In particular, we put forward that a loaded context that is closely related to the decision maker's real-life experience is more likely to orient her to associate the hypothetical scenario with her self-concept, and therefore, evoke strong emotional responses. Consequently, the decision maker is more likely to perceive it as the "evocative context" (and putatively more emotional). The decision maker is more engaged with the task and is likely to devote more attention to her decisions. Moreover, the moral standard and social norms embedded in such a context are more salient to the individual. Actions that violate certain moral obligations or injunctive norms would bring the decision

maker with considerable psychological costs. Behaviorally, the decision maker is less likely to engage in dishonest practices.

Hypothesis 1. Unethical behaviors should be less likely to happen in the evocative context, as compared with the meaningful context and the neutral context.

On the other hand, a loaded context that is distant from the individual's real-life experience is more likely to be perceived as the "meaningful context" (and putatively less emotional). Although the meaningful context is constructed with a specific scenario, it doesn't evoke strong emotions or significant psychological responses—It leads the decision maker to be unattached to the task. The moral standards and social norms in a meaningful (yet remote) context are ambiguous, or vague to the individual. The ambiguity in the moral standard plus the lack of personal involvement make it easier to find external justifications for a dishonest practice. To escape from the aversive state and strive for self-consistency, people would rationalize their unethical decisions. The reasoning can possibly be: "This is just a game; I would not do that in real life" (although the participant had no similar experience in real life), or "I'm curious about what the consequences are for choosing this; let me try it out." Because neither the meaningful context nor the neutral context evokes strong emotions, the psychological cost of engaging in dishonest behavior should be similar in these two conditions. Thus, we would expect the meaningful context and the neural context drives similar behavioral outcome.

Hypothesis 2. The meaningful context and the neutral context will lead to similar behavioral outcomes.

While people's behaviors are observable, the motives of the behaviors are not. In the field of social psychology, it has been widely accepted that behavior is guided by attitudes (e.g., Ajzen et al., 2018). In the current study, we are curious about if attitudes toward bribery can help explain unethical decision making. To complement the laboratory experiment, we conduct an independent survey to measure participants' attitude toward bribery. We want to test if people's attitude toward corruption can predict behavior in the bribery game.

EXPERIMENTAL DESIGN

Administration

The experiment is conducted at the school of business in Jianghan University (Wuhan, China). The experimental procedure is reviewed and approved by the research ethics committee at Jianghan University. To recruit participants, we distribute recruitment flyer to students during their self-study sessions. Students who are interested in participating will response to our recruit email on the flyer. The experimenter then sends them the electronic copy of the information sheet. Potential participants can take as much time as they need to make the decision. For students who decided to participate, we send them the invitation with detailed time and location.

Upon arrival, the students will first receive the consent form, and then orally indicate whether they agree to participant. Informed consent is obtained from all participants. Next, the participant will be randomly assigned with a unique experimental identification number. This number will be used to track their decisions and responses during the experiment. Since all data are collected anonymously, we do not ask the participants to provide signed consent.

In total, 340 students (92 male, 248 female), which consisted of freshmen or sophomores, participated in the experiment. Among the 340 participants, 250 of them (56 male, 194 female) are randomly invited to our lab to play a bribery game and followed by a short questionnaire asking about their decisions and reasoning in the game. For the rest of the 90 students (36 male, 54 female), we conduct an independent attitude survey to obtain the perceived attitudes toward unethical behaviors in each game context. All data is collected anonymously. It is very important to note that each participant only participates in either the bribery game plus the corresponding questionnaire, or the attitude survey.

To run the bribery game, we conduct 13 sessions with either 10 or 20 participants in each. It takes approximately 60 min (including check-in and payment processing) to run one session. All the sessions are conducted with computer-based materials, which are developed using z-tree (Fischbacher, 2007). During the experiment, all participants make decisions anonymously, and earn "points" (the fictitious experimental currency). At the conclusion, participants are paid in cash privately at the rate: 1 RMB (=0.16 US dollar) for every 100 points they earn. The average earnings are 30 RMB (including 5 RMB show-up fee)[1].

The Laboratory Bribery Game

We use a laboratory bribery game to simulate a decision-making scenario in which unethical behavior may occur. All the 250 participants who participate in the bribery game are randomized into 25 groups with ten participants in each.

In the beginning of the game, each participant is randomly assigned with a role. Within a group, five participants play as applicants (potential bribers, **player 1** below), the other five participants play as granter (potential bribee, **player 2** below). Each player 1 applies for five different grants (each grant values 1,000 game points); each player 2 is in charge of allocating the 1,000 game points among the five player 1s. In addition, each player 1 is randomly paired with a player 2. Prior to the player 2's point allocation decision, the two participants in a pair can interact with each other. We adopt a fixed-partner design to allow repeated interactions between the paired players. All the interactions are anonymous. After the role assignment, the participants start to make decisions. The process is as follow:

- At the beginning of each period, each player 1 receives 200 points as an initial endowment. Player 2 has no endowment.
- Player 1 first decides whether to make a private ***transfer*** to the player 2 in his/her pair. If the decision is to transfer, the participant must specify a whole integer in the range from 1 to 200 points.
- Following that decision, the player 2 may face one of the two cases:

[1]At Jianghan University, 30 RMB is approximately the cost of a one-person daily meal in the student dining hall.

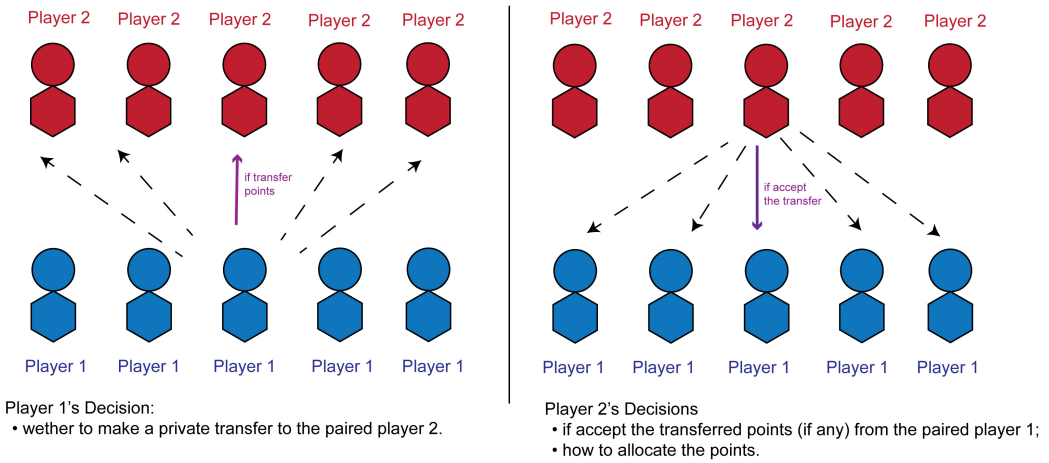


FIGURE 1 | The basic setup of the players' decisions.

- Case 1: the paired player 1 decided NOT to transfer point. In this case, the player 2 sees a feedback "*no point being transferred*" and has no decision to make at this step.
- Case 2: the paired player 1 decided to make a transfer. Then the player 2 sees the total points being transferred by the player 1, and then decides whether to ***accept*** or ***reject*** the bribe. If accepting it, then the amount offered is deducted from the player 1's account and added to the player 2's account. If the player 2 rejects the bribe, then both players' accounts remain unchanged.

- Last, the player 2 decides how to allocate the 1,000 points among the five player 1s. If abiding by the game rules, then each player 1 earns 200 points (equal split). If violating the game rules, then the player 1 in the pair earns 1,000 points, and the other player 1s earn nothing.
- After all the allocation decisions have been made, the player 1 sees feedback on how the points are allocated.
- **Figure 1** illustrates the players' decisions.

The game repeats for 15 periods with fixed partners. At the end of period 15, all participants will be reassigned with a different role, and then paired with a strange partner. The new pairs will then play the same game for another 15 periods. That is to say, if a player was the briber in the first half (period 1–15), she/he will be playing the bribee in the second half of the game (period 16–30). At the conclusion of the experiment, four periods (two from period 1–15, two from period 16–30) are randomly selected to determine the players' payment.

During the iterations, a pair of participants is identified as a "rule-breaking pair" if any offer from the player 1 is accepted by the player 2. If a pair has been identified as the "rule-breaking pair" at least once, then there is a 1% chance the punishment occurs: both players' earnings are cleared from their accounts. By the end of all the 30 periods, a lottery is played out to decide whether to punish the rule-breaking pairs. The extremely low probability reflects that most corrupt activities in reality are difficult to discover. As a matter of fact, many corrupt activities are even unobservable, and the severe penalty we impose represents the consequences arising from discovery of corrupt activities. **Figure 2** depicts the extensive form of the game in each period within each pair. Use T denotes the number of points offered by player 1. X and Y denote the possible penalty for the player 1 and the player 2, respectively.

Under the homo-economicus assumption, a rational decision maker is motivated by pure self-interest. The rational decision maker does not have to overcome moral qualms about unethical behavior. The theoretical equilibria of the game are not hard to obtain. Since this is a finite-repeated game, rational players will apply backward induction to solve for a unique subgame perfect Nash equilibrium. On an equilibrium path, a player 2 is indifferent between "abide by the rules" and "violate the rules" [2]. Accordingly, player 2 will play the two alternatives with the same probability (50%). Furthermore, player 2's expected payoffs for accepting the bribe is T-Y, which is greater than 0. Therefore, player 2 will accept any bribe being offered. Given that, the expected payoffs of a player 1 who offers T points to his or her partner is $(1200\text{-}2T\text{-}2X)/2$, which is lower than the expected payoffs of offering nothing $(1200/2 = 600)$. That is, not bribing is the dominant strategy for player 1. In equilibrium, player 1 does not offer a bribe to player 2, and player 2 violates the allocation rule with a probability of 50%. However, a growing literature has shown that actions that violate social norms can bring the decision maker with considerable psychological costs. We anticipate that participants' behavior will deviate from the theoretical equilibria.

Treatments

Three treatments are conducted with the same bribery game framework. The treatments only vary in the experimental instructions. In the first treatment, the game is presented as a

[2]Note that the determination of "rule-breaking" is based on the decision regarding whether or not to accept the bribe, rather than the decision regarding point allocation.

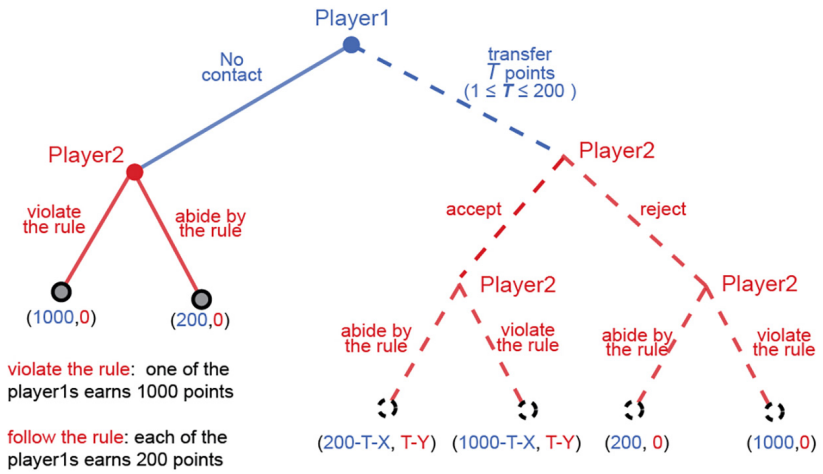


FIGURE 2 | The extensive form of the bribery game in each period.

Table 1 | The contexts and vocabulary used in the three treatments.

Treatments		Familiar context	Unfamiliar context	Neutral context
Earnings		Scholarship	Profits	Points
Player 1's role		Student	Bidder	Applicant
Player 1's alternatives	Alternative 1	Make a transfer	Make a transfer	Make a transfer
	Alternative 2	No contact	No contact	No contact
Player 2's role		Advisor	Bid-inviter	Granter
Player 2's alternatives	Alternative 1	Abide by the rule	Abide by the rule	Abide by the rule
	Alternative 2	Violate the rule	Violate the rule	Violate the rule

scholarship allocation scenario in a college[3] (the familiar-context treatment below). In the second treatment, the game is presented as a competitive bidding scenario among firms (the unfamiliar-context treatment below). In the third treatment, the game is presented in an abstract form without any specific scenario or role (the neutral-context treatment below). **Table 1** summarizes the roles and terminology for the alternatives in each of the treatments. All participants are randomly assigned to one of the three treatments. In total, 100 students participate in the familiar-context treatment, 110 students in the neutral-context treatment, and 40 students in the unfamiliar-context treatment.

Since all participants are college students, we conjecture that the college scenario is more likely to be perceived as an evocative context. Unethical decisions in this context will trigger strong emotional responses, bringing the decision maker considerable psychological costs. Consequently, corrupt conduct (offer bribe, accept bribe, or violate the rule) should be less likely to happen in the familiar-context treatment.

Another question we are curious about is whether the unfamiliar-context treatment and the neutral-context treatment may lead to different behavioral results. As discussed earlier, a meaningful but not evocative context will not trigger emotional responses. The participants in the meaningful (yet unfamiliar) context should bear the same psychological costs as in the neutral context. As a result, we conjecture that the unfamiliar-context treatment and the neutral-context treatment will produce the same behavior.

The Attitude Surveys

In addition to the laboratory bribery game, we also conduct an independent attitude survey to measure students' attitudes toward unethical behaviors. In the survey, we present the bribery relationship to the respondents, and then ask them to indicate their attitude on a 7-point Likert scale. Similar to the laboratory bribery game, the same interaction structure is framed with three different contexts (i.e., familiar context, unfamiliar context, neutral context). Please see the survey with familiar context below as an example[4].

> Imagine a scholarship allocation scenario in a college. In total five students applied to the same scholarship. There are 1,000 dollars available in the award pool. All the student applicants are equally qualified. **According to the college policy**, the academic advisor shall split the $1,000 dollars among the five

[3]At the Jianghan University (and many other colleges in China), the academic advisor is in charge of scholarship allocation.

[4]All the surveys are attached in the complementary materials. The original version of the surveys is in Chinese language (available upon request).

applicants. That is to say, each of the applicants shall receive an award of $200.

However, prior to the scholarship allocation decision, one of the five students talked to the academic advisor, sent him a gift that worth $200 (secretly and privately). As return, the academic advisor announced that student as the only person who won the scholarship, distributed all $1,000 to her. All other applicants earned nothing. The interaction between the student and the academic advisor will not be discovered by others.

Please select the response that indicates the degree to which you agree or disagree with the STUDENT and the ACADEMIC ADVISOR'S activities. There is no right or wrong answer, so try hard to be completely honest in your responses. You can state your opinions accurately as the information you submit will be completely confidential.

For the STUDENT:

1	2	3	4	5	6	7
Extremely Disagree	Moderately Disagree	Somewhat Disagree	Not Sure	Somewhat Agree	Moderately Agree	Extremely Agree

For the ACADEMIC ADVISOR:

1	2	3	4	5	6	7
Extremely Disagree	Moderately Disagree	Somewhat Disagree	Not Sure	Somewhat Agree	Moderately Agree	Extremely Agree

Besides, we also ask the respondents to indicate their sex. To the best of our knowledge, this is the first study that use independent attitude survey to complement laboratory bribery experiment.

In total 90 participants are invited to our lab to complete the survey. The 90 students are randomized into the three different contexts (with 30 respondents in each context). We adopt a between-subjects design, each respondent only participant in one of the three contexts. Given that the survey respondents and the laboratory game participants are randomly chosen from the same population, we assume that they should have similar attitudes toward unethical behaviors in the given contexts. Our design allows us to obtain measures for attitudes that are not influenced by decisions in the laboratory bribery game. Results from the attitude survey can inform us what people perceive as the "right thing to do" in each context. Ideally, the attitudes should be able to help predict people's behavior in the bribery game experiment.

ANALYSIS AND RESULTS

Attitudes Toward Corrupt Activities in Each of the Contexts

To analyze the survey data, we take people's attitude toward unethical conduct as the dependent variable. The first independent variable is role, which has two levels: player 1 or player 2; The second independent variable is context, which has three levels: familiar context, unfamiliar context, and neutral context. We first perform a two-way 2 (role: player 1 or player 2) × 3 (context: familiar, unfamiliar, neutral) mixed measures ANOVA with repeated measures on the "role" variable (because each respondent needs to indicate their attitude toward both players). The result is presented in **Table 2**. From this table, we learn that both the context variable and the role variable have significant main effects on attitude. However, there is no two-way interactions between the context variable and the role variable on attitude [$F_{(2,87)} = 0.735$, $p = 0.482$].

Table 2 | Attitudes toward unethical behavior: two-way Mixed ANOVA.

Effect	DFn	DFd	F-statistics	GES
Context	2	87	16.483**	0.191
Role	1	87	4.561*	0.019
Context: role	2	87	0.735	0.006

*Indicates the result is statistically significant at the p = 0.05 level.
**Indicates the result is statistically significant at the p = 0.01 level.

We then compare the attitudes to the two roles: The mean score toward corrupt conduct is 2.793 and 2.344 for player 1 and player 2. This difference is statistically significant ($p < 0.001$). That is, people rate player 2's unethical behavior more negatively. We also conduct a pairwise comparison between group levels to see how context impact attitudes on each role. The result (**Table 3**) indicates that the mean attitude score is significantly lower in the familiar context, as compare with the other contexts. The attitude score is not significantly different in the unfamiliar-context vs. neutral context comparison. The distribution of people's attitudes toward unethical behavior in each of the three contexts are presented in the boxplot in **Figure 3**. Keep in mind that because the respondents of the attitude survey did not participate in the laboratory bribery game, their responses are not influenced by the game.

Corrupt Activities in the Laboratory Bribery Game

To analyze the data from the bribery game, we first pool all the participants' data together, use exploratory analysis to examine how the contexts may change behavior; Next, we look at if the interaction patterns between the paired players are different across the contexts; Finally, we apply a random effect model to investigate how the contexts may impact the dynamic of decision making over time.

Exploratory Analysis

In general, the frequency of a player 1's bribery attempt is 37% across all treatment, and the frequency of a player 2 violating the rule when allocating resources is 14.33%. The difference is statistically significant (Fisher exact test $p < 0.001$). That is, player 2 is less likely to engage in corrupt activities.

The frequency of a player 1's bribery attempt is 31.13% in the familiar-context treatment, 39.81% in the unfamiliar-context treatment, and 41.83% in the neutral-context treatment. Fisher exact test results indicate that the familiar-context treatment has the lowest bribery rate (Fisher exact test $p < 0.0001$ in comparison to the unfamiliar-context and

Table 3 | Pairwise comparison of the mean attitude toward bribing behavior.

	Group 1		Group 2				
	Context	**Mean (std)**	**Context**	**Mean (std)**	**Mean-difference**	***p*-value**	**Adjusted *p*-value[a]**
Player 1	Familiar	1.57 (1.30)	Unfamiliar	3.77 (1.83)	−2.20	0.000	0.000
	Familiar	1.57 (1.30)	Neutral	3.07 (1.78)	−1.50	0.001	0.002
	Unfamiliar	3.77 (1.83)	Neutral	3.07 (1.78)	0.70	0.105	0.316
Player 2	Familiar	1.43 (1.01)	Unfamiliar	3.00 (2.03)	−1.57	0.000	0.001
	Familiar	1.43 (1.01)	Neutral	2.60 (1.67)	−1.17	0.001	0.022
	Unfamiliar	3.00 (2.03)	Neutral	2.60 (1.67)	0.40	0.35	1

[a]*Bonferroni corrected p-value.*

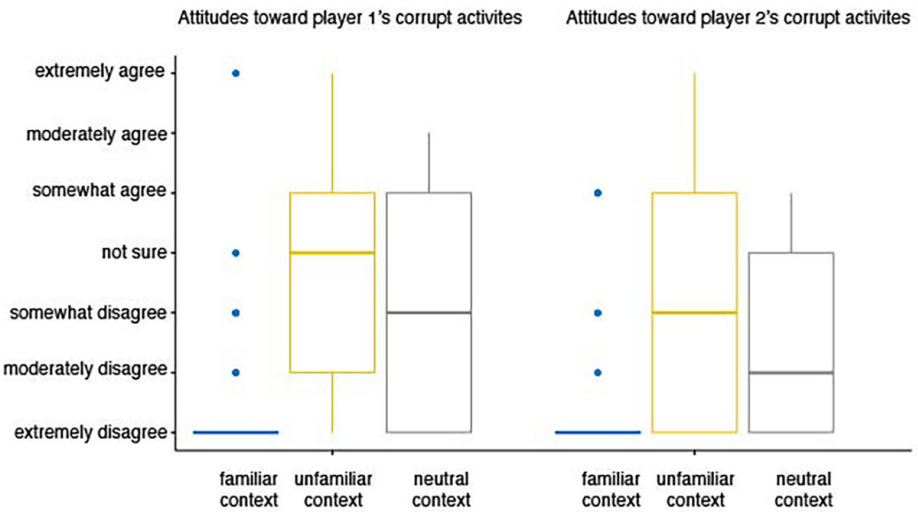


FIGURE 3 | Results of the attitude surveys.

$p < 0.0001$ in comparison to the neutral-context treatment). No evidence suggests that the player 1's bribery rate in the unfamiliar-context treatment is significantly different than in the neutral-context treatment (Fisher exact test $p = 0.4089$). **Table 4** summarizes the player 1 behavior. In the familiar-context treatment, 36% of individual player 1s never tried to bribe their partners; this proportion is 20% in the unfamiliar-context treatment (significantly lower than in the familiar-context treatment; Fisher exact test $p = 0.0480$), and 17.27% in the neutral-context treatment (significantly lower than in the familiar-context treatment; Fisher exact test $p = 0.0017$). In the experiment, some of the player 1s may have selected the bribery option by mistake (or perhaps to become familiar with the game). Among all player 1s in the familiar-context treatment, 43% made bribery attempts no more than 1 time (out of 15 periods); this number is 20% in the unfamiliar-context treatment (significantly lower than in the familiar-context treatment; Fisher exact test $p = 0.0079$), and 25.45% in the neutral-context treatment (significantly lower than in the familiar-context treatment; Fisher exact test $p = 0.0055$). Moreover, the proportion of participants who constantly bribe the partners is the lowest in the familiar-context treatment (Fisher exact test $p < 0.001$).

We then compare the outcomes in the unfamiliar-context treatment and the neutral-context treatment. We do not find any significant differences (proportion of participants who never offer bribe: $p = 0.8922$; proportion of participants who offer a bribe no more than 1 time: $p = 0.478$; proportion of participants who constantly offer a bribe: $p = 1.0$).

Table 5 summarizes the frequencies of the player 2s' corrupt activities The proportion of participants who never violate the rule is 64% in the familiar-context treatment and 17.50% in the unfamiliar-context treatment. These two proportions are significantly different (Fisher exact test $p < 0.0001$). The player 2s in the familiar-context treatment are also much more likely to abide by the rules than those in the neutral-context treatment (Fisher exact test $p < 0.0001$). The proportion of participants who never violate the rule is 17.50% in the unfamiliar-context treatment and 30% in the neutral-context treatment. Again, the difference is not statistically significant (Fisher exact test $p = 0.147$).

Table 4 | The frequency of the player 1s' (potential bribers) bribing attempts.

Never offer bribe (attempt=0/15)			No more than one time (attempts <= 1/15)			Constantly offer bribe (attempts >= 8/15)		
Familiar context	**Unfamiliar context**	**Neutral context**	**Familiar context**	**Unfamiliar context**	**Neutral context**	**Familiar context**	**Unfamiliar context**	**Neutral context**
36/100	8/40	19/110	43/100	8/40	28/110	29/100	15/40	41/110
36.00%	20.00%	17.27%	43.00%	20.00%	25.45%	29.00%	37.50%	37.27%

Table 5 | The frequency of the Player 2s' (potential bribees) unethical decisions.

Never violate the rule (attempt = 0/15)			No more than one time (attempts <= 1/15)			Constantly violate the rule (attempts >= 8/15)		
Familiar context	**Unfamiliar context**	**Neutral context**	**Familiar context**	**Unfamiliar context**	**Neutral context**	**Familiar context**	**Unfamiliar context**	**Neutral context**
64/100	7/40	33/110	74/100	9/40	56/110	5/100	4/40	12/110
64.00%	17.50%	30.00%	74.00%	22.50%	50.91%	5.00%	10.00%	10.91%

Upon completion of the bribery game, all the participants are asked to complete an open-ended questionnaire[5] about their decisions and reasoning in the bribery game. According to the questionnaire, 54% of the participants in the familiar-context treatment indicate that corrupt behaviors are typically disapproval in the college context. Among them, only 6% engaged in corruption in the experiment.

Interaction Between the Paired Players

Next, we examine the interactions between the player 1 and the player 2 in a pair. We find that a reciprocal relationship between the two players is less likely to be established in the familiar-context treatment (**Table 6**). Specifically, 66% of the time corrupt activity never occurs (i.e., the player 1 never offers his or her partner a bribe, and the player 2 never violates the rules) in the familiar-context treatment. This percentage is 50.83% in the unfamiliar-context treatment (significantly lower than in the familiar-context treatment, Fisher exact test $p < 0.0001$) and 53.20% in the neutral-context treatment (significantly lower than in the familiar-context treatment, Fisher exact test $p < 0.0001$). Moreover, the player 2s in the familiar-context treatment are more likely to reject the bribery from the other person (**Table 7**). In aggregate, 81.58% of the bribes from player 1 are rejected in the familiar-context treatment. This percentage is 63.35% in the unfamiliar-context treatment (significantly lower than in the familiar-context treatment, Fisher exact test $p < 0.0001$) and 77.12% in the neutral-context treatment (significantly lower than in the familiar-context treatment, Fisher exact test $p = 0.075$).

In addition, we notice that some player 2s violate the game rule without an offer from the player 1. Such behavior could be understood as "signaling." The essence of bribery relationship is a mutual exchange of favors relying on trust and reciprocity. Since the two individuals in a pair may interact with each other repeatedly, in early stage of the game, the player 2 may have an incentive to signal the player 1 that he is interested in establishing such relationship (in the hope that the player 1 will start to offer bribe in later interactions). From **Table 8**, we can see that only 4.16% of the interactions are initiated by player 2 in the familiar-context treatment (i.e., player 2 violates the game rules without an offer from player 1). This proportion is also the lowest among the three treatments. Again, we do not see different results in the unfamiliar-context treatment and the neutral-context treatment.

The Dynamic of Decision Making

To further examine whether the familiar context is inversely predicting the probability of engaging in corruption, we perform several regression analyses. In particular, consider the following random effect model:

$$y_{1it} = \alpha + \beta_1.familiar_i + \beta_2.\ unfamiliar_i + \beta_3.male_i + U_{it} + E_{it}$$

where:

y_{1it} is player 1's bribery decision at period t. $y_{1it} = 1$ if individual i offers a payment to the other person in period t and $y_{1it} = 0$ if otherwise.

$familiar_i$ and $unfamiliar_i$ are dummy variables, they indicate if individual i is in a particular context. We use the neutral context treatment as the compare group.

$male_i$ is a dummy variable indicates if individual i is male.

U_{it} is the individual-specific random effect (i.e., between-entity error).

ε_{it} is the error term.

α is the constant term.

We first use the model above to estimate how the contexts affect player 1's decisions, results are reported in column (1) in **Table 9**. Next, we add period and the group an individual is in as additional controls, and then estimate the model again. Results are listed in column (2) and (3).

Following that, we conduct a similar analysis for the player 2s. In addition to the existing independent variables, we add the total amount of points being offered to the model, because player 2's decision might be influenced by how much payment was offered.

[5]In the questionnaire, we ask the participants were you engaged in any bribery relationship in the game, and what is the rationale of your decision.

Table 6 | Proportion of pairs never commit any unethical decision.

Corruption never happened (no bribery, no violation)		
Familiar context	**Unfamiliar context**	**Neutral context**
990/1,500	305/600	863/1,650
66%	50.83%	53.20%

Table 7 | Proportions of offers being rejected by player 2.

The player 2 rejected the bribery from the player 1		
Familiar context	**Unfamiliar context**	**Neutral context**
381/467	159/251	507/657
81.58%	63.35%	77.12%

Table 8 | Proportion of interactions initiated by player 2.

The player 2 violate the rule without any bribery from player 1		
Familiar context	**Unfamiliar context**	**Neutral context**
43/1033	44/349	130/993
4.16%	12.61%	13.09%

Accordingly, the model becomes:

$$y_{2it} = \alpha + \beta_1.familiar_i + \beta_2.\ unfamiliar_i + \beta_3.male_i + \beta_4 offer_{it} + U_{it} + +E_{it}$$

where:

y_{2it} is player 2's decision at period t. $y_{2it} = 1$ if individual i violates the allocation rule at period t and $y_{2it} = 0$ if otherwise.

$offer_{it}$ represents the amount of point being offered in period t.

All the other independent variables are the same as in the previous model. We also change the model specification by adding periods and groups as controls. The estimation results are shown in column (5)—column (7) in **Table 9**.

As **Table 9** suggest, the familiar-context dummy is negatively related to the probability of engaging in corrupt activities for both players. This effect is highly robust to changes in specification. In addition, we find that male participants are more likely to engage in corrupt behavior than female participants.

Lastly, we are interested in if there is any interaction between the context effect and the gender effect. We then add the interaction terms of the gender and the contexts into the models (i.e., *familiar male*, and *unfamiliar male*), and then estimate the parameters. Regression results with the interaction terms are reported in column (4) and column (8) in **Table 9**. From the results, we do not find any interaction effect between gender and context for player 1. The interaction effect for player 2 is quite interesting. In particular, the marginal effect of being in the familiar context is –0.063 for male and is –0.045 for female. This result suggests that the familiar context makes both male and female less likely to violate the funding allocation rule, but it has a stronger effect on male than on female. The marginal effect of being in the unfamiliar context is 0.098 for male and is 0.064 for female. That is to say, compare with the neural context, the unfamiliar context makes people more likely to violate the funding allocation rule. One possible reason of this observation is that people may have higher tolerance for bribing behavior in a business competition setting. However, this finding also adds a caveat to the application of loaded context: it brings extra confounding variable into the experiment and reduce experimental control.

DISCUSSION

In this paper, we use bribery game as an example to look at how different contexts impact unethical decision making in laboratory economic studies. Past studies on this topic (e.g., Cooper et al., 1999; Abbink and Hennig-schmidt, 2006; Barr and Serra, 2009) often compare the distinction between neutral context and loaded context. Our work tries to extend this dichotomous view by taking into account individual's emotional responses. In particular, we propose that emotional responses and psychological costs evoked by the framing is the key to understand when and why context may alter behavior.

We carry out three different treatments: a familiar-context treatment, an unfamiliar-context treatment, and a neutral-context treatment. In addition, we also use an independent survey to measure people's attitudes toward unethical behaviors in each of the contexts. Since attitude is considered to be an effective indicator of behavior, observations from the survey can help us better understand the motivation of decisions. In summary, we find that the survey respondents in the familiar context express the strongest negative attitudes toward corruption. Attitudes toward unethical behavior are the same in the neutral context and the unfamiliar context. Behaviorally, corrupt activities are substantially fewer in the familiar-context treatment than in the other two treatments. In the unfamiliar-context treatment and the neutral-context treatment, we do not find essential differences in the participants' behaviors.

From the attitude survey, our first finding is that most student respondents hold negative attitudes toward corrupt activities across all contexts (cheers for humanity!). Further, we find that although both the familiar context and the unfamiliar context are heavily loaded with suggestive words and background stories, the former clearly evoke stronger emotional responses. When we compare the attitudes in the unfamiliar context and the neutral context, we do not see statistically different outcomes. Result from the pairwise comparison analysis indicates that the negative attitudes are amplified by the familiar (i.e., college) context among the student respondents. Moreover, we find the students hold stronger negative attitudes toward player 2's unethical behavior. Result from the mixed measures ANOVA shows that there is no interaction effect between the context variable and the role variable. Note that the survey respondents and the bribery game participants are randomly chosen from the same population, their attitudes toward bribery should be similar. Hence, we anticipant the results from the attitude survey can help predict behavior in the lab. First of all, we anticipate that

Table 9 | Regression analysis with random effect models.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Player 1				Player 2			
Model	Random effect models				Random effect models			
Dependent variable	**Offer payment**	**Offer payment**	**Offer payment**	**Offer payment**	**Rule violation**	**Rule violation**	**Rule violation**	**Rule violation**
Familiar context	–0.076***	–0.074***	–0.074***	–0.078***	–0.086***	–0.086***	–0.085***	–0.045**
	(0.016)	(0.016)	(0.016)	(0.019)	(0.012)	(0.012)	(0.012)	(0.014)
Unfamiliar context	–0.015	–0.016	–0.016	–0.030	0.031	0.031	0.031	0.064**
	(0.025)	(0.022)	(0.022)	(0.027)	(0.017)	(0.017)	(0.017)	(0.020)
Neutral context (compare group)	–	–	–	–	–	–	–	–
Amounts of points being offered	–	–	–	–	0.001***	0.001***	0.001***	0.001***
					(0.000)	(0.000)	(0.000)	(0.000)
Male	0.220***	0.216***	0.221***	0.204***	0.065***	0.064**	0.067**	0.159***
	(0.021)	(0.021)	(0.021)	(0.034)	(0.015)	(0.016)	(0.016)	(0.025)
Familiar × male				0.015				–0.177***
				(0.049)				(0.036)
Unfamiliar × male				0.060				–0.125*
				(0.064)				(0.047)
Constant	0.352***	0.306***	0.264***	0.261***	0.115***	0.066	0.081	0.058
	(0.022)	(0.048)	(0.060)	(0.060)	(0.010)	(0.034)	(0.045)	(0.046)
Control for periods	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Control for groups	No	No	Yes	Yes	No	No	Yes	Yes
Observations	250	250	250	250	250	250	250	250
R-squared	0.036	0.050	0.056	0.056	0.067	0.075	0.077	0.083

**Indicate the result is statistically significant at p = 0.05 level.*
***Indicate the result is statistically significant at p = 0.01 level.*
****Indicate the result is statistically significant at p = 0.001 level.*

the player 2s should be less likely to engage in corrupt activities than the player 1s; Secondly, we predict that the participants in the familiar-context treatment should be less likely to engage in unethical behavior; Thirdly, we expect that the unfamiliar-context treatment and the neutral-context treatment should lead to similar behavior outcome.

Findings from the laboratory bribery game confirmed all these predictions. The experimental data suggest that the possibility of engaging in unethical behavior (offer bribe, accept bribe, or violate the rule) is obviously, in a statistical sense, the lowest in the familiar-context treatment, for both the player 1s and the player 2s. The unfamiliar-context treatment and the neutral-context treatment produce the same behavior. Observations from the bribery game, together with evidence from the attitude survey, suggest that in unethical decision-making experiments, emotional responses evoked by the context can be used to explain participants' behavior. Since the familiar context evokes the strongest emotional responses among all the contexts, the norm-consistent behaviors (i.e., behave with integrity) are more predictive in the familiar-context treatment than in the others.

With this insight, let's try to reconcile the mixed findings from past studies. Abbink and Hennig-schmidt (2006) conduct a bribery game experiment structured as interactions between "firms" and "public officers." Two different instructions are used, one with neutral descriptions and words and the other with suggestive words. The main finding from the study is that contexts did not change student participants' behavior, and the authors attribute this finding to the participants' lack of "expertise." A similar bribery game by Barr and Serra (2009) with University of Oxford student as participants find that when the participant plays as bribee, context has no effect on bribe acceptance; meanwhile, when the participant plays as briber, the context alters the behaviors. The authors attribute these results to participants' "intrinsic motivation." Here, we think the aforementioned "expertise" or "intrinsic motivation" can be good explanations in their individual studies, the emotional responses evoked by the framing might provide a generic explanation for all experiments of this type. Based on our results, the experimenters can expect to observe behavior change only when the emotional responses and psychological costs evoked by dishonest practices are different across contexts.

Alatas et al. (2009) invited real public officers in Indonesian to participate in a bribery game experiment. They find that when the public officer participants play as the bribees, they are less likely to engage in unethical behavior. One interpretation is that when participants play a role that is the same as their real-life identity, they know better the consequences of their decisions.

Therefore, familiarity with the experimental role would help prevent unethical behavior from happening. We consider that familiarity with the identity is a special case of familiarity with the context. Individual's pre-game experience is not just limited to participants' real-life identity. Rather, it is an integration of one's real-life role, expertise, knowledge, worldview, and all factors that contribute to the individual's self-concept. As long as a participant is familiar with the context, she will link the experimental task to her self-concept. Consequently, behaviors that violate certain social norm would trigger stronger emotional responses.

LIMITATIONS AND FURTHER DIRECTION

A major limitation of the current study is that the experimental design cannot fully reveal the mechanism underlying context effect in unethical decision-making experiments. For instance, there are at least two other possible explanations for the observed results. First, it is possible that the familiar context amplifies the cognitive dissonance evoked by engaging in corrupt activities. A key element that determines the intensity of the dissonance is personal involvement—the more attention one devotes to the unethical decision, the greater the dissonance experienced. As Elliot Aronson (1999) suggests: "...cognitive dissonance theory makes its strongest and clearest predictions when the self-concept of the individual is engaged. . . . dissonance is greatest and clearest when it involves not just any two cognitions but, rather, a cognition about the self and a piece of our behavior that violates that self-concept." Another possible mechanism is that the familiar context changes behavior via norm salience. Cialdini et al. (1990) propose that only "activated" norms impact people's behavior. In the current experiment, it is possible that the social norm in the familiar-context treatment is more salient to the participants. Accordingly, the participants are more likely to follow the dominant norms (i.e., behave with integrity). In future studies, it would be interesting to further investigate the how emotion, cognitive dissonance, and social norm jointly (or separately) determine behavior.

Unbalanced sample is another limitation of this paper. In particular, the sample we collected is unbalanced in two senses: First, the number of participants in the unfamiliar-context treatment is fewer than the other two treatments (40 in the unfamiliar-context treatment, 100 in the familiar-context treatment, and 110 in the unfamiliar-context treatment); Second, female participants account for 73% of the sample. To the first point, the highly unbalanced participant number is caused by administrative reasons that out of our control. Such sample may jeopardize the power of the statistic tests, especially when the variables of interest have different variances across treatments. The good news for us is that even in the unfamiliar-context treatment, the sample size (n = 40) is still sufficient for the statistical tests we used. Unbalanced sample may also cause unequal variances between samples. To address this concern, we compared the variances of bribing decisions in the familiar-context treatment and the unfamiliar-context treatment and find no significant difference. To the second point, the unequal number of male and female is caused by both the gender imbalance of the school and our recruitment strategy. At Jianghan University (where the experiment was conducted), female students account for 60% of the student population. Moreover, due to our sampling strategy, it turned out female students are more likely to reply to our recruitment email. Although gender effect is not the main focus of this paper, it would be better if we can use a more representative sample to conduct the study. In future studies, it would be interesting to systematically explore how gender affect the ways people interpret contexts.

Additionally, our conclusion would be much convincing with a counterfactual experiment in a non-student population. In the current study, the underlying assumption is that the familiar context (i.e., scholarship allocation) can give the student participants a more self-relevant, emotional experience than the unfamiliar context (bidding competition in business setting). This implies that with participants who is more familiar with the bidding competition in business setting but less familiar with academic setting, the contexts may lead to different behavior patterns. In the future, we hope to test our theory with different samples.

CONCLUDING REMARKS

Social scientists from different fields (economics, psychology, sociology, political science, etc.) apply various approaches to investigate the motivation of dishonest behavior in games, yet interdisciplinary cooperation in this area is surprisingly rare. This lack of communication may result from the disagreement on some issues concerning fundamental research methodology.

Our research contributes to one of the persistent, but still far from settled questions on experimental methodology: what is the role of experimental context in laboratory unethical decision-making research? In economics, it has become standard to present the experimental tasks using the neutral-context instructions, even in the experiments that emphasize the values and ethics embedded in the context. Because people worry about the loaded-context instructions may impact behavior in an unpredictable manner. Past studies in bribery games show that the loaded context alters people's behavior in some cases but produces the same result as the neutral context in others. Nevertheless, using loaded-context instructions has clear advantages. For instance, the participants can better learn the experimental tasks and be more engaged; the experimenters can explicitly associate the loaded contexts with the research questions to better understand participants' motivation. By identifying factors through which the loaded context impacts behavior, we can actually use the properly framed context as a way to gain ecological validity.

We do not think our results should be seen as a whole rejection of the neutral-context design approach. Instead, the point we are trying to make is that we should always keep our experimental design as simple as possible, but not simpler. In reality, moral obligation and emotional responses play vital roles in unethical decision making; therefore, it is important to simulate these

non-monetary payoffs while conducting laboratory experiments. In unethical decision-making experiments, we think it is inappropriate to assume that experimental manipulation can be studied apart from the cultural and social norms that define its meaning. When the values and ethics associated with the contexts are unclear to participants, we put the ecological validity and reproducibility of the experiment at risk.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Jianghan University, Committee of Research Ethics. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

SW designed and conducted the experiment and performed the statistical analysis. TC and SW together wrote the manuscript. Both authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.675319/full#supplementary-material

REFERENCES

Abbink, K., and Hennig-schmidt, H. (2006). Neutral versus loaded instructions in a bribery experiment. *Exp. Econ.* 9, 103–121. doi: 10.1007/s10683-006-5385-z

Ajzen, I., Fishbein, M., Lohmann, S., and Albarracín, D. (2018). "The influence of attitudes on behavior," in *The Handbook of Attitudes*, 197–255.

Alatas, V., Cameron, L., Chaudhuri, A., Erkal, N., and Gangadharan, L. (2009). Subject pool effects in a corruption experiment: a comparison of Indonesian public servants and Indonesian students. *Exp. Econ.* 12, 113–132. doi: 10.1007/s10683-008-9207-3

Alekseev, A., Chamess, G., and Gneezy, U. (2017). When and why contextual instructions are important. *J. Econ. Behav. Organ.* 134, 48–59. doi: 10.1016/j.jebo.2016.12.005

Alm, J., McClelland, G., and Schulze, W. (1992). Why do people pay taxes? *J. Public Econ.* 48, 21–38. doi: 10.1016/0047-2727(92)90040-m

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *Am. Econ. Rev.* 85, 891–904.

Andreoni, J., and Miller, J. (2002). Giving according to GARP: an experimental test of the consistency of preferences for altruism. *Econometrica* 70, 737–753. doi: 10.1111/1468-0262.00302

Armantier, O., and Boly, A. (2014). On the effects of incentive framing on bribery: evidence from an experiment in Burkina Faso. *Econ. Gov.* 15, 1–15. doi: 10.1007/s10101-013-0135-0

Aronson, E. (1992). The return of the repressed: dissonance theory makes a comeback. *Psychol. Inq.* 3, 303–311. doi: 10.1207/s15327965pli0304_1

Aronson, E. (1999). "Dissonance, hypocrisy, and the self-concept," in *Readings About the Social Animal*, eds J. Aronson and E. Aronson (New York, NY: Worth), 219–236.

Banerjee, R. (2016). On the interpretation of bribery in a laboratory corruption game: moral frames and social norms. *Exp. Econ.* 19, 240–267. doi: 10.1007/s10683-015-9436-1

Bardhan, P. (2006). The economist ' s approach to the problem of corruption. *World Dev.* 34, 341–348. doi: 10.1016/j.worlddev.2005.03.011

Barr, A., and Serneels, P. (2004). *To Serve the Community or Oneself: The Public Servant ' s Dilemma Policy Research Working Paper;No.3187*. Washington, DC: World Bank.

Barr, A., and Serra, D. (2009). The effects of externalities and framing on bribery in a petty corruption experiment. *Exp. Econ.* 12, 488–503. doi: 10.1007/s10683-009-9225-9

Blanchette, I., and Caparos, S. (2013). When emotions improve reasoning: the possible roles of relevance and utility. *Think. Reason.* 19, 399–413. doi: 10.1080/13546783.2013.791642

Carpenter, J., Connolly, C., and Myers, C. K. (2008). Altruistic behavior in a representative dictator experiment. *Exp. Econ.* 11, 282–298. doi: 10.1007/s10683-007-9193-x

Chou, E., McConnell, M., Nagel, R., and Plott, C. (2009). The control of game form recognition in experiments: understanding dominant strategy failures in a simple two person "guessing" game. *Exp. Econ.* 12, 159–179. doi: 10.1007/s10683-008-9206-4

Cialdini, R., Raymond, R., and Carl, K. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *J. Pers. Soc. Psychol.* 58:1015. doi: 10.1037/0022-3514.58.6.1015

Cooper, B. D. J., Kagel, J. H., Lo, W. E. I., and Gu, Q. L. (1999). Gaming against managers in incentive systems: experimental results with chinese students and chinese managers. *Am. Econ. Rev.* 89, 781–804. doi: 10.1257/aer.89.4.781

Eckel, C. C., and Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games Econ.* 191, 181–191. doi: 10.1006/game.1996.0081

Fischbacher, U. (2007). z-Tree: zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10, 171–178. doi: 10.1007/s10683-006-9159-4

Goel, V., and Vartanian, O. (2011). Negative emotions can attenuate the influence of beliefs on logical reasoning. *Cogn. Emot.* 25, 121–131. doi: 10.1080/02699931003593942

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108. doi: 10.1126/science.1062872

Griggs, R., and Cox, J. (1982). The elusive thematic materials effect in Wason's selection task. *Br. J. Psychol.* 73, 407–420. doi: 10.1111/j.2044-8295.1982.tb01823.x

Laury, S. K., and Taylor, L. O. (2008). Altruism spillovers: are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good? *J. Econ. Behav. Organ.* 65, 9–29. doi: 10.1016/j.jebo.2005.05.011

Smith, V. (1976). Experimental economics: induced value theory. *Am. Econ. Rev.* 66, 716–731.

Wason, P. C., and Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Q. J. Exp. Psychol.* 23, 63–71. doi: 10.1080/00335557143000068

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 30 July 2021
doi: 10.3389/fpsyg.2021.684226

Masculinity and Lying

Marc Vorsatz[1], *Santiago Sanchez-Pages*[2*] *and Enrique Turiegano*[3]

[1] Department of Economic Analysis, National University of Distance Education (UNED), Madrid, Spain, [2] Department of Political Economy, King's College London, London, United Kingdom, [3] Department of Biology, Autonomous University of Madrid, Madrid, Spain

Dishonesty in communication has important economic implications. The standing literature has shown that lying is less pervasive than predicted by standard economic theory. We explore whether biology can help to explain this behavior. In a sample of men, we study whether masculine traits are related to (dis)honesty in a sender-receiver game. We study three masculine physical traits: the second-to-fourth digit ratio, facial morphometric masculinity and the facial width-to-height ratio. These biomarkers display significant associations with lying and deception in the game. We also explore the extent to which these effects operate through social preferences or through beliefs about the behavior of receivers.

Keywords: lying, deception, cheap-talk, masculinity, testosterone
JEL codes: C72, C91, D83, D87

OPEN ACCESS

Edited by:
Agne Kajackaite,
Social Science Research Center Berlin, Germany

Reviewed by:
Roel Van Veldhuizen,
Social Science Research Center Berlin, Germany
Kobe Millet,
Vrije Universiteit Amsterdam, Netherlands

***Correspondence:**
Santiago Sanchez-Pages
santiago.sanchez-pages@kcl.ac.uk

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *23 March 2021*
Accepted: *21 June 2021*
Published: *30 July 2021*

Citation:
Vorsatz M, Sanchez-Pages S and Turiegano E (2021) Masculinity and Lying. Front. Psychol. 12:684226. doi: 10.3389/fpsyg.2021.684226

1. INTRODUCTION

Truthful communication is a pillar of human interactions. Economic exchanges rely on language being trustworthy. Many buyers consult financial advisors before acquiring stocks or probe sellers on the quality of their goods. Honest communication is also crucial in policy making. Central banks make pronouncements which influence the actions of investors and stock traders. Regulatory bodies consult private entities before setting new standards. A third area where communication is crucial is organizations. Division managers, for example, report local market conditions to the their superiors who then use this information to devise their plans for the firm.

But better-informed agents often have incentives to misrepresent what they know in order to alter the decision-making process in their favor. Central bankers have an incentive to manipulate economic expectations (Stein, 1989). Private firms hired by financial agencies may recommend the adoption of standards to their own advantage (Melumad and Shibano, 1994). Low-level managers may bias their reports to maximize the profits of their division rather than of the entire firm (Dessein, 2002). Given that dishonesty in communication severely undermines trust (e.g., Gawn and Innes, 2018), it is of great importance to study its prevalence and determinants.

The experimental literature on strategic information transmission has shown that individuals engage in truthful communication above standard game-theoretical predictions (e.g., Gneezy, 2005; Cai and Wang, 2006). This literature also highlights that purely monetary cost-benefit calculations cannot explain such behavior. A substantial proportion of individuals refuse to tell lies that may benefit them at the expense of others (e.g., Sanchez-Pages and Vorsatz, 2007; Hurkens and Kartik, 2009), even if these lies can lead to Pareto-superior allocations (Erat and Gneezy, 2012)[1].

In this paper, we offer an exploratory study of the role biological factors play in explaining the individual heterogeneity observed in honesty in strategic communication. In particular, we focus on masculine physical traits. The development of masculine physiology-related traits during key life stages is associated to organizational changes in the neural circuitry of the brain which can in turn affect behavior (e.g., Sisk and Zehr, 2005).

[1] For a survey of the experimental literature on non-strategic communication, see Abeler and Raymond (2019).

The study of masculine traits is particularly relevant in the context of strategic information transmission because of two reasons. First, men are typically overrepresented in environments such as firms, finance and policy making where communication of this sort is pervasive[2]. Second, a growing body of literature has shown that biological mechanisms, and sexual hormones in particular, influence moral decision-making (e.g., Capraro, 2018). As strategic communication often entails the choice between truth-telling (an almost universal moral principle) and self-serving lies (widely deemed as antisocial), masculine traits are likely to relate to this choice.

The experimental literature has shown that individuals with more masculine facial features are less trustworthy (Stirrat and Perrett, 2010) and more likely to cheat in non-strategic settings (Haselhuhn and Wong, 2012; Geniole et al., 2015). Jia et al. (2014) find that CEOs with more masculine facial features are more likely to be subject to external audits and to be accused of financial wrongdoings. But to the best of our knowledge, the present paper is the first to explore whether markers of masculinity correlate with lying and deception in strategic communication.

We conduct a laboratory experiment based on the sender-receiver game in Sanchez-Pages and Vorsatz (2007) with a sample of 168 males. Participants are matched in pairs; one is assigned to be the sender, the other to be the receiver. Only the sender is informed about the state of the world which determines players' payoffs conditional on the action the receiver will take later on. The sender sends a non-verifiable message to the receiver about the state of the world. The receiver then decides which action to take and payoffs are realized. Preferences are opposed: the best outcome for the sender is the worst for the receiver and viceversa. As a result, the standard game-theoretical prediction is that senders' messages are entirely uninformative.

In our analysis, we make use of the distinction between lying and deception introduced by Sobel (2020). Lying refers to the content of messages. Deception entails misleading others to obtain an advantage, which in our design can be achieved by lying when one expects to be trusted but also by telling the truth when one expects to be distrusted (Sutter, 2009).

We study how lying and deception by senders in this game correlate with a set of masculine physical traits. Two of the biomarkers we consider are related to testosterone exposure at two developmental periods, *in utero* (second-to-fourth digit ratio) and during puberty (facial morphometric masculinity). The third one has been associated to antisocial and dominance behavior (facial Width-to-Height ratio). We discuss these markers in detail and the debates about their relevance in the following section.

Our results suggest the existence of a significant relationship between the masculine physical traits we study and (dis)honesty in strategic communication. We find that individuals exposed to higher levels of testosterone *in utero* and with more masculine facial features are more likely to send truthful messages. However, we also find that the latter engage more often in deception through truth-telling. In contrast, individuals exposed to higher prenatal levels of testosterone seem to display a stronger lie aversion as they are more likely to engage in costly truth-telling, i.e., send a truthful message when it is expected to be trusted.

Finally, we explore whether these associations between lying and the masculine physical features we consider operate mostly through social preferences (e.g., lying aversion) or through beliefs about the behavior of receivers. Results suggest that preferences are the main drivers of these effects.

The present paper contributes to the rapidly expanding literature on the influence of biometric traits and sexual hormones on economic behavior. Studies in this area have shown that differences in circulating and basal levels of sexual hormones influence risk preferences (e.g., Garbarino et al., 2011), social preferences (Buser, 2012a; Sanchez-Pages and Turiegano, 2013), bidding in auctions (Chen et al., 2013; Pearson and Schipper, 2013; Sanchez-Pages et al., 2014; Schipper, 2015), cooperation in social dilemmas (Sanchez-Pages and Turiegano, 2010; Cecchi and Duchoslav, 2018) and willingness to compete (Buser, 2012b; Wozniak and Harbaugh, 2014). The closest papers to ours in this strand of the literature have studied the effect of administered testosterone on non-strategic misreporting (Wibral et al., 2012) and on strategic gambling in poker (van Honk et al., 2015). In contrast to these two papers, we consider stable physiology-related traits rather than hormone infusions.

There are two other papers related to ours which explore the correlation between biological data and honesty in sender-receiver games. Using eye-tracking techniques, Wang et al. (2010) observed that senders look disproportionately at the payoffs corresponding to the true state of the world and that their pupils dilate when they send deceptive messages. On the other hand, Volz et al. (2015) studied the neural correlates of dishonesty using fMRI and found that brain activation patterns can reveal whether the sender intends to deceive the receiver.

Finally, our paper relates to the literature on gender differences in lying in sender-receiver games. If masculinity and femininity are viewed as a continuum, we would expect our results to reflect to some extent any gender differences observed in these studies. A recent meta analysis on deception games[3] by Capraro (2018) showed that male senders are more likely to lie than female senders when lies benefit them at the expense of the receiver and when lies hurt the sender but benefit the receiver. We obtain results along these lines when studying deception and costly truth-telling, in the sense that individuals exposed to more prenatal testosterone and with more masculine facial features engage more often in these behaviors. However, for the purpose of our study, gender is a too coarse marker of physiological differences as it is binary and it is heavily influenced by socialization.

[2]Only 1 in 3 financial advisors in the US, 1 in 4 board members in European companies and 1 in 5 US congresspersons are women (UN WOMEN, 2014; Bureau of Labor European Comission, 2016; Statistics, 2017).

[3]Deception games (e.g., Gneezy, 2005) differ from the family of sender-receiver games our design belongs to in that the receiver does not know the set of payoffs in deception games but does in ours.

2. MASCULINE PHYSICAL TRAITS

Masculinity can be defined as "a set of physical and behavioral traits that are male typical" (Lippa, 2016). These traits can distinguish men from women and/or order men by their degree of male typicality. Masculinity is thus not a latent trait but a set of dimensions that are typical of men. From all possible masculine traits proposed in the literature, the ones we choose in our study are based on rather stable physical features.

A widely studied physiological masculine trait is exposure to androgens -testosterone in particular- during key phases of development. Androgens produce distinctive changes in the male body, such as greater musculoskeletal development and the appearance of secondary sexual characteristics. More importantly, they have organizational effects on the brain, that is, they modify neural structures and can therefore influence adult behavior (Sisk and Zehr, 2005). In particular, testosterone seems to affect the structure of the amygdala, a cluster of neurons responsible for emotional reactions such as responses to interpersonal challenges and threats (van Honk et al., 2012).

There are two stages of development during which androgen exposure has crucial organizational effects on the brain: the prenatal period and puberty (Schulz et al., 2009; Berenbaum and Beltz, 2011). Androgens levels during these two periods have been proxied in the literature with two types of morphological features, the second-to-fourth digit ratio (2D:4D) and facial morphometric masculinity (fMM).

2.1. 2D:4D

The second-to-fourth digit ratio (2D:4D) is the ratio between the length of the index and the ring fingers. The available evidence suggests that the 2D:4D ratio is related to the ratio of amniotic testosterone/estrogen concentrations (Zheng and Cohn, 2011; Swift-Gallant et al., 2020). A lower 2D:4D ratio indicates higher relative exposure to masculine sexual hormones during foetal development. Men across countries have shorter ratios than women (Hönekopp et al., 2007; Grimbos et al., 2010). These differences are already present in human embryos (Galis et al., 2009). The underlying mechanism seems to be that both digit growth and the development of primary sexual characteristics are influenced by the Hox genes (Manning et al., 1998). Although early studies showed a correlation between 2D:4D and circulating (current) testosterone in adults, more recent ones have conclusively rejected that association (e.g., Honekopp and Watson, 2010).

Several studies have cast doubts on the validity of 2D:4D as a proxy for prenatal testosterone exposure. These studies find no correlation between 2D:4D and testosterone levels in umbilical blood or mother's blood (Hickey et al., 2010; van Leeuwen et al., 2020). However, these methods to measure foetal hormonal levels are imprecise. In mammals, testosterone levels at birth measured from the umbilical cord are substantially lower than during pregnancy. In addition, the role of the placenta in the process of blood exchange with the mother is to regulate the hormone levels the foetus is exposed to. In contrast, there is abundant indirect evidence of 2D:4D correlating with prenatal testosterone coming from studies of patients with congenital adrenal hyperplasia, Klinefelter's syndrome and androgen insensitivity syndrome (for meta analyzes, see Honekopp and Watson, 2010; Richards et al., 2020; Sadr et al., 2020). This evidence plus the lack of competing explanations (Swift-Gallant et al., 2020) lead us to believe that that 2D:4D remains the best available proxy for testosterone levels during foetal development[4].

The relationship between 2D:4D and strategic behavior is not fully understood yet. Earlier studies showed that men with lower 2D:4D are more prosocial and cooperative (Millet and Dewitte, 2006, 2009; van den Bergh and Dewitte, 2006). Later, large studies found no evidence of sexual hormones affecting decision making (e.g., Zethraeus et al., 2009; Ranehill et al., 2018) and economic preferences (Neyse et al., 2021). Recent studies suggest that the role of 2D:4D is very context-dependent (e.g., Ryckmans et al., 2015; Cecchi and Duchoslav, 2018), especially when the context challenges individual status (Millet, 2011; Manning et al., 2014; Millet and Buehler, 2018). One possible reason for this is that circulating testosterone, which seems to activate the neural circuitry affected by prenatal testosterone exposure, varies with environmental stimuli (Montoya et al., 2013). Several studies support the idea that the effects of testosterone levels are modulated by circuits established by prenatal exposure to sex hormones (van Honk et al., 2011a, 2012; Buskens et al., 2016). Evidence from brain imaging underscores that prenatal exposure to sex hormones matters for both neural and behavioral manifestations of testosterone in adult behavior (Chen et al., 2016). Meta analyzes linking this trait with violent and aggressive behaviors (Honekopp and Watson, 2011; Turanovic et al., 2017) also suggest that 2D:4D may affect economic decisions.

Following the standard procedures in the literature (e.g., Pearson and Schipper, 2012) we scanned both hands of all participants. Using the TPS morphometric software (Rohlf, 2015) on these images, two researcher assistants took digit length measures of the second and fourth digits of each hand from the flexion crease proximal to the palm to the top of the digit. Interrater correlation was $r = 0.747$ for the right hand and $r = 0.732$ for the left hand. The average of the four values is the first of our markers of interest. In our analyzes below we have transformed the variable so that higher values are meant to signify higher exposure to prenatal testosterone.

2.2. Facial Morphometric Masculinity

Although there are no direct measures relating pubertal hormone levels to the facial shape, it is well established that higher androgens levels during puberty are related to facial bone size and certain facial features (e.g., Marečková et al., 2011). Given that testosterone exposure in adolescence creates sex differences in the face shape, another masculine physical trait to be considered is the degree of difference between a man's face and a female face of reference.

[4]Two other criticisms of this measure is that sex differences in 2D:4D might be the result of an allometric shift in shape (Kratochvíl and Flegr, 2009; Lolli et al., 2017) and that 2D:4D changes across life (McIntyre et al., 2005; Trivers et al., 2006). However, very recently, Butovskaya et al. (2021) have observed in a very large sample (>7000) across different ethnicities and ages that 2D:4D is stable during life and that sex differences persist after controlling for allometry issues.

FIGURE 1 | Facial measures used to compute fWHR.

There is a wide variety of methods to measure facial dimorphism[5]. Specifically, we employ facial morphometric masculinity (fMM), which is in line with others employed in the literature (van Dongen, 2014; Ekrami et al., 2021). In previous studies, we found an association of fMM with rejections of low offers in the ultimatum game (Sanchez-Pages and Turiegano, 2013), and more aggressive bidding in the first price auction (Sanchez-Pages et al., 2014). One key advantage of morphometric methods is that they gather information from the entire facial shape rather than from specific distances or angles. Specifically, fMM corresponds to the Procrustes distance between the shape of the participant's face and a reference female face. This distance is computed from a number of landmark coordinates placed on the facial image. Two research assistants independently placed 39 of these landmarks (LMs) in the image resulting from averaging the two photographs of each subject. These LMs can be unambiguously identified in every photo (see **Figure 1**) and are thus comparable across individuals. Since we are interested in the changes in the facial shape induced by the exposure to testosterone during puberty, LMs were not placed on soft parts of the face, which are more prone to changes during life. We built the female reference image by averaging the photos of 100 females of similar age and background to the subjects in our sample. The TPS software (Rohlf, 2015) computed a fMM score for each individual with higher scores indicating a higher distance between the subject's face and the average female face, that is, higher facial masculinity. This software also implements a correction accounting for LM placement error across researchers. The resulting score is our second trait of interest.

[5]We surveyed many of them in Sanchez-Pages et al. (2014).

2.3. Facial Width-To-Height Ratio

The available evidence suggests that men with certain facial features tend to be more aggressive and less prosocial (Geniole et al., 2015; Haselhuhn et al., 2015). Some of these features are based on raters' perceptions whereas others are calculated from physiological markers. Perceived masculinity is problematic because subjective judgments tend to be influenced by perceived health and skin color. Objective measures are better suited for our purposes. The facial width-to-height ratio (fWHR), first described by Weston et al. (2007), is probably the most popular among these because it is very easy to compute: fWHR is the ratio between the width and the height of the face.

Individuals with higher fWHR engage more often in threat and dominance behaviors and are perceived as more threatening and dominant (Geniole et al., 2014). They also are more prone to engage in antisocial behavior (for meta analyzes see Haselhuhn et al., 2014, 2015) and display superior deception skills (Matsumoto and Hwang, 2021). Elite hockey players with higher fWHR are sanctioned with more penalty minutes over the season (Carré and McCormick, 2008). Since fWHR is also associated with dominance in non-human primates (Lefevre et al., 2014), some authors have argued that the trait serves, or at least served in our evolutionary past, as a signal of aggression and dominance in inter-male competition (Geniole et al., 2015; Wang et al., 2019). In economic games, this marker has been shown to correlate with the propensity to exploit others in the trust game (Stirrat and Perrett, 2010; Sanchez-Pages et al., 2014).

To construct the fWHR, we took two full frontal facial color photographs of our subjects at standardized light and distance conditions. They were asked to remove any facial adornments and were carefully instructed to look into the camera with a neutral expression. We later converted these images to a 8-bit gray-scale format. Using the TPS morphometric software and following the method described in Weston et al. (2007), two research assistants measured the maximum horizontal (bizygomatic) distance from the left to the right cheekbone and divided it by the vertical distance between the lip and brow (see **Figure 2**). The correlation between their measures was $r = 0.840$. The average of the ratios obtained by the two researchers is our third and final masculine trait of interest.

It is important to note at this point that the available evidence strongly suggests that fWHR is not sexually dimorphic (e.g., Kramer, 2017). This casts some doubts on the value of fWHR as a masculine trait. The literature suggests that this lack of sex differences might be driven by the influence of body weight on the facial shape. For that reason, we also collected height and weight measurements of our subjects to construct their Body Mass Index (BMI) and we included it as a control in all our specifications.

3. THE EXPERIMENT

3.1. Design

3.1.1. Equilibrium Predictions

Our experimental design is based on Sanchez-Pages and Vorsatz (2007). First, nature randomly selects one of two tables, A or B, with equal probability. The chosen table $\theta \in \{A, B\}$ determines

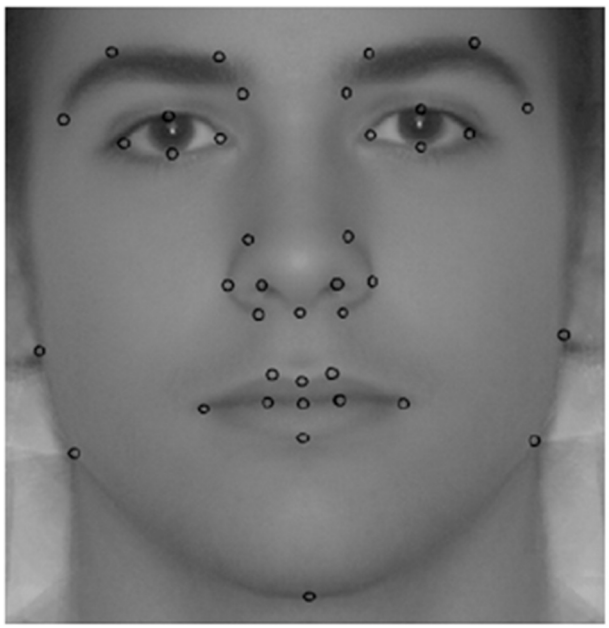


FIGURE 2 | Facial landmarks used to compute fMM.

how payoffs will be realized. There are two players, the *sender* and the *receiver*. Only the sender is informed about θ. After being informed about the table selected, the sender sends a message to the receiver telling him[6] which table nature selected. Formally, the sender chooses a mixed strategy profile $\{p(m \mid \theta)\}_{\theta=A,B}^{m=A,B}$ where $m \in \{A, B\}$ is the message sent with $p(A \mid \theta)+p(B \mid \theta) = 1$.

The receiver observes the message m and must choose a mixed strategy over his available actions, A and B. The action taken $s \in \{A, B\}$ is relevant for both players as it determines in conjunction with the table selected θ the payoffs they receive. The payoff structure is of divergent interests (Crawford and Sobel, 1982) as shown in the matrices in **Table 1**. This means that the best action for the receiver is the one that matches the table selected, i.e., $s = \theta$. The opposite holds for the sender. Lying occurs when the sender sends the message "The table selected is **Table 1A** (B)" when nature has actually selected **Table 1B** (A), i.e., when $m \neq \theta$.

The receiver holds a belief profile $\{\mu(s \mid m)\}_{s=A,B}^{m=A,B}$, where $\mu(m \mid m)$ is the probability with which the receiver believes that the message m is truthful and action $s = m$ will indeed earn him the highest payoff. Note that $\mu(A \mid m) + \mu(B \mid m) = 1$. Denote the mixed strategy of the receiver as $\{q(s \mid m)\}_{m=A,B}^{s=A,B}$, where $q(A \mid m) + q(B \mid m) = 1$. As it is customary in the literature, we will interpret that a receiver trusted (or followed) the sender's message if he took the action that maximized his payoff if the message was truthful, i.e., when $s = m$.

Under these preferences, the standard game-theoretical prediction is that the set of sequential equilibria of the game are all "babbling": Senders send each message with the same probability regardless of the table chosen, i.e., $p(A \mid A) = p(A \mid B) = p \in [0, 1]$, meaning that they lie with 50% probability. This renders messages completely uninformative, so receivers' posterior beliefs remain identical to the prior, i.e., $\mu(A \mid m) = \mu(B \mid m) = \frac{1}{2}$. Given this behavior on the part of senders, receivers should follow messages with 50% probability. Note that 1) risk attitudes do not alter this set of predictions and 2) the babbling equilibrium with $p = \frac{1}{2}$ is the unique logit agent quantal response equilibrium of the game (McKelvey and Palfrey, 1995)[7].

TABLE 1 | Payoff matrices.

Action A	Action B
Table A	
40 for the sender	100 for the sender
100 for the receiver	40 for the receiver
Table B	
100 for the sender	40 for the sender
40 for the receiver	100 for the receiver

3.1.2. A Behavioral Taxonomy

Let us now introduce some behavioral considerations. Suppose that the sender expects the receiver to trust his message with more (less) than 50% probability. In that case the sender should tell a lie (the truth) under standard preferences (and independently of his risk attitude). It is at this point where we should make a crucial distinction between lying and deception: whereas lying is related to the *content* of the message, deception relates to the *outcome* the message is trying to induce (Sobel, 2020). Obviously, lying occurs in our design when a subject sends an untruthful message, i.e., $m \neq \theta$. On the other hand, we will say that a sender engages in deception when he sends a message aiming to induce the receiver to take the inferior action, that is, the best action for the sender (note again that we are assuming that receivers take the action they believe maximizes their own payoff). Therefore, a sender can be deceptive in our experiment either by lying when he expects the receiver to trust his message with more than 50% chance or by telling the truth when he expects the receiver to follow his message with less than 50% chance[8].

On the other hand, a sender who tells the truth when he expects the receiver to trust his message with more than a 50% chance is not maximizing his expected payoff. We will say that this sender is a *strong truth-teller*. A sender who tells a lie when he expects the receiver to distrust his message with more than a 50% chance is not maximizing his expected payoff either and in addition he is lying. Given that such sender is paying a monetary cost and probably a psychic (lying) cost also to make the receiver obtain a higher payoff, we refer to this sender as an *altruistic liar*[9] **Table 2** summarizes this behavioral taxonomy.

[6]Because our sample only comprises men, we use male pronouns throughout the paper.

[7]See Sanchez-Pages and Vorsatz (2007) for a formal proof of these results.

[8]Sutter (2009) called the latter *sophisticated deception*.

[9]Erat and Gneezy (2012) call these lies *altruistic white lies*.

TABLE 2 | Behavioral taxonomy by messages and beliefs.

Message \ Belief	Trust < 50%	Trust > 50%
Truthful	Deception	Strong truth-telling
Untruthful	Lying	Lying
	Altruistic lying	Deception

3.2. Procedures

The study was conducted with undergraduate students at the Universidad Autónoma de Madrid (UAM), Spain, in early 2016. It was approved by the UAM Research Ethics Committee (reference CEI 62-1086). Subjects were recruited from the subject pool of the Madrid Laboratory for Experimental Economics (MADLEE) and with posters and flyers distributed within the Faculty of Sciences where the experimental sessions took place. The invitations and promotional materials mentioned that participants would be taken images of their faces and their hands and that these images could not be linked to any personal information. No mention to the all-male nature of the experiment was made during the recruitment process or the sessions.

A total of 168 males participated in 10 sessions composed by 12–24 subjects each. This sample size was meant to detect the associations between economic behavior and masculine physical traits identified in previous works[10]. All subjects except one identified themselves as Caucasian; we excluded that subject from our analysis as the fMM measure requires ethnic homogeneity. Another subject did not fill the belief elicitation question in one treatment. The sessions comprised two experiments run in a fixed order with a break in the middle to collect participant's morphometric data. After these measures were collected subjects were free to go if they preferred to not participate in that second component, whichwas unrelated to the one discussed here[11]. The duration of the experiment presented in this paper was 40–60 min, including the collection of physiological measurements.

Subjects were called one by one to the lab and took sit at individual tables. Instructions were then read aloud (see the **Supplementary Material**). During this debriefing, participants were reminded that experimenters were to take photos of their faces and scans of their hands after the session and that these images would be anonymized. Participants were invited to leave the experiment at that point if they did not consent with these images being taken; they could keep the show-up fee if they left. They were also told they were free to leave the session at any later stage.

Subjects participated in two treatments administered in a fixed order. Subjects received no feedback between them. First, they participated in a *control* treatment, where they played the sender-receiver game described above. After that, they played a *punishment* treatment, a version of the Punishment Game in Sanchez-Pages and Vorsatz (2007). That game is identical to the one in the control treatment up to the point where the receiver takes his action. Before payoffs are realized, the receiver is informed about the payoff outcome and the table selected by nature and he is given the option to accept the resulting payoff distribution or to reduce his and the sender's payoff to zero. It is easy to see that the set of sequential equilibria of this game under standard preferences is identical to the one in the control treatment as no purely payoff-maximizing receiver would reduce his own payoff.

Participants made their choices in the two roles within each treatment. When choosing as receivers, subjects observed a message from the sender and chose their action. When choosing as senders, they where informed about the table selected by nature and decided which message to send to the receiver. We used a simplified version of the strategy method to elicit these decisions.[12] Rather than eliciting their choices for each table (for senders) and each message (for receivers), participants were just presented one instance and were told that experimenters would infer from their choice that their behavior would have been analogs in the other eventuality. That is, that we would interpret that senders who lied (were honest) when the table selected was A would have also lied (been honest) if the table selected had been B, and viceversa. Similarly, when playing as receivers, subjects were told we would interpret that if they followed (distrusted) a message saying that the table selected was A, they would have equally followed (distrusted) a message reporting that the table selected was B (and viceversa).

In the punishment treatment, receivers were presented four additional choices. They had to decide whether they would accept or reduce the payoffs to zero for each of the four possible histories of the game, i.e., $\{m = \theta, s = \theta\}$, $\{m \neq \theta, s = \theta\}$, $\{m = \theta, s \neq \theta\}$ and $\{m \neq \theta, s \neq \theta\}$.

Participants recorded their choices in paper booklets, one booklet per treatment. Each page of the booklet presented a decision round. Subjects were not allowed to move to a new decision round until all participants had finished with that round. In each treatment, we elicited beliefs about the percentage of senders in the session who would send truthful messages and the

[10] Stirrat and Perrett (2010) detected a Spearman correlation $r_s = -0.34$ between fWHR and trustworthiness. We needed 106 observations to detect that correlation in the present study with $\alpha = 0.05$ and $\beta = 0.90$. Sanchez-Pages and Turiegano (2013) detected an effect of size $d = 0.409$ of fMM on rejections of a low offer in the Ultimatum Game; assuming that 56% of messages would be truthful, as observed in Sanchez-Pages and Vorsatz (2007), 210 observations were needed to detect the same effect size under identical α and β. Finally, Sanchez-Pages et al. (2014) found a Pearson correlation $\rho = -0.164$ between fMM and bids in a first-price auction. We needed 314 observations to detect that correlation.

[11] That second experiment was on cooperation and third party punishment. Results are reported in Rodriguez-Ruiz et al. (2019), although subjects in the present study who took part in that experiment constituted a small part of the overall sample. For the purpose of that second experiment, we collected data on upper body strength, fluctuating asymmetry, self-perceived attractiveness and sexual orientation. We do not test or report these measures as they are not masculinity traits nor have a previously described association with dishonesty.

[12] The evidence on the differential effect of the strategy method and the direct response method in sender-receiver games is scant and mixed. López-Pérez and Spiegelman (2013) found no significant differences whereas Minozzi and Woon (2020) observed increased overcommunication under the strategy method. We find that average behavior is similar to that in Sanchez-Pages and Vorsatz (2007), who employed the direct response method. Note, however, that we are interested on whether truth-telling relates to masculinity markers rather than on truth-telling rates themselves.

TABLE 3 | Descriptive statistics.

Variable	Mean	Std dev	Min	Max	*n*
1. 2D:4D	0.961	0.027	0.887	1.059	168
2. fWHR	1.929	0.116	1.584	2.260	168
3. fMM	0.093	0.021	0.047	0.156	167
4. BMI	23.458	3.008	15.570	34.478	168
5. Age	21.940	2.299	18	29	168

TABLE 4 | Frequency of behavioral types by treatment.

Behavior \ Treatment	Control (%)	Punishment(%)
Lying	37.5	25
Deception	55.8	46.5
By lying	29.2	17.6
By truth-telling	26.6	28.9
Strong truth-telling	35.7	47.9
Altruistic lying	8.5	5.6

percentage of receivers who would follow the sender's message: We paid 100 extra points to subjects whose guess was within a 5 percentage points band around the actual target percentage. In the punishment treatment, we also elicited beliefs about the percentage of receivers participants expected to reduce payoffs in each of the four possible histories. The order of decision rounds within each treatment was: (1) choice as receiver; (2) choice as sender; (3) elicitation of beliefs about expected truth and trust rates. In the punishment treatment there were two additional rounds: (4) punishment choices and (5) elicitation of beliefs about punishment rates.

At the end of each session, participants were called one by one to an adjacent room where morphometric measurements were taken in private by one experimenter and two research assistants. After this, one treatment was selected for payment. Roles were randomly assigned within each anonymously matched pair of participants and payoffs were determined according to their decisions. Subjects were paid their earnings in cash in addition to a 5€ show-up fee for this experiment. The exchange rate between points in the experiment and money was 100 points=1€. Average earnings were 7.82€.

4. RESULTS

4.1. Descriptive Statistics

Table 3 contains descriptive statistics for the three masculine physical traits we consider. They correlate only slightly. The Spearman's correlation coefficient between fMM and fWHR is 0.126 ($p = 0.099$, $n = 167$) and between fMM and 2D:4D is -0.128 ($p = 0.105$, $n = 167$). No significant correlation exists between fWHR and 2D:4D. These weak correlations are in line with previous studies (Sanchez-Pages et al., 2014), and were expected since masculinity is not a latent concept but a set of traits typical of males. fWHR is non-dimorphic so it was not expected to correlate with the other two traits, which are sexually dimorphic. 2D:4D is a measure of prenatal testosterone and fMM of adolescent testosterone. These two periods of exposure to sexual hormones independently influence adult behavior (Berenbaum and Beltz, 2011)[13].

[13]The evidence on prenatal hormone effects in human and non-human primates shows that androgens are the masculinizing agent whilst oestrogens affect both sexes *in utero*. In contrast, there are dramatic sex differences in both androgen and oestrogen exposure during puberty. It has been proposed that puberal exposure acts by refining the organizational effect of sexual hormones during early development (Montoya et al., 2013).

4.2. Aggregate Behavior

The percentages of untruthful messages in the control (37.5%) and punishment (25%) treatments were well below the theoretical prediction of 50% (Proportion test, $p = 0.001$ and $p < 0.001$, respectively, $n = 168$). They were also significantly different from each other ($p = 0.013$, $n = 336$). Trust rates were 63.1% in the control treatment and 58.9% in the punishment one. Both rates were significantly higher than 50% ($p < 0.001$ and $p = 0.020$, respectively, $n = 168$) and similar to those in Sanchez-Pages and Vorsatz (2007), but not different from each other. Beliefs about trust rates were very accurate, 61.4% and 58.1% in the control and punishment treatments, respectively. The distributions of beliefs were not different across treatments (Mann-Whitney, $p = 0.245$, $n = 335$).

Table 4 summarizes the proportion of subjects in each category of our behavioral taxonomy by treatment. Note that frequencies do not add up to 100% vertically because some behavioral classifications overlap. The first result stemming from this table is that deception is more common than lying. Altruistic lies are rare whereas deception by truth-telling is quite frequent and seems unaffected by the threat of punishment. The second result is that the possibility of punishment reduces the frequency of lies; as mentioned earlier, the difference in the percentage of untruthful messages between the two treatments is statistically significant. Selfish lying accounts for just over half of all instances of deception in the control treatment but only accounts for about a third in the punishment one. The third and last result is that a substantial proportion of subjects can be classified as strong truth-tellers. The frequency of this behavior differs across treatments ($p = 0.033$, $n = 335$), suggesting that the possibility of punishment induced senders to switch from selfish lying to costly truth-telling. The threat of punishment was indeed very real: The punishment rate after history {lie,trust} was substantial, 27.38%.[14]

4.3. (Dis)honesty

We next use regression analysis to study the association between lying and deception on the one hand and the three masculine physical traits we consider on the other. In **Table 5** below, we present the results of five random-effects regressions with robust standard errors clustered at the session level. These

[14]The rest of punishment rates were 9.52% for {truth,distrust}, 4.77% for {lie, distrust} and 2.97% for {truth,trust}.

models pool the data from the two treatments and include a dummy variable for the punishment one. The three masculine traits and beliefs about the trust rate among receivers (when used as a control) are standardized. Coefficients should then be interpreted as the change in the outcome variable produced by a one standard deviation change in the corresponding independent variable.

Beliefs. Column (1) studies the association between masculine physical traits and participants' beliefs about trust rates in their session. All three markers display a significant coefficient, although in varying degrees of significance and directions. An increase in fMM (fWHR) by a standard deviation increases the expected trust rate by 3.5 (2.3) percentage points (pp). However, higher exposure to prenatal androgens as measured by 2D:4D, decreases the expected trust rate by 3.2 pp.

Result 1: Higher levels of fMM are associated with senders expecting more receivers to follow their message. Higher exposure to prenatal testosterone and higher fWHR are associated with the opposite.

Note that the coefficients of interest in column (1) are relatively small and on those for fMM and fWHR are weakly significant. This already suggests that the association between beliefs and our masculine traits is not strong. We will come back to this issue below.

Lying. In the rest of columns of **Table 5**, the dependent variable is a dummy with value one if the subject lied. We chose a linear probability model (LPM) for these regressions because we are interested in the marginal effects of the masculine traits. These marginal effects are intuitively measured by the coefficients of a LPM: changes in a variable corresponds to a percentage points (pp) change in the probability of lying. LPMs, however, present two problems: heteroskedasticity in errors (by construction), and estimation bias, which has been shown to increase with the proportion of predicted probabilities outside the [0, 1] interval (Horrace and Oaxaca, 2006). We use clustered robust standard errors to avoid the first issue. On the other hand, only 0.5% of our predicted probabilities is negative and none is above one, suggesting that our LPM estimates are fairly unbiased as well. Nonetheless, we also ran random-effects probit models (see **Table A1** in the Appendix), which yielded similar results.

The specification in column (2) estimates the total effect of the three masculinity markers we study on lying. Again, the traits display sizeable effects at different degrees of significance. 2D:4D is associated to a decrease in the probability of lying by 6.1 pp. This regression thus suggests that individuals exposed to more testosterone *in utero* engaged less in lying. The estimates for the other traits are non-negligible but less significant.

Column (3) includes subjects' expected trust rates as a control. The coefficient for fWHR increases in significance and absolute value whereas the one for 2D:4D decreases. An increase of one standard deviation in fWHR now leads to a reduction in lying by 6.6 pp, and an increase in the latter to a decrease by 5.4 pp.

Result 2. Higher fWHR and higher exposure to prenatal testosterone are associated to less lying.

The positive coefficient for beliefs implies that participants who believed that a higher fraction of receivers would trust messages were more likely to lie.[15] The coefficient is highly significant and its size is substantial: a standard deviation increase in the expected trust rate translates into a 6.7 pp increase in the probability of lying.

TABLE 5 | Random-effects models.

	Belief	Lie	Lie	Lie (Trust<50%)	Lie (Trust>50%)
	(1)	(2)	(3)	(4)	(5)
fWHR	2.276*	–0.058*	–0.066*	–0.091**	–0.044
	(0.100)	(0.083)	(0.051)	(0.031)	(0.206)
fMM	3.497*	0.039*	0.032	0.088	0.038
	(0.079)	(0.064)	(0.107)	(0.228)	(0.254)
2D:4D	–3.169**	–0.061**	–0.054*	0.023	0.095**
	(0.027)	(0.023)	(0.051)	(0.296)	(0.049)
Punishment	–3.193	–0.125***	-0.112**	-0.068	-0.146*
	(0.395)	(0.008)	(0.017)	(0.223)	(0.062)
Belief			0.067***	–0.018	0.037
			(0.002)	(0.866)	(0.593)
Observations	333	334	333	103	191

All specifications control for the BMI and age of the subject. Robust standard errors. clustered at the session level. Variables are standardized. p-*values in parentheses.* *** *denotes* $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Preferences vs. beliefs. The regression in column (3) is also important because it allows us to explore the extent to which the association between our masculine traits and lying behavior operates through beliefs about the behavior of others, through preferences, or both (Eisenegger et al., 2012). Assuming that sender's behavior depends on preferences, i.e., lying aversion, and beliefs about receivers' behavior and that, in turn, both preferences and beliefs vary with masculine traits implies that column (2) estimate the total association of our biomarkers with dishonesty. When in column (3) we control for participants' beliefs about trust rates we would be estimating the indirect association of the trait via preferences as we would be switching off the beliefs channel. This implies that the differences in estimates between those in columns (2) and (3) allows us to measure the size of the effect of our masculine traits on lying operating through beliefs. The sizes of these effects are all very small, approximately 0.08 pp for a one standard deviation increase in fWHR and –0.07pp for a one standard deviation increase in fMM and 2D:4D. This would corroborate the following result:

Result 3: Masculine markers have a statistically significant but weak association with lying via beliefs about the behavior of receivers.

Let us mention that this identification strategy rests on two assumptions. The first one is that beliefs are measured without error. This is important because, as Gillen et al. (2019) have shown, measurement error in a control variable (beliefs in this case) that correlates both with the dependent variable (lying) and

[15]This is in line with Gneezy (2005) and Peeters et al. (2015) who found that the higher the expected costs of truth-telling, that is, the more trusting receivers are expected to be, the more likely are senders to lie.

other controls (masculine traits) alters estimates. The presence of a substantial measurement error in elicited beliefs would distort the estimates in column (3) and affect our inference of effect sizes. Secondly, we are assuming that the beliefs about trusting rates that individuals report do not depend on their decision as senders. However, it might be the case that the action participants take as senders influence the beliefs they report.

To partially ameliorate these concerns, we report in the Appendix the results of an instrumental variable approach where we substitute elicited beliefs by the residuals from the estimation in column (1) (see **Table A2**). These residuals are thus the expected trust rates left unexplained by the masculine physical traits we study. The estimates resulting from this exercise are analogs to those in column (3) and can thus be interpreted as the association between the masculine traits and lying via preferences. The coefficients for fMM and 2D:4D become more significant, reinforcing the idea the effect of these traits operate mostly through lying aversion. That said, this IV approach is not a panacea and this result should be taken as suggestive.

Deception. At this point, the distinction between lying and deception becomes important. A higher likelihood of sending a truthful message does not necessarily indicates stronger prosociality. If the receiver is expected to distrust the sender, telling the truth becomes a form of sophisticated deception (Sutter, 2009). Columns (4) and (5) account for the different ethical and monetary implications of lying depending on receivers' expected trust rates. These models restrict the analysis to subjects who believed that less (more, respectively) than 50% of receivers would trust their message. We leave out senders who believed that exactly 50% of receivers would follow messages as these senders would be indifferent between lying or not. Due to the reduction in observations, we lose some precision. Still, estimates show that the association between fWHR and lying observed in column (3) is only significant for senders who expected receivers to distrust messages, although coefficients in columns (4) and (5) are not significantly different from each other.[16] In addition, the coefficient for 2D:4D is only significant for senders whose expected trust rates were above 50%. In addition, it is statistically different from the coefficient in column (4); a *t*-test of the equality of the coefficients returns a *p*-value of 0.008.[17]

Result 4: Higher fWHR is associated with more deception by telling the truth, whereas higher exposure to prenatal testosterone is positively associated with strong truth-telling.

To address the possibility of false positives, we also run a series if bivariate regressions where each of the three masculine traits is regressed on the dependent variables. **Table A3** in the Appendix shows that almost all coefficients are of the same magnitude and significance as those in Table 5. The only difference is that the association between 2D:4D and beliefs is no longer significant in the corresponding bivariate regression. For the sake of transparency, we also include in the Appendix the analysis of receivers' beliefs, trusting decisions and punishment after history {lie, trust} (**Table A4** and **A5**). There, a significant positive association emerges between fWHR and trusting by receivers, especially when senders were expected to have lied in more than 50% of occasions.

[16]We ran a version of this model were the three masculine traits were interacted with a dummy taking the value 1 if the subject expected more than 50% of receivers to follow messages. This model returned a *p*-value of 0.145 for the t-test of the coefficient on the interaction of the expected trust dummy and the subject's fWHR being zero.

[17]This *p*-value is from the *t*-test of the coefficient on the interaction between the expected trust dummy and the subject's 2D:4D being equal to zero in the interacted version of the model.

5. DISCUSSION

Our results suggest that biology, and masculine physical traits in particular, can help explain the observed individual heterogeneity in honesty in strategic communication. This association seems to operate more strongly through preferences than through beliefs about the behavior of others. Two of the masculine traits markers we study (2D:4D and fWHR) are negatively correlated with lying, but the picture changes when we bring into consideration the expected consequences of messages. This is consistent with recent research on the role of testosterone in social interactions. This hormone makes the seek of social status and dominance a salient motivation, but this goal translates into aggression and competitiveness in some contexts and into prosocial behavior in others (Eisenegger et al., 2011, 2012; Millet, 2011; van Honk et al., 2011b). In our experiment, senders could obtain higher status either by outsmarting the receiver or by following the moral imperative of truth-telling, especially when that was costly.

We next elaborate on how our results relate with other results previously observed in the literature.

5.1. The Reduced Empathy Hypothesis

According to the dual-process theory, moral decisions trigger immediate emotional responses such as harm aversion and empathy (see Montoya et al., 2013, and references therein). When striving for status, an awareness of the emotions of others might be detrimental to oneself. In that case, instrumental considerations must override emotional responses in order to clear the path for payoff maximization. Research on behavioral endocrinology suggests that prenatal testosterone exposure is indeed positively associated with decreased empathy, even from a very early age (e.g., Knickmeyer et al., 2005), thus facilitating narrow utility maximization over emotional decision making. Recent evidence also shows that testosterone suppresses the activity of the ventromedial prefrontal cortex (vmPFC), a brain region implicated in moral decision making; it is known that individuals with vmPFC lesions are significantly less likely to experience regret, guilt or embarrassment after violating social norms (Carney and Mason, 2010).

Previous research on the association between masculine facial features and economic behavior is also consistent with this *reduced empathy hypothesis*. Stirrat and Perrett (2010) found that subjects with higher fWHR are more likely to exploit the trust of others in the trust game. Haselhuhn and Wong (2012) and Geniole et al. (2014) found that fWHR is positively related to cheating in non-strategic settings. And Jia et al. (2014) observed a positive relationship between the fWHR of CEOs and their probability of engaging in fraudulent accounting practices.

Our results related to masculine facial features as measured by fWHR are consistent with the reduced empathy hypothesis. Subjects with higher fWHR were more likely to engage in deception by telling the truth (Sutter, 2009), which entails hurting the receiver for profit.

5.2. The Status Signaling Hypothesis

Social status is likely to be associated with prosocial behavior when prosociality can signal dominance or higher standing (Eisenegger et al., 2012). For instance, Millet and Dewitte (2009) found that individuals exposed to higher amniotic testosterone concentrations are indeed more generous in the dictator game. Two mechanisms might drive this association. First, it seems that testosterone enhances self-image concerns, leading individuals to make choices which make them feel proud and to avoid those considered dishonorable (Wibral et al., 2012). Second, higher androgens levels suppress the immune system and are feasible only for healthy individuals. In that case, traits and behaviors associated with higher testosterone exposure become signals of superior health and genetic fitness (Puts et al., 2012).

Our results related to prenatal testosterone levels are consistent with this form of status signaling. We find a strong association between 2D:4D and truth-telling. This suggests that androgens exposure during foetal development is related to a pure preference for truth-telling, or alternatively, to a stronger lying aversion (Kartik, 2009; Millet, 2011).

In addition, we find a positive association between prenatal testosterone exposure and costly truth-telling, which is consistent with the idea that honesty in communication under divergent preferences can be seen as a signal of status. Sending a truthful message when most receivers are expected to believe it has a monetary cost compared to lying. Such behavior can in turn be rationalized as a costly signal of higher moral or resource standing (Gintis et al., 2001). This interpretation is also consistent with Weston et al. (2007) and van Honk et al. (2015), who find that testosterone administration reduces selfish misreporting and bluffing, and with the literature suggesting that amniotic testosterone concentrations are positively related with prosociality and altruism (e.g., van den Bergh and Dewitte, 2006; Millet and Dewitte, 2009).

5.3. The Dual Role of fWHR

The literature on the fWHR suggests that this trait is linked to antisocial behaviors such as untrustworthiness and cheating. But it remains unclear the extent to which this association is due to individuals adopting these behaviors because of their physical appearance. Wang et al. (2019) suggest that fWHR might have been associated with antisocial behaviors in ancestral environments. The more imposing appearance of men with wider faces may lead them to be less concerned with retaliations to their aggressions (Geniole et al., 2015). That might explain the more aggressive financial policies of CEOs with higher fWHR (Mills and Hogan, 2020). In our experiment, males with wider faces may have felt that their deception was less likely to meet a punishment, as it was the case in their daily life, leading them to engage more in deception.

A body of studies offer a more nuanced view on fWHR. Wong et al. (2011) show that firms in the Fortune 500 whose male CEOs have higher fWHR enjoy higher returns-to-assets ratios. This might be due to these individuals being more exploitative, but also to them being more cooperative. In fact, Stirrat and Perrett (2012) show that men with wider faces contribute more to public goods under inter-group competition. In addition, Lewis et al. (2012) find that US presidents with higher fWHR had a higher drive for achievement but were not more aggressive in their policies. This suggests that the link between fWHR and economic behavior might be contingent on the context: Aggression or cheating might be a bad strategy for presidents but it is perhaps useful and/or socially forgiven in business and finance. Alternatively, it might be that fWHR is associated with prosociality when the own group (e.g., a country, a firm) is competing against another. Inter-group competition was absent in our design; that might have shut down any possible association between this trait and honesty.

6. CONCLUSION

In this article, we have offered an exploratory study of the biological roots of lying and deception in strategic communication. Our results suggest that individual differences in honesty in sender-receiver games can be partially explained by physiological factors. We have also explored whether these association operate more strongly through preferences or through beliefs about the behavior of receivers. Exposure to sexual hormones during foetal development increases truth-telling whereas fWHR, which is related to aggression and dominance in a variety of contexts, predicts more lying and more sophisticated deception. We observed these associations in an environment where the preferences of senders and receivers were completely opposed. Future research should explore whether masculine physical traits or other biological markers may a have different relationship with honesty under other preference configurations and contexts.

In addition, further studies on the relationship between biomarkers and dishonesty should include female participants. A next step could be to explore whether hormone levels may relate differently with lying and deception among females. It would also be relevant to study whether the associations between dishonesty and physiology-related traits are mediated by the gender identity of the sender or the receiver once disclosed to the other party.

Finally, it is important to note that our study cannot tease out the direct biological effect of visible masculine traits from their effect on how individuals are perceived and treated by their peers. Any trait an individual displays mixes biological influences (i.e., genes, which respond to the environment through endocrine and nervous system signalling, whose organization also depends on genes and the environment) and the events the individual experiences (including abiotic factors and their interaction with other living beings). From that point of view, the conjecture that prenatal hormone levels may influence human behavior is especially interesting (Beltz et al., 2011; Berenbaum, 2018). There exist some differences in preferential activities associated with differences in prenatal hormone levels, such as the interest in hitting rather than swinging objects or differences in the attention devoted to objects and people. These differences interact with the social environment to produce

behavioral differences in adulthood. In other words, biological and social processes interact with each other and jointly affect development. On the other hand, hormone levels in adolescence influence facial features which in turn may influence how people are perceived and treated, leading to further differences in behavior. In sum, it is extremely difficult to disentangle the direct biological effect of masculine physical traits unless individuals are continuously monitored. This is another open avenue for future research.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Autonomous University of Madrid Research Ethics Committee (reference CEI 62-1086). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

MV acknowledges funding from the Spanish Ministry of Science and Technology research grant ECO2015-65701-P and SS-P and ET from grant ECO2015-66281-P.

ACKNOWLEDGMENTS

We are grateful to the editor AK and two referees for their comments; to Raul Lopez-Perez and the Madrid Laboratory for Experimental Economics (MADLEE) for their help in the organization of the experiment; and to Marta Iglesias-Julios and Jose Antonio Muñoz-Reyes for their excellent research assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.684226/full#supplementary-material

REFERENCES

Abeler, J, Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica* 87, 1115–1153. doi: 10.3982/ECTA14673

Beltz, A. M., Swanson, J. L., and Berenbaum, S. A. (2011). Gendered occupational interests: Prenatal androgen effects on psychological orientation to things versus people. *Hormones Behav*. 60, 313–317. doi: 10.1016/j.yhbeh.2011.06.002

Berenbaum, S. A. (2018). Beyond pink and blue: the complexity of early androgen effects on gender development. *Child Dev. Perspect*. 12, 58–64. doi: 10.1111/cdep.12261

Berenbaum, S. A., and Beltz, A. M. (2011). Sexual differentiation of human behavior: effects of prenatal and pubertal organizational hormones. *Front. Neuroendocrinol*. 32, 183–200. doi: 10.1016/j.yfrne.2011.03.001

Buser, T. (2012a). The impact of the menstrual cycle and hormonal contraceptives on competitiveness. *J. Econ. Behav. Organ*. 83, 1–10. doi: 10.1016/j.jebo.2011.06.006

Buser, T. (2012b). Digit ratios, the menstrual cycle and social preferences. *Games Econ. Behav*. 76, 457–470. doi: 10.1016/j.geb.2012.07.006

Buskens, V., Raub, W., Van Miltenburg, N., Montoya, E. R., and Van Honk, J. (2016). Testosterone administration moderates effect of social environment on trust in women depending on second-to-fourth digit ratio. *Sci. Rep*. 6, 27655. doi: 10.1038/srep27655

Butovskaya, M., Burkova, V., Apalkova, Y., Dronova, D., Rostovtseva, V., Karelin, D., et al. (2021). Sex, population origin, age and average digit length as predictors of digit ratio in three large world populations. *Sci. Rep*. 11, 1–17. doi: 10.1038/s41598-021-87394-6

Cai, H., and Wang, J. (2006). Overcommunication in strategic information transmission games. *Games Econ. Behav*. 95, 384–394. doi: 10.1016/j.geb.2005.04.001

Capraro, V. (2018). Gender differences in lying in sender-receiver games: a meta-analysis. *Judg. Decis. Making* 13, 345–355. doi: 10.31234/osf.io/jaewt

Carney, D. R., and Mason, M. F. (2010). Decision making and testosterone: when the ends justify the means. *J. Exper. Soc. Psychol*. 46, 668–671. doi: 10.1016/j.jesp.2010.02.003

Carré, J. M., and McCormick, C. M. (2008). In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proc. R. Soc. B Biol. Sci*. 275, 2651–2656. doi: 10.1098/rspb.2008.0873

Cecchi, F., and Duchoslav, J. (2018). The effect of prenatal stress on cooperation: evidence from violent conflict in Uganda. *Eur. Econ. Rev*. 101, 35–56. doi: 10.1016/j.euroecorev.2017.09.015

Chen, C., Decety, J., Huang, P. C., Chen, C. Y., and Cheng, Y. (2016). Testosterone administration in females modulates moral judgment and patterns of brain activation and functional connectivity. *Hum. Brain Mapp*. 37, 3417–3430. doi: 10.1002/hbm.23249

Chen, Y., Katuščák, P., and Ozdenoren, E. (2013). Why can't a woman bid more like a man? *Games Econ. Behav*. 77, 181–213. doi: 10.1016/j.geb.2012.10.002

Crawford, V., and Sobel, J. (1982). Strategic information transmission. *Econometrica* 50, 1431–1451. doi: 10.2307/1913390

Dessein, W. (2002). Authority and communication in organizations. *Rev. Econ. Stud*. 69, 811–838. doi: 10.1111/1467-937X.00227

Eisenegger, C., Haushofer, J., and Fehr, E. (2011). The role of testosterone in social interactions. *Trends Cogn. Sci*. 15, 263–271. doi: 10.1016/j.tics.2011.04.008

Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., and Fehr, E. (2012). New evidence on testosterone and cooperation. A reply. *Nature* 485: E5–E6. doi: 10.1038/nature11137

Ekrami, O., Claes, P., Van Assche, E., Shriver, M. D., Weinberg, S. M., Marazita, M. L., et al. (2021). Fluctuating asymmetry and sexual dimorphism in human facial morphology: a multi-variate study. *Symmetry* 13, 304. doi: 10.3390/sym13020304

Erat, S., and Gneezy, U. (2012). White lies. *Manag. Sci*. 58, 723–733. doi: 10.1287/mnsc.1110.1449

Galis, F., Ten Broek, C. M., Van Dongen, S., and Wijnaendts, L. C. (2009). Sexual dimorphism in the prenatal digit ratio (2D:4D). *Arch. Sex. Behav*. 39, 57–62. doi: 10.1007/s10508-009-9485-7

Garbarino, E., Slonim, R., and Sydnor, J. (2011). Digit ratios (2d:4d) as predictors of risky decision making for both sexes. *J. Risk Uncertain*. 42, 1–26. doi: 10.1007/s11166-010-9109-6

Gawn, G., and Innes, R. (2018). Do lies erode trust? *Int. Econ. Rev*. 59, 137–161. doi: 10.1111/iere.12265

Geniole, S. N., Denson, T. F., Dixson, B. J., Carré, J. M., and McCormick, C. M. (2015). Evidence from meta-analyses of the facial width-to-height ratio as an evolved cue of threat. *PLoS ONE* 10:e0132726. doi: 10.1371/journal.pone.0132726

Geniole, S. N., Keyes, A. E., Carré, J. M., and McCormick, C. M. (2014). Fearless dominance mediates the relationship between the facial width-to-height ratio and willingness to cheat. *Pers. Individ. Dif.* 57, 59–64 doi: 10.1016/j.paid.2013.09.023

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the Caltech cohort study. *J. Polit. Econ.* 127, 1826–1863. doi: 10.1086/701681

Gintis, H., Smith, E. A., and Bowles, S. (2001). Costly signaling and cooperation. *J. Theor. Biol.* 213, 103–119. doi: 10.1006/jtbi.2001.2406

Gneezy, U. (2005). Deception: the role of consequences. *Am. Econ. Rev.* 95, 384–394. doi: 10.1257/0002828053828662

Grimbos, T., Dawood, K., Burriss, R. P., Zucker, K. J., and Puts, D. A. (2010). Sexual orientation and the second to fourth finger length ratio: a meta-analysis in men and women. *Behav. Neurosci.* 124, 278–287. doi: 10.1037/a0018764

Haselhuhn, M. P., Ormiston, M. E., and Wong, E. M. (2015). Men's facial width-to-height ratio predicts aggression: a meta-analysis. *PLoS ONE* 10:e0122637. doi: 10.1371/journal.pone.0122637

Haselhuhn, M. P., and Wong, E. M. (2012). Bad to the bone: facial structure predicts unethical behaviour. *Proc. R. Soc. B Biol. Sci.* 279, 571–576. doi: 10.1098/rspb.2011.1193

Haselhuhn, M. P., Wong, E. M., Ormiston, M. E., Inesi, M. E., and Galinsky, A. D. (2014). Negotiating face-to-face: Men's facial structure predicts negotiation performance. *Leader. Q.* 25, 835–845. doi: 10.1016/j.leaqua.2013.12.003

Hickey, M., Doherty, D. A., Hart, R., Norman, R. J., Mattes, E., Atkinson, H. C., et al. (2010). Maternal and umbilical cord androgen concentrations do not predict digit ratio (2D: 4D) in girls: a prospective cohort study. *Psychoneuroendocrinology* 35, 1235–1244. doi: 10.1016/j.psyneuen.2010.02.013

Hönekopp, J., Bartholdt, L., Beier, L., and Liebert, A. (2007). Second to fourth digit length ratio (2D:4D) and adult sex hormone levels: new data and a meta-analytic review. *Psychoneuroendocrinology* 32, 313–321. doi: 10.1016/j.psyneuen.2007.01.007

Honekopp, J., and Watson, S. (2010). Meta-analysis of digit ratio 2D:4D shows greater sex difference in the right hand. *Am. J. Hum. Biol.* 22, 619–630. doi: 10.1002/ajhb.21054

Honekopp, J., and Watson, S. (2011). Meta-analysis of the relationship between digit ratio 2D:4D and aggression. *Pers. Individ. Dif.* 51, 381–386. doi: 10.1016/j.paid.2010.05.003

Horrace, W. C., and Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Econ. Lett.* 90, 321–327. doi: 10.1016/j.econlet.2005.08.024

Hurkens, S., and Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. *Exper. Econ.* 12, 180–192. doi: 10.1007/s10683-008-9208-2

Jia, Y., van Lent, L., and Zeng, Y (2014). Masculinity, testosterone, and financial misreporting. *J. Account. Res.* 52, 1195–1246. doi: 10.1111/1475-679X.12065

Kartik, N. (2009). Strategic communication with lying costs. *Rev. Econ. Stud.* 76, 1359–1395. doi: 10.1111/j.1467-937X.2009.00559.x

Knickmeyer, R., Baron-Cohen, S., Raggatt, P., and Taylor, K. (2005). Foetal testosterone, social relationships, and restricted interests in children. *J. Child Psychol. Psychiatry* 46, 198–210. doi: 10.1111/j.1469-7610.2004.00349.x

Kramer, R. S. (2017). Sexual dimorphism of facial width-to-height ratio in human skulls and faces: a meta-analytical approach. *Evol. Hum. Behav.* 38, 414–420. doi: 10.1016/j.evolhumbehav.2016.12.002

Kratochvíl, L., and Flegr, J. (2009). Differences in the 2nd to 4th digit length ratio in humans reflect shifts along the common allometric line. *Biol. Lett.* 5, 643–646. doi: 10.1098/rsbl.2009.0346

Lefevre, C. E., Wilson, V. A.D, M.orton, F. B., Brosnan, S. F., Paukner, A., and Bates, T. C. (2014). Facial width-to-height ratio relates to alpha status and assertive personality in capuchin monkeys. *PLoS ONE* 9:e93369. doi: 10.1371/journal.pone.0093369

Lewis, G. J., Lefevre, C. E., and Bates, T. C. (2012). Facial width-to-height ratio predicts achievement drive in US presidents. *Pers. Individ. Dif.* 52, 855–857. doi: 10.1016/j.paid.2011.12.030

Lippa, R. A. (2016). "Biological influences on masculinity," in *APA Handbook of Men and Masculinities*, eds Y. J. Wong and S. R. Wester (Washington, DC: APA Books).

Lolli, L., Batterham, A. M., Kratochvíl, L., Flegr, J., Weston, K. L., and Atkinson, G. (2017). A comprehensive allometric analysis of 2nd digit length to 4th digit length in humans. *Proc. R. Soc. B Biol. Sci.* 284, 20170356. doi: 10.1098/rspb.2017.0356

López-Pérez, R., and Spiegelman, E. (2013). Why do people tell the truth? Experimental evidence for pure lie aversion. *Exper. Econ.* 16, 233–247. doi: 10.1007/s10683-012-9324-x

Manning, J. T., Kilduff, L., Cook, C., Crewther, B., and Fink, B. (2014). Digit ratio (2D: 4D): a biomarker for prenatal sex steroids and adult sex steroids in challenge situations. *Front. Endocrinol.* 5:9. doi: 10.3389/fendo.2014.00009

Manning, J. T., Scutt, D., Wilson, J., and Lewis-Jones, DI (1998). The ratio of 2nd to 4th digit length: a predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. *Hum. Reproduct.* 13, 3000–3004. doi: 10.1093/humrep/13.11.3000

Marečková, K. W., einbrand, Z., Chakravarty, M. M., Lawrence, C., Aleong, R., Leonard, G., et al. (2011). Testosterone-mediated sex differences in the face shape during adolescence: subjective impressions and objective features. *Hormones Behav.* 60, 681–690. doi: 10.1016/j.yhbeh.2011.09.004

Matsumoto, D., and Hwang, H. C. (2021). Facial width-to-height ratios and deception skill. *Pers. Individ. Dif.* 174, 110683. doi: 10.1016/j.paid.2021.110683

McIntyre, M. H., Ellison, P. T., Lieberman, D. E., Demerath, E., and Towne, B. (2005). The development of sex differences in digital formula from infancy in the fels longitudinal study. *Proc. R. Soc. B Biol. Sci.* 272, 1473–1479. doi: 10.1098/rspb.2005.3100

McKelvey, R., and Palfrey, T. (1995). Quantal response equilibria for extensive form games. *Exper. Econ.* 1, 9–41. doi: 10.1006/game.1995.1023

Melumad, N. D., and Shibano, T. (1994). The securities and exchange commission and the financial accounting standards board: regulation through veto-based delegation. *J. Account. Res.* 32, 1–37. doi: 10.2307/2491385

Millet, K. (2011). An interactionist perspective on the relation between 2D:4D and behaviour: an overview of (moderated) relationships between 2D:4D and economic decision making. *Pers. Individ. Dif.* 51, 397–401. doi: 10.1016/j.paid.2010.04.005

Millet, K., and Buehler, F. (2018). A context dependent interpretation of inconsistencies in 2D: 4D findings: the moderating role of status relevance. *Front. Behav. Neurosci.* 11:254. doi: 10.3389/fnbeh.2017.00254

Millet, K., and Dewitte, S. (2006). Second to fourth digit ratio and cooperative behavior. *Biol. Psychol.* 71, 111–115. doi: 10.1016/j.biopsycho.2005.06.001

Millet, K., and Dewitte, S. (2009). The presence of aggression cues inverts the relation between digit ratio (2D:4D) and prosocial behaviour in a dictator game. *Br. J. Psychol.* 100, 151–162. doi: 10.1348/000712608X324359

Mills, J., and Hogan, K. M. (2020). CEO facial masculinity and firm financial outcomes. *Corporate Board* 16, 39–46. doi: 10.22495/cbv16i1art4

Minozzi, W., and Woon, J. (2020). Direct response and the strategy method in an experimental cheap talk game. *J. Behav. Exper. Econ.* 85, 101498. doi: 10.1016/j.socec.2019.101498

Montoya, E. A., Terburg, D., Bos, P. A., Will, G.-J, B.uskens, V., Raub, W., and van Honk, J. (2013). Testosterone administration modulates moral judgments depending on second-to-fourth digit ratio. *Psychoneuroendocrinology* 38, 1362–1369. doi: 10.1016/j.psyneuen.2012.12.001

Neyse, L., Johannesson, M., and Dreber, A. (2021). 2D:4D does not predict economic preferences: Evidence from a large, representative sample. *J. Econ. Behav. Organ.* 185, 390–401. doi: 10.1016/j.jebo.2021.02.029

Pearson, M., and Schipper, B. C. (2012). The visible hand: finger ratio (2D:4D) and competitive bidding. *Exper. Econ.* 15, 510–529. doi: 10.1007/s10683-011-9311-7

Pearson, M., and Schipper, B. C. (2013). Menstrual cycle and competitive bidding. *Games Econ. Behav.* 78, 1–20. doi: 10.1016/j.geb.2012.10.008

Peeters, R., Vorsatz, M., and Walzl, M. (2015). Beliefs and truth-telling: a laboratory experiment. *J. Econ. Behav. Organ.* 113, 1–12. doi: 10.1016/j.jebo.2015.02.009

Puts, D. A., Jones, B. C., and DeBruine, L. M. (2012). Sexual selection on human faces and voices. *J. Sex Res.* 49, 227–243. doi: 10.1080/00224499.2012.658924

Ranehill, E., Zethraeus, N., Blomberg, L., von Schoultz, B., Hirschberg, A. L., Johannesson, M., et al. (2018). Hormonal contraceptives do not impact

economic preferences: evidence from a randomized trial. *Manag. Sci.* 64, 4471–4965. doi: 10.1287/mnsc.2017.2844

Richards, G., Browne, W. V., Aydin, E., Constantinescu, M., Nave, G., Kim, M. S., et al. (2020). Digit ratio (2D: 4D) and congenital adrenal hyperplasia (CAH): systematic literature review and meta-analysis. *Hormones Behav.* 126, 104867. doi: 10.1016/j.yhbeh.2020.104867

Rodriguez-Ruiz, C., Munoz-Reyes, J. A., Iglesias-Julios, M., Sanchez-Pages, S., and Turiegano, E. (2019). Sex affects the relationship between third party punishment and cooperation. *Sci. Rep.* 9, 1–9. doi: 10.1038/s41598-019-40909-8

Rohlf, J. H. (2015). The tps series of software. *Hystrix* 26, 9–12. doi: 10.4404/hystrix-26.1-11264

Ryckmans, J., Millet, K., and Warlop, L. (2015). The influence of facial characteristics on the relation between male 2D:4D and dominance. *PLoS ONE* 10:e0143307. doi: 10.1371/journal.pone.0143307

Sadr, M., Khorashad, B. S., Talaei, A., Fazeli, N., and Hönekopp, J. (2020). 2D:4D suggests a role of prenatal testosterone in gender dysphoria. *Arch. Sex. Behav.* 49, 421–432. doi: 10.1007/s10508-020-01630-0

Sanchez-Pages, S., Rodriguez-Ruiz, C., and Turiegano, E. (2014). Facial masculinity: how the choice of measurement method enables to detect its influence on behaviour. *PLoS ONE* 9:e112157. doi: 10.1371/journal.pone.0112157

Sanchez-Pages, S., and Turiegano, E. (2010). Testosterone, facial symmetry and cooperation in the prisoners' dilemma. *Physiol. Behav.* 99, 355–361. doi: 10.1016/j.physbeh.2009.11.013

Sanchez-Pages, S., and Turiegano, E. (2013). Two studies on the interplay between social preferences and biological characteristics. *Behaviour* 150, 713–735. doi: 10.1163/1568539X-00003077

Sanchez-Pages, S., and Vorsatz, M. (2007). An experimental study of truth-telling in a sender–receiver game. *Games Econ. Behav.* 61, 86–112. doi: 10.1016/j.geb.2006.10.014

Schipper, B. (2015). Sex hormones and competitive bidding. *Manag. Sci.* 61, 249–266. doi: 10.1287/mnsc.2014.1959

Schulz, K. M., Molenda-Figueira, H. A., and Sisk, C. L (2009). Back to the future: the organizational-activational hypothesis adapted to puberty and adolescence. *Hormones Behav.* 55, 597–604. doi: 10.1016/j.yhbeh.2009.03.010

Sisk, C. L., and Zehr, J. L. (2005). Pubertal hormones organize the adolescent brain and behavior. *Front. Neuroendocrinol.* 26, 163–174. doi: 10.1016/j.yfrne.2005.10.003

Sobel, J. (2020). Lying and deception in games. *J. Polit. Econ.* 128, 907–947. doi: 10.1086/704754

Stein, J. C. (1989). Cheap talk and the Fed: a theory of imprecise policy announcements. *Am. Econ. Rev.* 79, 32–42.

Stirrat, M., and Perrett, D. (2010). Valid facial cues to cooperation and trust: male facial width and trustworthiness. *Psychol. Sci.* 21, 349–354. doi: 10.1177/0956797610362647

Stirrat, M., and Perrett, D. (2012). Face structure predicts cooperation. *Psychol. Sci.* 23, 718–722. doi: 10.1177/0956797611435133

Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *Econ. J.* 119, 47–60. doi: 10.1111/j.1468-0297.2008.02205.x

Swift-Gallant, A., Johnson, B. A., Di Rita, V., and Breedlove, SM (2020). Through a glass, darkly: human digit ratios reflect prenatal androgens. *Hormones Behav.* 120, 104686. doi: 10.1016/j.yhbeh.2020.104686

Trivers, R., Manning, J., and Jacobson, A. (2006). A longitudinal study of digit ratio (2D:4D) and other finger ratios in Jamaican children. *Hormones Behav.* 49, 150 – 156. doi: 10.1016/j.yhbeh.2005.05.023

Turanovic, J. J., Pratt, T. C., and Piquero, A. R. (2017). Exposure to fetal testosterone, aggression, and violent behavior: a meta-analysis of the 2D: 4D digit ratio. *Aggression Violent Behav.* 33, 51–61. doi: 10.1016/j.avb.2017.01.008

van den Bergh, B., and Dewitte, S. (2006). Digit ratio (2D:4D) moderates the impact of sexual cues on men's decisions in ultimatum games. *Proc. R. Soc. B Biol. Sci.* 273, 2091–2095. doi: 10.1098/rspb.2006.3550

van Dongen, S. (2014). Associations among facial masculinity, physical strength, fluctuating asymmetry and attractiveness in young men and women. *Ann. Hum. Biol.* 41, 205–213. doi: 10.3109/03014460.2013.847120

van Honk, J., Montoya, E. R., Bos, P. A., van Vogt, M., and Terburg, D. (2012). New evidence on testosterone and cooperation. *Nature* 485, E4–E5. doi: 10.1038/nature11136

van Honk, J., Schutter, D. J., Bos, P. A., Kruijt, A. W., Lentjes, E. G., and Baron-Cohen, S. (2011a). Testosterone administration impairs cognitive empathy in women depending on second-to-fourth digit ratio. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3448–3452 doi: 10.1073/pnas.1011891108

van Honk, J., Terburg, D., and Bos, P. A. (2011b). Further notes on testosterone as a social hormone. *Trends Cogn. Sci.* 15, 291–292. doi: 10.1016/j.tics.2011.05.003

van Honk, J., Will, G. J., Terburg, D., Raub, W., Eisenegger, C., and Buskens, V. (2015). Effects of testosterone administration on strategic gambling in poker play. *Sci. Rep.* 6, 80–96. doi: 10.1038/srep18096

van Leeuwen, B., Smeets, P., Bovet, J., Nave, G., Stieglitz, J., and Whitehouse, A. (2020). Do sex hormones at birth predict later-life economic preferences? Evidence from a pregnancy birth cohort study. *Proc. R. Soc. B Biol. Sci.* 287, 20201756. doi: 10.1098/rspb.2020.1756

Volz, K. G., Vogeley, K., Tittgemeyer, M., von Cramon, D. Y., and Sutter, M. (2015). The neural basis of deception in strategic interactions. *Front. Behav. Neurosci.* 9:27. doi: 10.3389/fnbeh.2015.00027

Wang, D., Nair, K., Kouchaki, M., Zajac, E. J., and Zhao, X. (2019). A case of evolutionary mismatch? Why facial width-to-height ratio may not predict behavioral tendencies. *Psychol. Sci.* 30, 1074–1081. doi: 10.1177/0956797619849928

Wang, J. T.-Y., Spezio, M., and Camerer, C. F. (2010). Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *Am. Econ. Rev.* 100, 984–1007.doi: 10.1257/aer.100.3.984

Weston, E. M., Friday, A. E., and Lio, P. (2007). Biometric evidence that sexual selection has shaped the hominin face. *PLoS ONE* 2:e710. doi: 10.1371/journal.pone.0000710

Wibral, M., Dohmen, T., Klingmüller, D., Weber, B., and Falk, A. (2012). Testosterone administration reduces lying in men. *PLoS ONE* 7:e46774. doi: 10.1371/journal.pone.0046774

Wong, E. M., Ormiston, M. E., and Haselhuhn, M. P. (2011). A face only an investor could love: CEO facial structure predicts firm financial performance. *Psychol. Sci.* 22, 1478–1483. doi: 10.1177/0956797611418838

Wozniak, D., and Harbaugh, H. T. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *J. Labor Econ.* 32, 161–198. doi: 10.1086/673324

Zethraeus, N., Kocoska-Maras, L., Ellingsen, T., von Schoultz, B., Hirschberg, A. L., and Johannesson, M. (2009). A randomized trial of the effect of estrogen and testosterone on economic behavior. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6535–6538. doi: 10.1073/pnas.0812757106

Zheng, Z., and Cohn, M. J. (2011). Developmental basis of sexually dimorphic digit ratios. *Proc. Natl. Acad. Sci. U.S.A.* 108, 16289–16294. doi: 10.1073/pnas.1108312108

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 August 2021
doi: 10.3389/fpsyg.2021.684078

Actions and the Self: I Give, Therefore I am?

Tobias Regner[1]* *and Astrid Matthey*[2,3]

[1] Department of Economics, University of Jena, Jena, Germany, [2] Economic and Social Environmental Issues, Max Planck Institute of Economics, Jena, Germany, [3] Umweltbundesamt/German Environment Agency, Dessau-Roßlau, Germany

Self-signaling models predict less selfish behavior in a probabilistic giving setting as individuals are expected to invest in a pro-social identity. However, there is also substantial evidence that people tend to exploit situational excuses for selfish choices (for instance, uncertainty) and behave more selfishly. We contrast these two motivations (identity management and self-deception) experimentally in order to test which one is more prevalent in a reciprocal giving setting. Trustees' back transfer choices are elicited for five different transfer levels of the trustor. Moreover, we ask trustees to provide their back transfer schedule for different scenarios that vary the implementation probability of the back transfer. This design allows us to identify subjects who reciprocate and analyze how these reciprocators respond when self-image relevant factors are varied. Our results indicate that self-deception is prevalent when subjects make the back transfer choice. Twice as many subjects seem to exploit situational excuses than subjects who appear to invest in a pro-social identity.

JEL classifications: C72, C91, D80, D91

Keywords: social preferences, pro-social behavior, moral wiggle room, self-image concerns, self-signaling, reciprocity, experiments

OPEN ACCESS

Edited by:
Steffen Huck,
Social Science Research Center Berlin, Germany

Reviewed by:
Menusch Khadjavi,
Vrije Universiteit Amsterdam, Netherlands
Sébastien Massoni,
Queensland University of Technology, Australia

***Correspondence:**
Tobias Regner
tobias.regner@uni-jena.de

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *22 March 2021*
Accepted: *14 June 2021*
Published: *10 August 2021*

Citation:
Regner T and Matthey A (2021) Actions and the Self: I Give, Therefore I am? Front. Psychol. 12:684078. doi: 10.3389/fpsyg.2021.684078

1. INTRODUCTION

Many people behave pro-socially—if the only other choice is selfish behavior. But what if the situation is less transparent? What if circumstances exist that allow a selfish choice while simultaneously a pro-social image in front of one self can be kept? Dana et al. (2007) find that giving rates are significantly reduced when moral excuses for selfish behavior are available. People seem to make use of 'moral wiggle room', a term coined by them, and evidence from a series of studies (e.g., Larson and Capra, 2009; Haisley and Weber, 2010; Hamman et al., 2010; Matthey and Regner, 2011; Feiler, 2014; Grossman, 2014; van der Weele, 2014; Exley, 2015) confirms such a self-serving bias in dictator game giving.

Bayesian self-signaling models propose a different type of behavior in an allocation decision. Individuals derive utility the more they believe they are a pro-social type. However, they are inherently unsure whether they actually are a pro-social type (or instead pro-self). Thus, they may give in order to send a positive signal to their self. Grossman (2015) develops such a model and tests it experimentally in a binary probabilistic dictator game. The model predicts more giving in the low

probability treatment as the expected cost of sending a pro-social signal is cheapened. However, results do not lend supporting evidence[1].

Bénabou and Tirole (2011) consider further channels how actions and the self relate to each other. Besides identity management (beliefs about one's type are malleable through actions) they also allow, for instance, for the possibility that an action may be regarded as uninformative about the self. Hence, individuals may attribute their selfish action to the context, instead of having to connect selfish behavior to their self-image. The salience of the context—to what extent one's action has an effect on the outcome—systematically varies the informativeness of an action. As a consequence, individuals would have a higher tendency to invest in their identity, when informativeness is high. Likewise, they would tend to succumb to the temptation of situational excuses, when informativeness is low. Naturally, it is in the self's 'eye of the beholder', whether an action is perceived to convey information about one's underlying character (potentially leading to an identity investment) or whether situational excuses are invoked and used to bias the signal to the self (making the identity damage of a selfish choice acceptably small). Thus, it is an empirical question which motivation is more dominant. Our study's goal is to consider both self-signaling channels (identity management and self-deception), test them experimentally, and thus shed more light on their relative prevalence.

We set up a probabilistic giving environment in which the predictions of the two approaches contrast each other. Identity management predicts higher transfers with increased uncertainty about the implementation of the transfer due to the pro-social signal becoming cheaper. Instead, self-deception predicts lower transfers, because uncertainty about the actual implementation could serve as a situational excuse.

Since the effect of moral wiggle room on giving in dictator games is well established, we decided to move our test bed to a less explored domain, namely, when also reciprocal concerns may motivate individuals[2]. For this purpose we conduct a modified trust game. Trustees' back transfer choices are elicited for five different transfer levels of the trustor. Moreover, we ask trustees to provide their back transfer schedule for different scenarios. While in scenario 1 the back transfer will be implemented for sure, in scenarios 2–4 there is a positive probability that the back transfer fails. In such a case the trustee gets to keep the available amount. After trustees have chosen their back transfer schedules for all scenarios, they are informed that they can select the scenario they would like to get implemented.

This design allows us to identify subjects who reciprocate (based on the back transfer schedule in scenario 1) and analyze how these reciprocators behave. Two situational excuses for selfish behavior are present in our design. First, the fact that in scenarios 2–4 the transfer could fail may serve as an excuse to return less in these scenarios (alternatively, the decreased implementation chance of the transfer could induce subjects to return more via identity management). Second, having to choose a scenario can imply the temptation of picking a favorable scenario—one that results in a monetary gain (in expectations)—while the trustor might not receive anything.

Our within-subjects design allows us to analyze back transfer choices at the individual level. Thus, we can distinguish between trustees motivated by self-deception and identity management. While our results show that behavior consistent with self-deception is more common when subjects make the back transfer choice (twice as many subjects decrease than increase their transfers under uncertainty), they also indicate that both motivational processes appear relevant for human decision making. Furthermore, as a substantial fraction of subjects makes a self-serving scenario choice, our results indicate that reciprocators make use of moral wiggle room if situational excuses exist.

The paper is organized as follows. In section 2 we describe the experiment and present behavioral predictions. Results are reported and discussed in section 3. We conclude in section 4.

2. EXPERIMENT

2.1. Design

The experiment consisted of a variant of the trust game (Berg et al., 1995). Both trustor and trustee received an endowment of 10 Euro. As the first step, the trustor could send either 0, 2.50, 5, 7.50, or 10 Euro to the trustee. This transfer was tripled and added to the trustee's account, who could then return any amount available on the account to the trustor. That is, depending on the trustor's transfer trustees could return up to 10, 17.50, 25, 32.50, or 40 Euro. All subjects played in both roles. They knew that it was determined randomly at the end of the experiment whether a subject acted as trustor or trustee. Trustees' decisions were elicited using the strategy method, that is, a trustee decided how much to send back to the trustor for all possible transfers. Hence, all trustees made five back transfer decisions, one of which was to become relevant according to the trustor's actual transfer. When entering their back transfer choices, trustees were informed about the respective amount they would receive at each transfer level. Trustors only learned the outcome, not the choice of the trustee.

Trustees knew that they make the back transfer choices for different scenarios. In scenario 1, the trustee's transfer was carried out with certainty, that is, it reached the trustor for sure and was subtracted from the trustee's account. In scenario 2, the transfer was carried out with 90% probability. With the remaining 10% probability, the trustee would keep the available amount. In this

[1]Also van der Weele and von Siemens (2020) test Bayesian self-signaling experimentally but find no empirical support. Grossman and van der Weele (2017) develop a Bayesian self-signaling model in the context of information acquisition where individuals can willfully ignore to get informed. Their related experimental test supports the model.

[2]Results of existing studies which analyze reciprocal behavior when there is a situational excuse for not giving point in different directions. van der Weele et al. (2014) apply the 'plausible deniability' treatment from Dana et al. (2007) to second-mover behavior in a trust/moonlighting game. They find no behavioral differences in comparison to a baseline and conclude that moral wiggle room has no effect on the incidence of reciprocal behavior. Regner (2018) find a significantly higher rate in three treatments that feature moral wiggle room manipulations (between 37.5 and 45%) compared to the baseline rate of selfish choices (6.25%). Malmendier et al. (2014) analyze an "exit option"—that is, to what extent subjects are willing to avoid an allocation choice even if it is costly—in the context of reciprocity. They find that subjects do sort out in the context of positive reciprocity but sorting out is significantly higher without reciprocity. Thus, our study provides an additional data point.

case, the trustor would be left with her endowment minus the amount she sent to the trustee, independently of the size of the trustee's back transfer. In scenario 3, the trustee's transfer was carried out with 80% probability, with 20% the trustee kept the entire amount. In half of our sessions we added a fourth scenario in which the trustee's transfer was or was not carried out with equal probability. Scenario 4 was employed to test whether the availability of an option with a much smaller transfer probability would serve as an excuse to choose a scenario with a transfer probability below 1 (but above 50%) rather than the certain transfer.

Overall, subjects therefore made five back transfer decisions (for each possible amount sent by the trustor) per scenario. After trustees completed all choices for one scenario, they were asked for their back transfers in the next scenario. Choices from previous scenarios were still visible. **Supplementary File 1** shows a screenshot of the decision interface for the scenario 1 choice and (**Supplementary File 1**) one for the scenario 3 choice when a subject has already entered back transfers for scenarios 1 and 2. It illustrates the sequential nature of entering the back transfer schedules for the scenarios and the fact that subjects were reminded of their choices in previous scenarios. We chose to provide choices in all previous scenarios in case a subject would like to take the same decision across scenarios. Since not just one decision but an entire back transfer schedule consisting of five choices would have to be remembered, we decided the interface should provide a reminder.

Subjects were instructed that the scenario to be implemented "would be decided" after they made all choices. No specific decision mechanism was mentioned. After subjects had made all decisions, they were shown an overview screen with their transfers for all scenarios. They were informed that they could choose themselves which scenario they wanted to apply. At this point, the uncertainty about the back transfer implementations had not yet been resolved, thus, subjects still did not know whether their chosen back transfer would occur or not. Hence, they had the chance to decide whether their transfer would reach the trustor with certainty or not. Finally, we asked subjects a set of additional questions on general dispositions and socio-demographics in a post-experimental questionnaire.

2.2. Behavioral Predictions

Assuming pure self-interest the unique subgame perfect Nash equilibrium of our game predicts that the trustee never returns any positive amount. Therefore, the trustor, anticipating this, does not transfer anything. Subjects with reciprocal concerns (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) may choose to send/return positive amounts. Based on existing evidence from trust games we expect that a substantial amount of subjects decides to reciprocate. More specifically, we expect that some trustees return positive amounts when the back transfer is certain, and weakly increase their back transfer with the amount received.

Given subjects reciprocate, we are interested in the way they behave when self-image relevant factors are varied. We use the model of Bénabou and Tirole (2011) to guide our analysis[3]. A key component of it is that beliefs about one's pro-sociality type are malleable through actions as imperfect recall is assumed. Thus, identity management becomes possible: if the cost of sending a pro-social signal (via performing a pro-social action) is small enough, pro-self types decide to invest in their identity by imitating a pro-social type[4]. Bénabou and Tirole (2011) also allow for the possibility that inferences from actions about one's type are malleable. Such inferential wiggle room exists, if the informativeness of an action about one's self-image is imperfect[5]. Inferential wiggle room allows self-deception: if the salience of a situation is low enough (reducing the signal strength of an action), pro-social types tend to make a selfish choice as the monetary gain outweighs the negative effect on the self-image.

Our experimental design manipulates self-image related aspects at two stages. First, trustees could exploit the fact that the back transfer is not executed for sure in scenarios 2–4[6]. Second, trustees could succumb to the temptation of choosing a scenario that benefits them.

Trustees may use the possible failure of the back transfer as an excuse to return less in comparison to their scenario 1 back transfer. They may tell themselves that their transfer may fail anyways and their choice will not matter for the trustor. Hence, the situations in scenarios 2 to 4 allow trustees with a desire not to appear selfish toward themselves to engage in self-deception. Essentially, their self-serving interpretation of the scenario's risk allows them to be more selfish. Exley (2015) studies choices between a certain amount and risky lotteries. She varies the recipient of both (self vs. a charity) and finds evidence of the use of risk as an excuse to give less. Haisley and Weber (2010) find such a self-serving bias caused by uncertainty in a related study involving dictator game choices under ambiguity, and Garcia et al. (2020) in the context of charitable giving. Also Di Tella et al. (2015) provide evidence

[3]Several approaches exist to model self-image concerns. See also the literature on cognitive dissonance (Festinger, 1962; Aronson, 1992; Beauvois and Joule, 1996; Konow, 2000; Spiekermann and Weiss, 2016), identity (Akerlof and Kranton, 2000), self-impression management (Murnighan et al., 2001), self-concept maintenance (Mazar et al., 2008) and other self-signaling models (Bodner and Prelec, 2003; Bénabou and Tirole, 2006; Tirole et al., 2016).

[4]Grossman (2015) focuses on this aspect of self-signaling. He tests his model experimentally in a binary probabilistic dictator game. The transfer's implementation probability is either certain or 1/3. The model predicts that giving in the low probability treatment is higher as the expected outcome-utility cost of the pro-social choice is cheapened. However, his results do not offer conclusive supporting evidence.

[5]See their parameter ν, the informativeness of an action, and their footnote 10. See also Tirole et al. (2016) who endogenize the signal strength of an action by incorporating narratives in their self-signaling model. Such a narrative, provided by others or "self-authored," may serve as an excuse for immoral behavior. It allows individuals to rationalize an immoral action even though it only contains a minimal degree of subjective plausibility.

[6]Generally, how much trustees return in scenarios 2–4 is not only affected by their social but also by their risk preferences (with respect to their own payoff and the trustor's). Trautmann and Vieider (2012) review the literature on social influences on risk attitudes and find no consistent patterns when risky decisions are taken on behalf of others. Based on a meta-analysis, Polman and Wu (2020) identify sub-domains in which there is an effect size in favor of a risky shift (e.g., non-financial choices, when emotions are involved) or a cautious shift (e.g., medical choices, decisions involving children) when people choose for others.

for self-serving interpretations and increased selfishness. See Shalvi et al. (2015) and Gino et al. (2016) for overviews of self-serving justifications, respectively motivated reasoning, used in the domain of ethical/moral behavior[7]. Thus, we expect that some trustees engage in self-deception when they make the back transfer choice. Moreover, the effect of uncertainty should be more pronounced the lower the implementation probability of the back transfer is. Thus, we expect a positive relationship between the probability and their back transfer choice.

Hypothesis 1. *In comparison to scenario 1 a non-trivial fraction of reciprocating trustees transfer back less when the transfer could fail (scenarios 2 to 4).*

Bayesian self-signaling would predict the opposite effect on trustees' back transfer choices. A decrease of the transfer's implementation probability cheapens the pro-social signal to the self. Investing in identity becomes more affordable and, in turn, more pro-social choices should result. If trustees engage in identity management, we expect a positive relationship between the probability and their back transfer choice[8].

Hypothesis 2. *In comparison to scenario 1 a non-trivial fraction of reciprocating trustees transfer back more when the transfer could fail (scenarios 2 to 4).*

A second situational excuse arises when subjects are informed that they can choose a scenario themselves. Given equal positive back transfers across scenarios, this choice implies a trade-off between the original scenario 1 that implements the back transfer for sure and a scenario that is favorable to the trustee since the transfer may fail. If this moral wiggle room affects the decision of reciprocators, a substantial amount of reciprocating trustees chooses a scenario that involves uncertainty with respect to the implementation of the back transfer.

Hypothesis 3. *When the choice of a scenario that involves uncertainty results in an expected monetary gain, a substantial amount of reciprocating trustees does not choose scenario 1.*

Finally, trustees may also have a desire not to appear selfish to others and, hence, may care about the effect of their choice on the trustor. The positive chance of a transfer failure in scenarios 2–4 allows them to return nothing as the trustor could not distinguish whether getting zero is the consequence of the trustee's choice or due to the failure of the transfer. Thus, returning nothing in scenarios 2–4 is compatible with an image of not appearing selfish to others. The reasoning follows Andreoni and Bernheim (2009) and is also in line with the prediction for the "Probability and Outcome" treatment in Grossman (2015)[9]. The choice of a scenario that involves uncertainty may be due to the manipulation's effect on the desire not to appear selfish toward oneself or toward others. Thus, our design cannot distinguish between the two at this stage. As the instructions do not explicitly mention that the chosen scenario is not communicated to the trustor, we cannot rule out that some trustees falsely believed trustors will be informed about the scenario they chose. This would eliminate the situational excuse (with respect to the scenario choice) for trustees motivated by a desire not to appear selfish to others. Only the situational excuse that affects the desire not to appear selfish toward oneself would remain.

2.3. Participants and Procedures

Using the ORSEE software (Greiner, 2004) 128 subjects were recruited among students from various disciplines at the local university[10]. In each session gender composition was approximately balanced and subjects took part only in one session. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007) and took, on average, 60 min. The average earnings in the experiment have been €14.17 (including a €2.50 show-up fee).

Upon arrival at the laboratory subjects were randomly assigned to one of the computer terminals. Each computer terminal is in a cubicle that does not allow communication or visual interaction among the participants. Subjects were given time to privately read the instructions and were allowed to ask for clarifications. In order to check the understanding of the instructions subjects were asked to answer a set of control questions. After all subjects had answered the questions correctly the experiment started. At the end of the experiment subjects were paid in cash according to their performance. Privacy was guaranteed during the payment phase.

3. RESULTS

Our analysis starts with a big picture look at the effect of transfer level and scenario on trustees' back transfer decisions. We proceed by identifying the subjects who elicit reciprocal concerns. Then, we analyze reciprocators' back transfers across scenarios as well as their scenario choices in order to test how they behave when self-image relevant factors are varied.

3.1. Analysis

We first perform random-effects panel regressions with the back transfer as the dependent variable. The panel includes all choices of a trustee (five transfer levels in three/four different scenarios). Standard errors are robust and clustered at the individual level. See **Table 1** for results. The specification in column I includes

[7]Note that uncertainty does not generally lead to more selfishness. Engel and Goerg (2018) find that dictators' transfers increase with the risk their recipient receives their endowment (expected values remaining equal).

[8]Note that an alternative motivation to increase transfers could be the desire to make sure that the same expected amount reaches the trustor, independently of the scenario.

[9]Note, however, that in our design we cannot be completely sure that trustees assume the trustor will be informed about the possibility that the transfer may fail when trustees make their back transfer choice. They may assume that trustors only learn the outcome. In that case, social-signaling does not predict an effect of a variation of the implementation probability on the incidence of zero back transfers. See Grossman (2015) for the formal argument.

[10]Pre-registering was not yet common when our study was conducted. However, we adhere to the principles proposed by Simmons et al. (2011) to tackle the problem of false-positive publications. With respect to the sample size, we decided ex-ante to collect observations from 4 sessions with 32 subjects each.

TABLE 1 | Determinants of amount returned.

	I		II		III	
Transfer	1.22***	(0.063)	1.22***	(0.063)	1.25***	(0.062)
Four scenarios	0.76	(0.65)	0.74	(0.65)	0.74	(0.65)
Implementation probability	1.38***	(0.49)				
Scenario 2 (90)			−0.25***	(0.093)	−0.21***	(0.080)
Scenario 3 (80)			−0.40***	(0.13)	−0.32***	(0.11)
Scenario 4 (50)			−0.69***	(0.24)	−0.64***	(0.23)
Transfer of 10					−0.099	(0.29)
Transfer of 10 × Scenario 2 (90)					−0.19	(0.16)
Transfer of 10 × Scenario 3 (80)					−0.40**	(0.19)
Transfer of 10 × Scenario 4 (50)					−0.22	(0.49)
Constant	−3.56	(2.46)	−2.04	(2.30)	−2.14	(2.29)
Adjusted R^2	0.46		0.46		0.46	
Observations	2,240		2,240		2,240	

Random-effects panel regression with robust standard errors; in II the scenario with certainty (1) serves as the baseline; significance levels: ∗ ∗ ∗ = 1%, ∗∗ = 5%, ∗ = 10%.

TABLE 2 | Categorization of subjects' scenario 1 back transfers.

Type	Number of subjects	Mean of returned amount when receiving				
		0	2.5	5	7.5	10
Purely selfish	11	0	0	0	0	0
Conditional cooperators	109	0.27	4.02	7.53	10.99	14.8
Humpback-shaped	8	0	4	6.37	7.87	3.87

the transfer received and the implementation probability as explanatory variables. Coefficients for both are positive and highly significant. A control dummy for the treatment with four scenarios is not significant.

The specification in column II replaces the implementation probability with dummies for scenarios 2–4. All scenario dummies are negatively correlated with the back transfer. Further tests of the dummy coefficients show that back transfers in scenario 3 and 4 are lower than back transfers in scenario 2 ($p = 0.055$, $p = 0.058$) but not significantly. Also back transfers in scenario 4 are not significantly lower than back transfers in scenario 3 ($p = 0.129$). Overall, there is a positive correlation between the trustor's transfer and the amount trustees chose to return. On average, subjects reciprocate. Moreover, on average, subjects seem to reduce the amount they send back when the scenario implies uncertainty about the implementation of their back transfer.

In a further specification shown in column III we test whether sending the maximum of 10 has an effect on trustees' behavior. For this purpose we add a dummy variable for a transfer of 10 as well as interaction terms with the three scenario dummies. The interaction between the transfer of 10 dummy and the dummy for scenario 3 is negative and significant at the 5%-level, while none of the other additional regressors is significant. Thus, results do not indicate increased selfishness among trustees when less than the full amount is transferred. A Hausman test validates the choice of a random-effects model over a fixed-effects one ($p = 0.49$). If control variables (age, gender) are included in the regression, they are not statistically significant and the reported results are not affected.

We continue the analysis at the individual level. Following Fischbacher et al. (2001) we categorize subjects based on what they return (given a transfer of 0, 2.5, 5, 7.5, or 10) when they make a choice under certainty (scenario 1), see also **Table 2**. Eleven subjects do not return anything, ever. The back transfers of 109 subjects are increasing weakly monotonically with the amount received and they are classified as conditional cooperators. Eight subjects elicit a humpback-shaped back transfer pattern. They first increase their back transfers with the amount received, but then decrease them. Our analysis considers conditional cooperators (even if they return only very little) as well as only partially reciprocating subjects (humpback-shaped pattern) as reciprocators. At the end of our analysis we will test for the robustness of our results, if humpback-shaped and selfish reciprocators are excluded.

What is reciprocating trustees' *behavior across scenarios*? More specifically, how did they behave in scenarios 2–4, that is, when there is a positive probability that their back transfer could fail? **Table 3** reports the percentage of reciprocating subjects who returned less/same/more in scenarios 2–4 (compared to scenario 1) for each amount received[11]. We perform Wilcoxon

[11]One subject selected the maximum back transfers in scenario 1, excluding them from returning more in other scenarios.

TABLE 3 | Pairwise comparison of reciprocating subjects' back transfers.

Amount received	0	2.5	5	7.5	10
Scenario 2 (90%)	6/90.6/3.4	16.2/77.8/6**	16.2/76.9/6.9**	17.1/71.8/11.1	13.7/78.6/7.7
Scenario 3 (80%)	7.7/89.7/2.6*	24.8/67.5/7.7***	27.4/60.7/11.9***	27.4/60/13.6**	23.1/65.8/11.1**
Scenario 4 (50%)	7/93/0**	29.8/57.9/12.3**	33.3/56.2/10.5***	31.6/54.4/14**	31.6/57.9/10.5***

In each cell x/y/z indicates the percentage of reciprocating subjects who returned less/same/more in the respective scenario in comparison to scenario 1. There are 117 reciprocators in scenarios 2 and 3, 57 in scenario 4. Significance of Wilcoxon signed-rank tests of choices under uncertainty compared to scenario 1 choices: ∗ ∗ ∗ = 1%, ∗∗ = 5%, ∗ = 10%.

TABLE 4 | Categorization based on the back transfer schedules across scenarios.

	Total subjects	Back transfers always 0	Decreased amount across scenarios	Returned same amount in all scenarios	Increased amount across scenarios
3 scenarios	64	4	21	30	9
4 scenarios	64	7	20	25	12
All	128	11	41	55	21
Aggregate back transfers in scenario 1, mean (st. error)	128	0	36.29 (2.61)	36.18 (2.08)	38.05 (3.71)
Transfer choice as trustor, mean (st. error)	128	2.5 (1.21)	4.88 (0.33)	4.73 (0.41)	4.76 (0.29)

signed-rank tests for each transfer level of scenarios 2–4 in order to compare reciprocating subjects' choices under uncertainty to their scenario 1 choices. The majority of subjects does not change the back transfer, yet there is a general tendency to return less under uncertainty. For a relatively high chance of transfer success (90%, scenario 2) the tendency to decrease the back transfer is only significant (at the 5%-level) for amounts received of 2.5 or 5. For an 80% chance of transfer success (scenario 3) the tendency to decrease the back transfer is significant (at least at the 5%-level) for all amounts received except 0. In scenario 4 (implementation probability 50%) the proportion of subjects who decrease is significant (at least at the 5%-level) for all amounts received.

We proceed to categorize subjects based on their choices across scenarios. For this purpose we compute, for every transfer level, the difference between back transfers in scenario 1 and 2, 2 and 3, and, if applicable, 3 and 4. The sum of these partial differences expresses how a subject reacted to the variation of the transfer implementation probability. We distinguish between three different behavioral patterns. Some trustees decreased their back transfers with the likelihood that the transfers fails. For each transfer level some trustees returned the same amount independently of the scenario. Finally, some increased their back transfers the more probable it gets that their transfer does not get implemented. **Table 4** provides frequencies of these behavioral patterns. The categories appear to be similarly represented in sessions with three and four scenarios. A χ^2 test ($p = 0.63$) does not reject that the distribution of types is the same. Out of 128 subjects (all sessions pooled), 11 never return anything, 41 decreased, 55 did not change and 21 increased the back transfer across scenarios[12]. **Table 4** also reports the mean aggregate back transfers in scenario 1 of each category, that is, the sum of the five back transfer choices. Aggregate back transfers under certainty are not significantly different across categories. Moreover, reciprocators' transfer choices as trustor do not differ across categories (4.88, 4.73, and 4.76), while the ones of purely selfish trustees are significantly lower (2.5).

Reciprocating trustees' behavior across scenarios indicates that 41 subjects reduced their back transfers with the likelihood that the transfers fail. Did these subjects tend to return zero with a positive failure probability or did they make use of the excuse in a more subtle way? Overall, the majority seems to return only slightly less, although few subjects drop their back transfer to zero in uncertain scenarios. **Figure 1** shows histograms of back transfers for scenarios 1 to 3 for a transfer of 7.5 (**Figure 1A**) and 10 (**Figure 1B**). It serves to illustrate the behavioral pattern among subjects who decreased their back transfers. Under certainty, given a transfer of 10 returning 15 corresponds to sending back half of what has been received and

[12]The categorization aggregates over choices at all five transfer levels. Hence, it could be that a subject's behavior is inconsistent across transfer levels. One out of 55 subjects categorized as returning the same amount did in fact decrease the back transfers by 2.5 at a transfer level of 2.5 and increased them by 2.5 when receiving 5. All others never deviated from their scenario 1 back transfers. Among subjects categorized as increasing the amount two slightly decreased their back transfer at a transfer level of 0. All others never lowered the back transfer. Out of 41 subjects categorized as decreasing the back transfers one subject increased the amount returned at a transfer level of 5 and one subject was inconsistent. All others never increased the back transfer. If selfish and humpback-shaped reciprocators are considered, 21 subjects are categorized as selfish, 33 decrease, 47 return the same and 19 increase the amount.

this is the most popular choice of trustees. In scenarios 2 and 3, the number of trustees returning 15 drops sharply and smaller back transfers become more common. The number of subjects returning zero increases when the failure chance of the back transfer is positive, but also those of subjects who decide to return less. We observe a similar pattern for a transfer of 7.5.

The way reciprocators handle the variation of the back transfer success rate across scenarios has implications for our analysis of the *scenario choice*. For subjects who returned the very same positive amount independently of the scenario, being able to pick a scenario unambiguously creates moral wiggle room. A subject who increased transfers across scenarios may have done so in order to invest in identity, thus conveying a signal of being pro-social. In such a case, being in a position to select a scenario with an implementation probability <1 may not be advantageous for the subject's expected utility. Finally, subjects who decreased amounts may have already exploited moral wiggle room when they made their back transfer choices in scenarios with uncertainty. They would only benefit from these choices by actually picking a scenario with an implementation probability <1. It is not clear how they would react to a "second serving" of moral wiggle room, though[13]. Hence, our analysis of the scenario choice focuses on the 55 subjects who did not vary the back transfers across scenarios.

Figure 2 shows histograms of the scenario choice for the four categories: purely selfish, decreasing, same, and increasing back transfers across scenarios. For trustees who returned the same amounts across scenarios (**Figure 2**, bottom left) the scenario choice involved the unambiguous opportunity to reap a monetary gain (in expectations). In this category, 33 of 55 subjects selected scenario 1. In contrast, 22 of them made use of the moral wiggle room and picked a scenario that did not guarantee the back transfer. Allowing for some noise in the decision making (i.e. some trustees, say 25%, select a scenario other than 1 by chance), a one-sided binomial test confirms that this fraction is significantly greater than the noise level ($p = 0.01$) and supports hypothesis 2. When subjects increased back transfers across scenarios, their choice of the scenario should not matter to them and we may expect a uniform distribution. This seems to be the case (**Figure 2**, bottom right), χ^2 tests for three scenarios ($p = 0.67$) and four scenarios ($p = 0.57$). Subjects who decreased amounts (**Figure 2**, top right) appear to have already exploited moral wiggle room when they made their back transfer choices in scenarios 2 to 4. No clear pattern with respect to their scenario choice seems evident. Finally, the scenario choice of purely selfish subjects (**Figure 2**, top left) has no consequence for their payoffs.

3.2. Discussion

In our experiment, 41 reciprocating subjects decreased their back transfer when the failure chance of the transfer was positive, an indication that they made use of this situational excuse. However, 16 of them eventually made a scenario choice that is clearly disadvantageous to them (in expected payoffs terms), while 25 selected a scenario that favors their expected payoffs. Out of 55 reciprocators who returned the same positive amount independently of the situation 22 selected a scenario that implied a positive chance that the back transfer fails to reach the trustor. The remaining 33 selected scenario 1 and made sure the back transfer reaches the trustor. They made no use of moral wiggle room in the scenario 2–4 back transfer choices and resisted the moral wiggle room provided by the scenario choice. Finally, 21 reciprocators increased the back transfer across scenarios, thus, resisting our first and evading our second manipulation. Summarizing, 47 of 117 (40%) reciprocators exploited moral wiggle room, while 70 (60%) resisted (to some extent)[14].

The back transfer choice across scenarios may be affected by two self-signaling channels and in our experiment we find evidence for both behavioral patterns[15]. However, behavior consistent with self-deception is more common as 41 subjects seem to engage in it compared to 21 whose behavior is consistent with identity management.

If the transfer choice was binary, as in the model of Grossman (2015), identity management predicts low pro-sociality types take the pro-social choice given the cost of the signal cheapens sufficiently. High types are not expected to change their behavior. They already take the pro-social choice under certainty and cannot improve on that. Alternatively, uncertainty about implementation of the back transfer would trigger self-deception processes as the self would perceive the situation as an excuse to behave more selfishly. In a binary context, high types would engage in self-deception (if the psychological cost is small enough), while low types already take the selfish choice under certainty. This implies that in a binary setting the direction of the effect of p would depend on the prevalence of low/high pro-sociality types. By design, only low types can invest in identity and self-deception is exclusive to high types. Consequently, identifying either of the behavioral pattern requires sufficient low/high types in the role of the decision maker.

In our experiment, subjects have more than two transfer options to choose from leaving both types the theoretical possibility to go either way. Unless subjects choose an extreme in scenario 1, they can adjust their transfer in both directions under uncertainty. Nevertheless, low types presumably have a higher tendency to respond with identity management and, likewise, high types are more prone to engage in self-deception. However, in our data we do not detect significant differences in the average scenario 1 back transfer across subjects who increase/decrease back transfers under uncertainty. It seems that identity management is not limited to low pro-sociality types and low as well as high types engage in self-deception.

[13]They may consciously choose the scenario that maximizes their expected payoff. However, having to pick a scenario that clearly favors them may be too much to still appear pro-social. Then, a choice of a less favorable scenario would result. Their choice might also be affected by moral balancing keeping them from engaging in self-deception two times in a row.

[14]When excluding humpback-shaped and selfish reciprocators (mean of aggregate back transfers in scenario 1 <20), 60 out of 99 (61%) resisted.

[15]An anonymous reviewer from a previous journal submission pointed us to another potential situational excuse. Trustees may think their trustor does not deserve to receive a back transfer, if the trustor's transfer is less than the maximum of 10. Our data do not indicate such a bias, though.

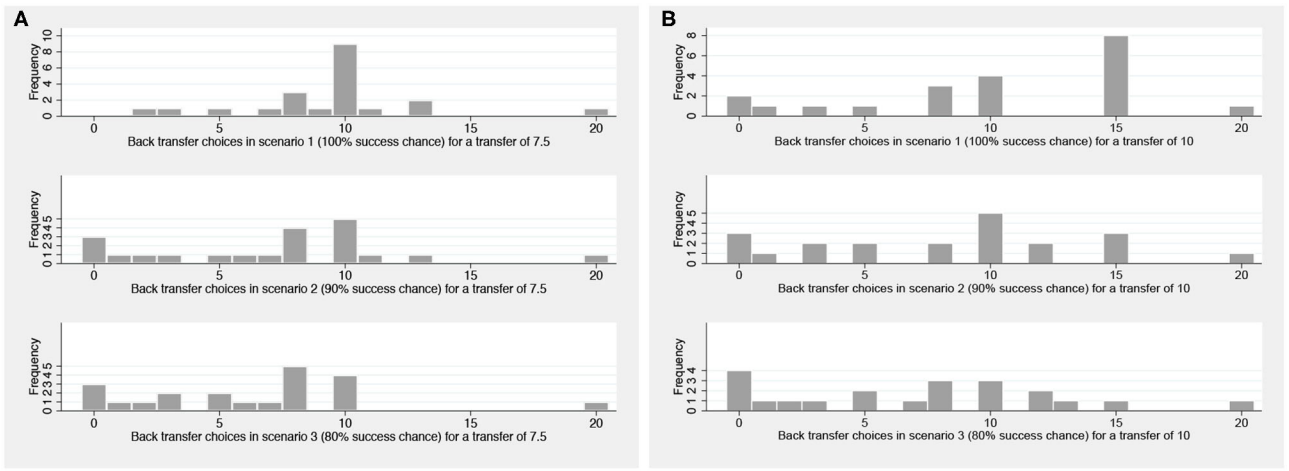


FIGURE 1 | Histograms of back transfers for scenarios 1–3 (for subjects who decreased the amount). **(A)** for a transfer of 7.5 and **(B)** for a transfer of 10.

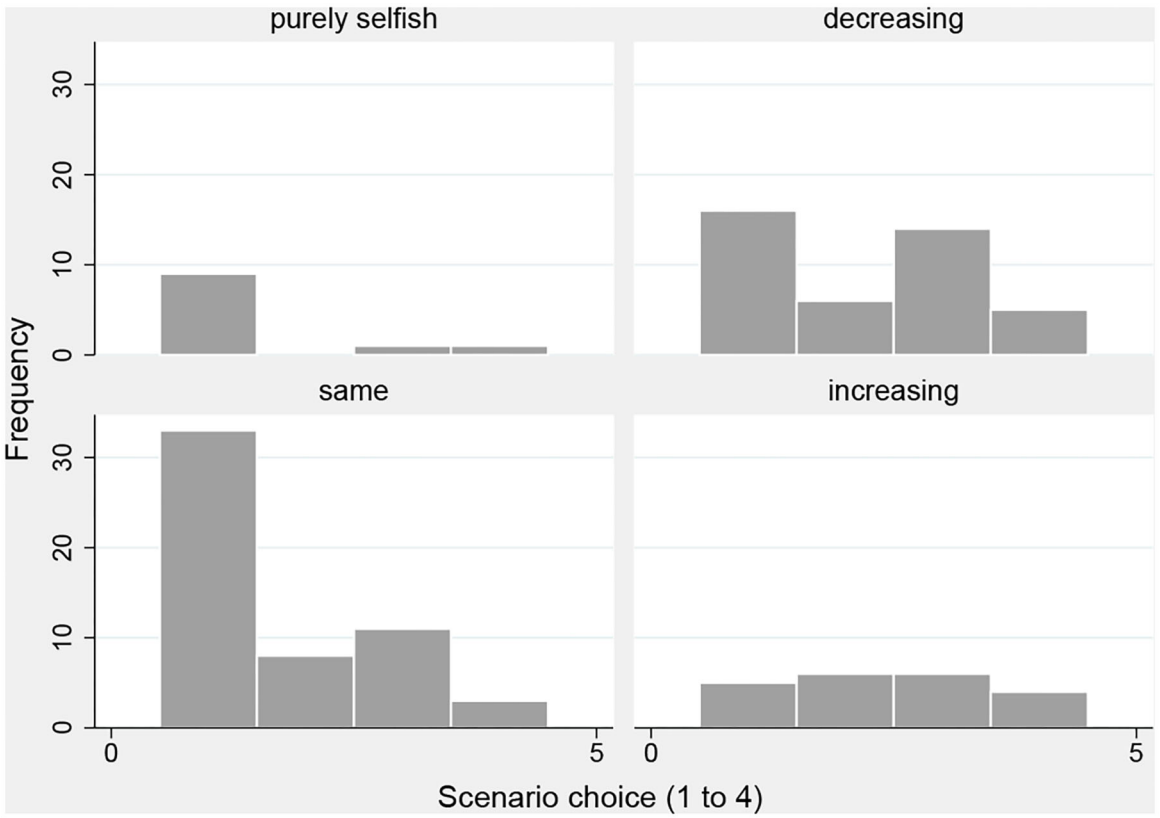


FIGURE 2 | Scenario choices by categories.

Finally, we discuss the robustness of our results in terms of the design choices we made. It is known that the use of the strategy method may encourage reciprocal behavior due to experimenter demand effects (Zizzo, 2010). In fact, we find more conditional cooperation (reciprocators) among our subjects than in Fischbacher et al. (2001), yet still within the range of results in similar studies. However, there is no indication that our within subjects variation of the scenario biases behavior in any

particular way[16]. Moreover, it is worth to note that two of our design choices made our experiment a tougher test environment for moral wiggle room to prevail than comparable experiments. First, while the side-by-side interface for entering back transfer schedules makes it easier for subjects who would like to enter the same positive amounts across scenarios to do so, it may become more difficult for subjects who have a tendency to engage in self-deception to actually do so. Since scenario 1 choices are still visible, the context of the choices under uncertainty is more salient than without the reminder. Second, we let our subjects play both roles which means that trustees are familiar with the trustor's perspective of the situation. This potential awareness about the other role may make it harder to exploit moral wiggle room in comparison to a design in which subjects only play one role.

Last but not least, we would like to stress that our implemented design does allow us to compare behavior consistent with self-deception and behavior consistent with identity management. However, it does not contrast a treatment in which self-deception is possible with a treatment that rules out self-deception. Likewise, it does not feature a treatment in which identity management is not possible. Consequently, our design is able to identify whether one effect dominates the other (at the individual level). It does not quantify the net effect of self-deception, respectively, identity management, though. In a similar design as ours, further treatments could serve as benchmarks to test the prevalence of behavior consistent with self-deception (identity management) against. More specifically, in such a self-deception only treatment the trustee would always have to pay the back transfer—independently of p—while the trustor may not receive the back transfer. This would take the chances of possibly benefitting from uncertainty off the table. In an identity management only treatment, the trustor would always get the back transfer, while there is a positive probability that the trustee does not have to pay the back transfer. This remains for future research.

4. CONCLUSION

We conducted a modified trust game in order to analyze how reciprocators respond to systematic changes of self-image relevant factors. In our experiment a substantial amount (40%) of reciprocating subjects behaved less pro-social when we introduced moral excuses for selfish behavior. That is, when the context of their choice became less salient, they succumbed to the temptation of keeping more.

This behavioral pattern is particularly interesting for the trustees' back transfer choices. Uncertainty about implementation of the back transfer may not only be perceived as a situational excuse to behave more selfishly (self-deception) but may also be interpreted as an opportunity to invest in a pro-social self-image (identity management). The two predicted effects go in opposing directions. Our results show that twice as many subjects decrease than increase their transfers under uncertainty. It seems that self-deception is prevalent when subjects make the back transfer choice. However, some trustees do increase their back transfers with more uncertainty about the implementation. It appears that self-image concerns have an ambiguous nature, in the sense that self-signaling processes can go either way: via self-deception they can lead to less giving, via identity management they can induce more giving.

Are there characteristics that distinguish individuals who are prone to self-deception from those who may invest in identity? It seems reasonable to assume that individuals who give more (pro-social types) are more likely to engage in self-deception, while those who give less (pro-self types) tend to be generous in order to boost their ego. However, our analysis does not provide support for this. Self-deceiving behavior and identity management are both used across the entire spectrum of scenario 1 back transfers.

Finally, our evidence also suggests that the effect of situational excuses extends beyond the setting of a dictator game where it has been established so far to the one of a trust game. It seems that the preference to reciprocate is also affected by the availability of situational excuses, just as the preference to give. See also Malmendier et al. (2014) and Regner (2018) for similar findings, while van der Weele et al. (2014) find no effect of a moral wiggle room manipulation in the context of reciprocity. Note, however, that our use of the strategy method—a design feature motivated by being able to test self-deception vs. identity management—can be seen as a relatively weak reciprocity environment (Casari and Cason, 2009). Although our analysis considers only subjects who do actually reciprocate, the direct response method would be regarded as a stronger setting to induce reciprocity.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Max Planck Institute of Economics, Jena. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We would like to thank audiences at IMEBE in Madrid, the University of Louvain-La-Neuve, the University of Giessen, the University of Grenoble and at the ESA meeting in Prague for their feedback. We are grateful to Paolo Crosetto for

[16]We cannot exclude, though, that presenting the scenarios in the same order may have affected subjects' choices.

valuable comments. Adrian Liebtrau provided excellent research assistance. We thank the Max Planck Institute of Economics, Jena, Germany, for financial support. TR gratefully acknowledges support by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-project number 628902.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.684078/full#supplementary-material

REFERENCES

Akerlof, G. A., and Kranton, R. E. (2000). Economics and identity. *Q. J. Econ*. 115, 715–753. doi: 10.1162/003355300554881

Andreoni, J., and Bernheim, B. D. (2009). Social image and the 50-50 norm: a theoretical and experimental analysis of audience effects. *Econometrica* 77, 1607–1636. doi: 10.3982/ECTA7384

Aronson, E. (1992). The return of the repressed: dissonance theory makes a comeback. *Psychol. Inq*. 3, 303–311. doi: 10.1207/s15327965pli0304_1

Beauvois, J., and Joule, R. (1996). *A Radical Theory of Dissonance*. London: Taylor & Francis.

Bénabou, R., and Tirole, J. (2006). Incentives and prosocial behavior. *Am. Econ. Rev*. 96, 1652–1678. doi: 10.1257/aer.96.5.1652

Bénabou, R., and Tirole, J. (2011). Identity, morals, and taboos: beliefs as assets. *Q. J. Econ*. 126, 805–855. doi: 10.1093/qje/qjr002

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games Econ. Behav*. 10, 122–142. doi: 10.1006/game.1995.1027

Bodner, R., and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *Psychol. Econ. Decis*. 1, 105–126.

Casari, M., and Cason, T. N. (2009). The strategy method lowers measured trustworthy behavior. *Econ. Lett*. 103, 157–159. doi: 10.1016/j.econlet.2009.03.012

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econ. Theory* 33, 67–80. doi: 10.1007/s00199-006-0153-z

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: avoiding altruism by distorting beliefs about others' altruism. *Am. Econ. Rev*. 105, 3416–3442. doi: 10.1257/aer.20141409

Dufwenberg, M., and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games Econ. Behav*. 47, 268–298. doi: 10.1016/j.geb.2003.06.003

Engel, C., and Goerg, S. J. (2018). If the worst comes to the worst: Dictator giving when recipient's endowments are risky. *Eur. Econ. Rev*. 105, 51–70. doi: 10.1016/j.euroecorev.2018.03.011

Exley, C. L. (2015). Excusing selfishness in charitable giving: The role of risk. *Rev. Econ. Stud*. 82, 587–628. doi: 10.1093/restud/rdv051

Falk, A., and Fischbacher, U. (2006). A theory of reciprocity. *Games Econ. Behav*. 54, 293–315. doi: 10.1016/j.geb.2005.03.001

Feiler, L. (2014). Testing models of information avoidance with binary choice dictator games. *J. Econ. Psychol*. 45, 253–267. doi: 10.1016/j.joep.2014.10.003

Festinger, L. (1962). *A Theory of Cognitive Dissonance, Vol. 2*. Stanford, CA: Stanford University Press.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ*. 10, 171–178. doi: 10.1007/s10683-006-9159-4

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Econ. Lett*. 71, 397–404. doi: 10.1016/S0165-1765(01)00394-9

Garcia, T., Massoni, S., and Villeval, M. C. (2020). Ambiguity and excuse-driven behavior in charitable giving. *Eur. Econ. Rev*. 124:103412. doi: 10.1016/j.euroecorev.2020.103412

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated bayesians: Feeling moral while acting egoistically. *J. Econ. Perspect*. 30, 189–212. doi: 10.1257/jep.30.3.189

Greiner, B. (2004). *The Online Recruitment System Orsee 2.0 - a guide for the Organization of Experiments in Economics*. Mimeo, Department of Economics, University of Cologne.

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Manag. Sci*. 60, 2659–2665. doi: 10.1287/mnsc.2014.1989

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *J. Econ. Behav. and Organ*. 117, 26–39. doi: 10.1016/j.jebo.2015.05.008

Grossman, Z., and van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *J. Eur. Econ. Assoc*. 15, 173–217. doi: 10.1093/jeea/jvw001

Haisley, E. C., and Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games Econ. Behav*. 68, 614–625. doi: 10.1016/j.geb.2009.08.002

Hamman, J. R., Loewenstein, G., and Weber, R. A. (2010). Self-interest through delegation: an additional rationale for the principal-agent relationship. *Am. Econ. Rev*. 100, 1826–1846. doi: 10.1257/aer.100.4.1826

Konow, J. (2000). Fair shares: accountability and cognitive dissonance in allocation decisions. *Am. Econ. Rev*. 90, 1072–1091. doi: 10.1257/aer.90.4.1072

Larson, T., and Capra, C. M. (2009). Exploiting moral wiggle room: Illusory preference for fairness? a comment. *Judgm. Decis. Mak*. 4, 467–474.

Malmendier, U., te Velde, V. L., and Weber, R. A. (2014). Rethinking reciprocity. *Annu. Rev. Econ*. 6, 849–874. doi: 10.1146/annurev-economics-080213-041312

Matthey, A., and Regner, T. (2011). Do i really want to know? a cognitive dissonance-based explanation of other-regarding behavior. *Games* 2, 114–135. doi: 10.3390/g2010114

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res*. 45, 633–644. doi: 10.1509/jmkr.45.6.633

Murnighan, J. K., Oesch, J. M., and Pillutla, M. (2001). Player types and self-impression management in dictatorship games: two experiments. *Games Econ. Behav*. 37, 388–414. doi: 10.1006/game.2001.0847

Polman, E., and Wu, K. (2020). Decision making for others involving risk: a review and meta-analysis. *J. Econ. Psychol*. 77:102184. doi: 10.1016/j.joep.2019.06.007

Regner, T. (2018). Reciprocity under moral wiggle room: Is it a preference or a constraint? *Exp. Econ*. 21, 779–792. doi: 10.1007/s10683-017-9551-2

Shalvi, S., Gino, F., Barkan, R., and Ayal, S. (2015). Self-serving justifications doing wrong and feeling moral. *Curr. Dir. Psychol. Sci*. 24, 125–130. doi: 10.1177/0963721414553264

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci*. 22, 1359–1366. doi: 10.1177/0956797611417632

Spiekermann, K., and Weiss, A. (2016). Objective and subjective compliance: a norm-based explanation of 'moral wiggle room'. *Games Econ. Behav*. 96, 170–183. doi: 10.1016/j.geb.2015.11.007

Tirole, J., Falk, A., and Bénabou, R. (2016). *Narratives, Imperatives and Moral Reasoning*. Technical report, Mimeo.

Trautmann, S. T., and Vieider, F. M. (2012). "Social influences on risk attitudes: applications in economics," in *Handbook of Risk Theory* (Dordrecht: Springer), 575–600.

van der Weele, J. J. (2014). *Inconvenient Truths: Determinants of Strategic Ignorance in Moral Dilemmas*. Available at SSRN 2247288.

van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *Am. Econ. J.* 6, 256–264. doi: 10.1257/mic.6.3.256

van der Weele, J. J., and von Siemens, F. A. (2020). Bracelets of pride and guilt? an experimental test of self-signaling. *J. Econ. Behav. Organ*. 172, 280–291. doi: 10.1016/j.jebo.2020.02.001

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Exp. Econ*.13, 75–98. doi: 10.1007/s10683-009-9230-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 13 August 2021
doi: 10.3389/fpsyg.2021.701627

Can We Commit Future Managers to Honesty?

Nicolas Jacquemet[1], *Stéphane Luchini*[2]*, *Julie Rosaz*[3] and *Jason F. Shogren*[4]

[1] Paris School of Economics, Univ. Paris 1 Panthéon-Sorbonne, Paris, France, [2] Aix Marseille Univ, CNRS, AMSE, Marseille, France, [3] Univ Lyon, Université Lumière Lyon 2, GATE UMR 5824, Ecully, France, [4] Department of Economics, University of Wyoming, Laramie, WY, United States

In a competitive business environment, dishonesty can pay. Self-interested executives and managers can have incentive to shade the truth for personal gain. In response, the business community has considered how to commit these executives and managers to a higher ethical standard. The MBA Oath and the Dutch Bankers Oath are examples of such a commitment device. The question we test herein is whether the oath can be used as an effective form of ethics management for future executives/managers—who for our experiment we recruited from a leading French business school—by actually improving their honesty. Using a classic Sender-Receiver strategic game experiment, we reinforce professional identity by pre-selecting the group to which Receivers belong. This allows us to determine whether taking the oath deters lying among future managers. Our results suggest "yes and no." We observe that these future executives/managers who took a solemn honesty oath as a Sender were *(a)* significantly more likely to tell the truth when the lie was detrimental to the Receiver, but *(b)* were not more likely to tell the truth when the lie was mutually beneficial to both the Sender and Receiver. A joint product of our design is our ability to measure in-group bias in lying behavior in our population of subjects (comparing behavior of subjects in the same and different business schools). The experiment provides clear evidence of a lack of such bias.

Keywords: commitment, lying, In-group bias, managers, honesty, Oath, business ethics

JEL classification: C92, D03, D63.

OPEN ACCESS

Edited by:
Nora Szech,
Karlsruhe Institute of Technology (KIT), Germany

Reviewed by:
Eberhard Feess,
Victoria University of Wellington, New Zealand
Hannes Rau,
Heidelberg University, Germany

***Correspondence:**
Stéphane Luchini
stephane.luchini@univ-amu.fr

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *28 April 2021*
Accepted: *15 June 2021*
Published: *13 August 2021*

Citation:
Jacquemet N, Luchini S, Rosaz J and Shogren JF (2021) Can We Commit Future Managers to Honesty? Front. Psychol. 12:701627. doi: 10.3389/fpsyg.2021.701627

1. INTRODUCTION

Recent scandals in the business community have raised serious concerns about whether the competitive culture fosters dishonest behavior for the sake of personal profit (see e.g., Cohn et al., 2014). This trepidation has created a growing interest for professional oaths in business (de Bruin, 2016). Observers inside and outside the business community have suggested that future managers, like graduating MBA students, should take a voluntary MBA oath—a commitment to an ethical standard of integrity and honesty (Anderson and Escher, 2010, see, e.g., http://mbaoath.org/). Started in 2001 by a coalition of 2,000 MBA students from the Harvard Business School, the MBA Oath initiative now covers graduates, advisors and alumni signers from over 500 MBA programs around the world.

But to our knowledge there exists no formal assessment of whether and how a voluntary solemn oath impacts the integrity of future business executives and managers. Building on our recent research on the behavioral impacts of a truth-telling oath (Jacquemet et al., 2017b, 2020, 2021), herein we explore whether future managers respond with more honesty to a voluntary oath that promotes truth-telling. We do so in the context of a laboratory experiment by recruiting

students from a renowned business school in France and ask them to perform a classic Sender-Receiver strategic gaming experiment in which dishonesty pays[1]. Each participant sees two rolls of a computerized dice and is asked to communicate the results to another person whose choice will determine the final payoff (see Erat and Gneezy, 2012). We consider two cases: Selfish lies—which are self-beneficial but detrimental to the receiver—and Pareto lies—which are mutually beneficial, i.e., a win-win "white lie." This game structure allows us to both investigate a rich set of lying behavior thanks to changes in the payoff structure, and to reduce the risk that truth-telling occurs because of sophisticated deception (Sutter, 2009).

Our main treatment variable is a truth-telling oath that participants are offered to sign before they learn the exact nature of the subsequent experiment. The oath procedure has been designed by Jacquemet et al. (2013) in such a way that compliance is voluntary, and most subjects do choose to comply—all subjects do sign the oath in this experiment, while the average is closer to 95% putting together all truth-telling oath experiments that has been carried out over the years. According to accumulated evidence in social psychology, compliance with the oath can commit subjects to truth-telling in subsequent decisions that are aligned with the content of the oath (see, e.g., Joule and Beauvois, 1998; Cialdini, 2007). Jacquemet et al. (2019) show that a requirement for the truth-telling oath to be effective is to remind people when "a lie is a lie" (a condition called "loaded environment")—because a neutral environment gives subjects more "room to wiggle" and to rationalize lying behavior under oath. They show that, without an oath, selfish lies (resp. Pareto lies) decrease from 41.7% (resp. 68.3%) to 35.7% (resp. 60.0%) when lying is made explicit. In the neutral environment, the oath has no effect on Pareto lies and decreases the proportion of selfish lies to 36.7%. But when lies are made explicit, the oath decreases the proportion of Pareto and selfish lies to 36.7 and 16.7%, respectively. We design our experiment to make lying explicit, and implement the "loaded environment" condition of Jacquemet et al. (2019). This design choice also rules out the possibility that the oath affects truth-telling behavior because the wording of the oath gives subjects a social cue about the appropriate behavior (Haley and Fessler, 2005; Rigdon et al., 2009)[2].

As a well-targeted subject pool of future executives and managers, we recruited students from a renowned French business school as our oath takers. We assign these future managers to always be in the role of the Sender, and we contrast the lying behavior between future managers under oath to those in a no-oath condition to measure the behavioral effect of a commitment to honesty within this subgroup of population. The obvious challenge to the identification of the effect of professional identity is self-selection into a particular profession, leading to a spurious correlation between behavior and professional identity that goes through unobserved, group-specific, individual heterogeneity. The usual strategy to overcome this issue is to implement exogenous variations in the provision of environmental cues associated with professional identity (Benjamin et al., 2010). According to "self-categorization theory" in social psychology (Turner, 1985), this manipulation makes professional identity more salient and leads subjects to rely more on the norms associated with this identity (this idea that behavior is induced by the norms associated with the identity to which people give more weight due to the circumstances of the choice is at the core of the economics of identity literature initiated by Akerlof and Kranton, 2000, 2010). Following Shih et al. (1999), our instrument to make identity more salient is the group identity of the matched partner: from the same business school or from a different field of study in another school. We purposefully pair our future managers with a partner from either the same business school as the identity inducement condition or from another discipline at another university as the control. This design choice generates data on the effect of the oath on future managers whose professional identity is salient; this feature also provides evidence as to whether an in-group bias drives lying behavior among future managers.

Our results are 3-fold. First, without the oath, future managers lie in both the Pareto and Selfish Lie cases—we observe more dishonesty for Pareto lies (79%) relative to Selfish lies (33%). The magnitude of Selfish lying is similar to that of students in other non-business fields, around 33%. Second, lying is of the same magnitude whether future managers are matched with a peer from the same school or not; we do not observe significant in-group/out-group effects in lying behavior. Third, the oath significantly reduces lying for the Selfish lie case (lying declined by 70%); but the oath had no significant effect on reducing Pareto lies (lies dropped by 14%). This lack of behavioral response, however, does not mean subjects are insensitive to the oath when telling mutually beneficial lies. Using "happiness" as a proxy for subject's internal response, we find that the oath makes lying psychologically more costly—making lying under oath more problematic than without—although not to an extent that is sufficient to change behavior when lying is payoff maximizing for both sides.

2. DESIGN OF THE EXPERIMENT

The experiment closely follows the extension by Jacquemet et al. (2019) of the sender-receiver game first introduced in Erat and Gneezy (2012). The design relies on three treatment variables: the type of lie (Selfish/Pareto, within-subjects), the group identity (in-group/out-group, between subjects) and the oath (no oath/oath, between subjects), implemented using a $2 \times 2 \times 2$ factorial design.

[1]While the control over the environment offered by a laboratory experiment is better suited to testbed the effect of the oath on future managers, it raises the obvious concern of whether these results would extend to a non-laboratory setting. On this issue, we refer the reader to the literature that correlates lying in lab experiments to unethical behavior in the field (e.g., Potters and Stoop, 2016; Dai et al., 2017; Hanna and Wang, 2017; Cingl and Korbel, 2020) and the ones that investigate both truth-telling behavior under oath in the context of field experiments (Carlsson et al., 2013; Koessler et al., 2019; Jacquemet et al., 2021) ot the long-lasting effects of promises (Peer and Feldman, 2021).

[2]An alternative interpretation of the effect of the oath is the idea that it simply primes subjects to truth-telling. Such an effect is not only unlikely based on the existing literature (Pashler et al., 2013), but is also ruled out by this design choice since the rules of the game themselves make an explicit reference to lying.

Sender-Receiver Game

Two players, a sender and a receiver (labeled 'player A' and 'player B' in the written instructions, see the **Appendix**, section Experimental Instructions, for an English translation of the original instructions in French) are randomly matched. The computer randomly draws a 6-sided die, and informs the sender about the outcome. The sender is then asked to choose between 6 possible messages to send to the receiver: "*The outcome of the roll of die was* [1, 2, .., 5, 6]." Our game replicates the 'loaded environment' condition of Jacquemet et al. (2019). Accordingly, we explicitly label untruthful communication a "lie" and truthful communication the "truth" in the written instructions. Based on the content of the message, the receiver is asked to choose a number in the set [1, 2, 3, 4, 5, 6], which determines the payment of both subjects between two payment options, X and Y. Only the sender knows the actual payoffs generated by each option. If the number chosen by the receiver matches the die roll, both subjects are paid based on option *X*; otherwise, *Y* is implemented. This is common knowledge to all subjects.

Types of Lie

Following Erat and Gneezy (2012), we use the combination of payoffs associated with each option as an experimental device to distinguish different types of lies. The payoffs are always set to (20; 20) for the sender and the receiver if the sender's choice matches the die roll (option *X*). In the "selfish lie" condition, the payoffs implemented by option *Y* are (21; 15): if the sender chooses to lie and the receiver follows the sender's message, the lie imposes a loss of 5 on the receiver, while the sender gains 1. We summarize the type of lie accordingly based on the variation in payoff induced by a lie as $\mathbb{T}[-5; 1]$. In the Pareto lie, by contrast, both the sender and the receiver benefit from the lie: the payoffs implemented by option *Y* are (30; 30), and the lie is accordingly denoted $\mathbb{T}[10; 10]$. We facilitate inter-study comparison by purposefully selecting the values of the payoff parameters in each treatment to closely follow those used in Erat and Gneezy (2012), and Jacquemet et al. (2019). This choice leads to an asymmetry in the sender's benefit from lying. These two conditions are implemented within subjects in a random order to control for order effects. The roles are fixed, but subjects are randomly rematched with a different subject between the two conditions. To avoid the confounding effect of changes in wealth over the two repetitions of the game, only one condition is binding to determine the actual payment given to subjects at the end of the experiment. Subjects receive no additional information about the other player's decisions or payoffs until the end of the experiment.

Group Identity

We manipulate group identity within pairs thanks to the school in which participants currently study. The lab is located close to a leading business school, whose students represent a large share (47% as of march 2021) of the lab's subject pool. The master in management offered by this business school lies in the world top-10 according to Quacquarelli Symonds, 2020[3] ranking and the school itself is part of the 2021 Financial Times top-100 business schools over the world. The experiment focuses on the lying behavior of future managers trained in this school. We always assign the role of sender to a student from this business school. In the in-group treatment (IN), the receiver is also a future manager coming from the same business school. In the out-group treatment (OUT), the receiver is a student from another school or university, and specialized in a discipline other than Management. These affiliations are made salient on the decision screen for subjects in both roles: once the role has been announced on the screen, a message appears informing participants that "*The player A (B) with whom you will interact studies (*OUT*: does not study) at [Name of the business school].*"

Oath

Before entering the laboratory, each subject is first invited (one by one) to enter an adjacent office. The other subjects could neither hear nor see what happened in the office, as the door was always closed before the start of the procedure. In the NO-OATH condition while in this office, subjects randomly draw a sheet of paper from an envelope presented to them by the experimenter. The paper indicates the name of the seat they are assigned to in the lab. They are then invited to enter the lab using a side door located between the lab and the office, and the monitor invites the next subject to enter the office.

This no-oath procedure is also applied to receivers in the OATH condition. In contrast, the Senders are exposed to the truth-telling oath procedure designed by Jacquemet et al. (2013). Once they entered the office, subjects are first presented with a form untitled "Solemn oath" (see the **Appendix**, section English Translation of the Original Oath Form in French, for an English translation of the original form in French). They are asked to read the form and to decide "*freely whether they want to sign it or not*" (the experimenter follows a written script to make sure the subjects are all exposed to exactly the same procedure). The monitor makes clear to subjects that they are free to sign the form, and that neither participation to the experiment nor experimental earnings are conditional on their decision. Whatever their choice, subjects must give the form back to the experimenter, are thanked and invited to draw their seat according to the procedures implemented in the NO-OATH condition. To avoid communication between subjects prior to the experiment, one monitor stayed in the laboratory during the entire process and helps them find their seat in the room. Subjects receive no information about whether *(i)* other subjects were exposed to the oath procedure, or *(ii)* whether anyone else decided to sign the oath or not[4].

Control Variables

A key driver of senders' behavior in our experiment rests in the potential for group-specific attitude toward lying. While senders all belong to the same group, the group of receivers in our experiment can differ by business school, which allow us to measure such heterogeneity. To that end, receivers participate to a simplified (3-sided) version of the dice under the cup task

[3] https://www.qs.com/

[4] An obvious methodological concern with.

introduced by Fischbacher and Föllmi-Heusi (2013). Subjects then roll a three-sided dice available on their desk. They are told they can roll the dice as many times as they wish, but must report the outcome of the last trial. They are paid for this part according to their report: they earn 0 if they report 1, 1 if they report 2, and 2 if they report 3. Since our aim is to measure *ex ante* heterogeneity, this task is implemented at the start of the experiment, before the sender-receiver game. Senders are not exposed to this task; which allows us to compare their behavior to other experiments using the same sender-receiver game. At the end of the experiment, subjects are asked to fill in several questionnaires aimed at further measuring individual heterogeneity: their level of happiness (7 points Likert scale); their self-reported honesty (7 points Likert scale); the perceived honesty of other subjects (7 points Likert scale); and two measures of cognitive abilities: a 10 items version of the Raven (2008)'s progressive matrices test, and the Cognitive Reflection Test (Frederick, 2005, three reflection questions that must be answered within 60 s). This is followed by Gough (1979)'s Creative Personality Scale (a self-report personality inventory for creativity assessment), which is unrelated to the current paper. Last, we elicit the feeling of closeness to other students based on the "Inclusion of the Other in the Self" scale (IOS, a visual, 7-points, task that measures closeness thanks to the overlap between two circles introduced as representing the other and oneself; Aron et al., 1992; Gächter et al., 2015): subjects are asked about their closeness first with students from the business school and then with students not studying at the business school. The experiment ends with a socio-demographic questionnaire asking participants about their gender, their age, their level of study and the number of times they already participated to an experiment.

2.1. Procedures

To implement the group identity treatment variables, subjects were invited separately to the same sessions depending on whether they registered in our subject pool database (managed using HROOT, Bock et al., 2014) as students currently enrolled in the business school or not. We have separate registration lists that allow us to distinguish the group to which subjects belong. Upon arrival, participants are called one by one by their name to check the registration information: they enter the private (oath) office at this stage. The first 10 participants whom we call all are enrolled in the business school list, and are assigned to a computer whose pre-determined role is set to sender. This allows us to control the group identity of senders in all conditions, and to implement the oath procedure on senders only in the OATH conditions.

Once in the laboratory, participants are informed about the instructions of each step of the experiment on their screen (the experiment is computerized using a software developed on Z-tree, Fischbacher, 2007). They can push a button located inside their cubicle to ask a question in private to a monitor at any point in time. All payments in the experiment are expressed in *Experimental Currency Unit*. The exchange rate is 1 ECU = 0.3 euros. Participants are paid a fixed fee for of 5 euros for answering the post-experimental survey, which is added to the payment that results from their decision in the payoff condition of the sender-receiver game that is randomly drawn (and the outcome from the dice-under the cup task for receivers). The average individual payoff is 12.02 euros for an average 1 h participation. We ran 13 sessions (with 260 participants, among whom 130 are senders) of the experiment at the GATE-Lab between September 2019 and February 2020. **Table 1** provides the allocation of sessions and participants across the four between-subjects experimental treatments. All 63 senders who participated to an OATHsession agreed to sign the oath. This ensures the behavioral effect of the oath cannot be attributed to self-selection into compliance with the oath request.

2.2. Manipulation Checks

Before moving to the results of the experiment, we use our control variables to provide an overview of the identifying variations induced by the experimental design. The first key dimension in our experiment is to compare interactions of senders with future managers to interactions with non-future managers based on the group identity treatment manipulation. Among the 160 participants who play as a receiver in the experiment, 72 are assigned to the IN conditions and are not enrolled in the same business school as senders. The field of study of these OUT participants make it highly unlikely they belong to the same group as participants from the target business school: 68% of them are enrolled in one of the two engineering school that are located close to the laboratory. The remaining participants study chemistry, medicine, biology, law, political science and arts. Only three participants study fields that are close to business studies: two in economics, and one in management.

In our design, the manipulation of the group identity of the receiver is instrumental and aims to reinforce the self-identity of senders as future managers. We check the internal validity of the consequences of this group assignment by comparing the answers to the two IOS questions among the senders (following, e.g., Harris et al., 2015). The results unambiguously support that perceived closeness reacts to the treatment manipulation: future managers feel closer to their fellow, with an average closeness equal to 3.96, than to subjects from the other school (3.09, the difference is highly significant, $p < 0.001$, according to paired-sample Wilcoxon rank sum test). This difference prevails whether senders participate to the NO-OATH (4.13 vs. 3.30, $p < 0.001$) or to the OATH condition (3.78 vs. 2.86, $p < 0.001$).

An important confounding effect in sender-receiver games is the possibility that lying arises as an attempt to counteract the willingness of receivers not to follow the message received (called "sophisticated deception" by Sutter, 2009). To ascertain that senders will not react to treatments because they expect receivers to react differently to their message, we check whether receivers behavior is similar between treatments. Among future managers (IN condition) 75.0% of receivers decide to follow the message they receive. This proportion is slightly higher among receivers in the OUT conditions, who follow the message 70.1% of the time

TABLE 1 | Sample sizes.

	Total	No-oath-Out	No-oath-In	Oath-Out	Oath-In
Nb. of sessions (senders)	13 (160)	3 (34)	3 (33)	4 (38)	3 (25)

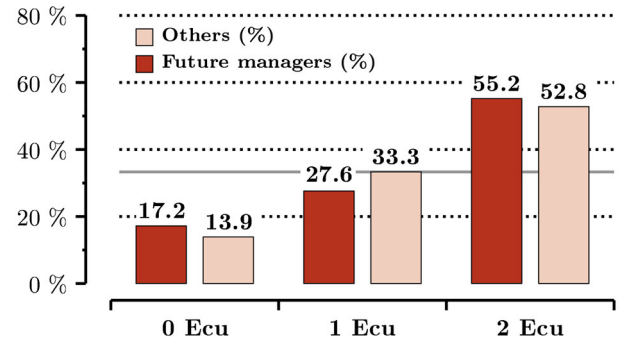


FIGURE 1 | Empirical distribution of dice draws among receivers.

($p = 0.678$, χ^2 bootstrap test)[5]. We observe the same lack of difference in receivers' behavior regarding the implementation of the oath: 69.8% of receivers in OATH and 74.6% of receivers in NO-OATH decide to follow the message ($p = 0.736$, χ^2 bootstrap test).

In **Figure 1**, we provide evidence that group identity does not translate into differences in individual attitudes toward lying, based on the outcomes from the dice under the cup task performed by receivers. We plot the distribution separately within the group of future managers and non-future managers. The horizontal line in gray displays the theoretical benchmark—the uniform distribution that would result from perfectly truthful reports. Among both groups, the empirical distribution of responses clearly departs from the benchmark: the proportion of reported draws that give rise to no earnings is under-represented ($p = 0.014$ for future managers, $p < 0.001$ for other subjects; two-sided proportion test) whereas the report that pays the most is over-represented ($p < 0.001$ and $p = 0.010$; two-sided proportion test). The two almost perfectly balance, as the middle report is in line with the theoretical expectation for both groups ($p = 0.430$ and $p = 0.607$; two-sided proportion test). Importantly, lying behavior is overall similar between future managers and other subjects ($p = 0.530$, two-sided χ^2 bootstrap test). Under the assumption that receivers' behavior is representative of their group, we conclude that the behavior of senders in our experiment cannot be attributed to group-specific attitudes toward lying.

Last, our experiment incidentally provides evidence on in-group-bias in lying behavior by focusing on the NO-OATH conditions. We observe the induced differences in perceived identity do not translate into different lying behavior in the sender-receiver game. Irrespective of the type of lie, 56.0% of messages sent by future managers in NO-OATH are dishonest. We observe a small difference in the proportion of dishonest messages between the IN and the OUT conditions. When future managers send messages to future managers, 51.5% are dishonest as compared to 60.3%. **Figure 2** provides a more detailed overview of the pattern of lying in the two conditions. Both the joint distribution of lies over the two games ($p = 0.731$, bootstrap χ^2 test) and the two marginal distributions of selfish and Pareto lies ($p = 0.963$ and $p = 0.650$; bootstrap proportion tests) are statistically no different across the two NO-OATH conditions.

[5]We test the differences between treatments at the individual level to account for potential within-subject correlation in receivers' behavior across the two lying conditions.

3. RESULTS

3.1. Unconditional Treatment Effects

We first look at the overall effect of the oath on lying behavior in the sender-receiver game by pooling the two (i.e., IN and OUT) OATH and the two NO-OATH conditions. Overall, offering future managers the possibility to sign a truth-telling oath decreases lying by 26.8%. Irrespective of the type of lie, the overall proportion of dishonest messages decreases from 56% in NO-OATH to 41% in OATH ($p = 0.070$, one-sided bootstrap test). **Figure 3** reports the joint and marginal distributions of lying across the two sender-receiver games. The joint distribution clearly shows that the oath induces a drastic increase in the share of fully honest future managers (who send an honest message in both games): their proportion in OATH is more than three times higher ($p < 0.001$, one-sided bootstrap test). This increase in the share of fully honest messages is compensated by a decrease in the share of each one of the three possible patterns of lie. As a results of these sharp differences, the joint distribution is significantly different between OATH and NO-OATH ($p = 0.004$, bootstrap χ^2 test). The marginal distributions, displayed on the right-hand side of the figure, indicate that the oath is much more powerful on selfish lies, that happen at the expense of the receiver, than on Pareto lies, that are mutually beneficial to the sender and the receiver. The share of selfish lies is more than twice lower among subjects under oath ($p = 0.012$, one-sided proportion bootstrap test). The slight decrease in Pareto lies is not significant ($p = 0.153$, one-sided proportion bootstrap test), and such behavior remains widespread even under oath.

In **Figure 4**, we compare the marginal effect of the oath on lying behavior between IN and OUT. For each of the two, we report the joint and marginal distributions of the difference in the proportion of lies between NO-OATH and OATH. The marginal effect of the oath on the share of subjects who decide not to lie is very much alike in the two situations. The main difference rests in how this change is obtained. In OUT, full honesty mainly results from a drastic drop in the share of full liars, whereas in IN it mainly comes from a decrease in the share of subjects who decide to lie only when this behavior is selfish. This is confirmed by the comparison of the marginal distributions: the decrease in the share of Pareto lies is much higher in OUT, while the marginal effect of the oath is similar on selfish lies in both conditions.

3.2. Conditional Treatment Effects

We check the robustness of our unconditional results by estimating a multinomial logit model that controls for individual covariates. The dependent variable is lying behavior as defined by the joint distribution: "no lie", "selfish only", "Pareto only" and "two lies". We use honesty ("no lie") as a reference, so that the coefficients for each of the 3 remaining outcomes can be interpreted in a natural way—a negative sign indicates a decrease in the corresponding lying behavior. We introduce

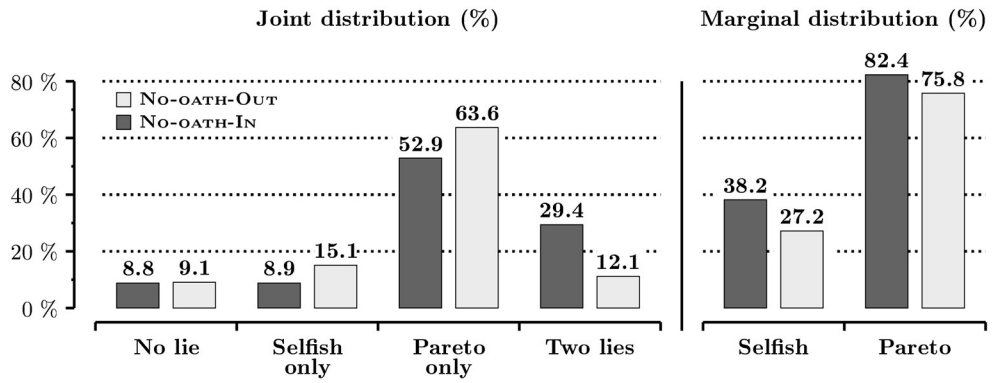


FIGURE 2 | Distributions of lies in the sender-receiver game in the NO-OATH conditions.

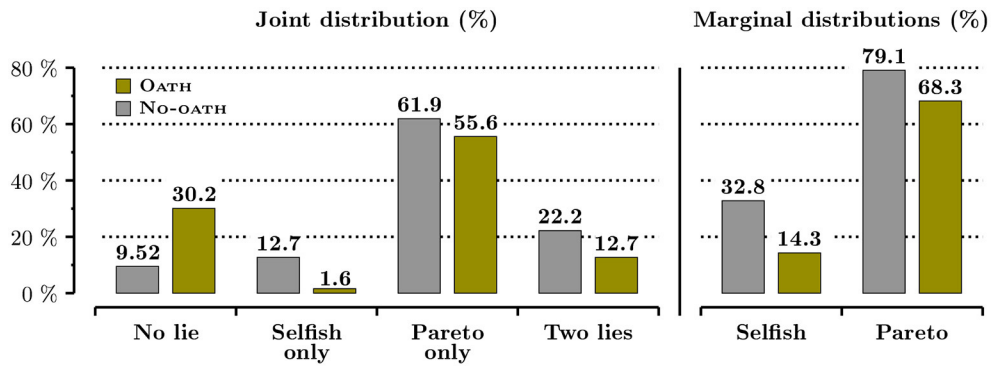


FIGURE 3 | The effect of the oath on the distribution of lying behavior in the sender-receiver game.

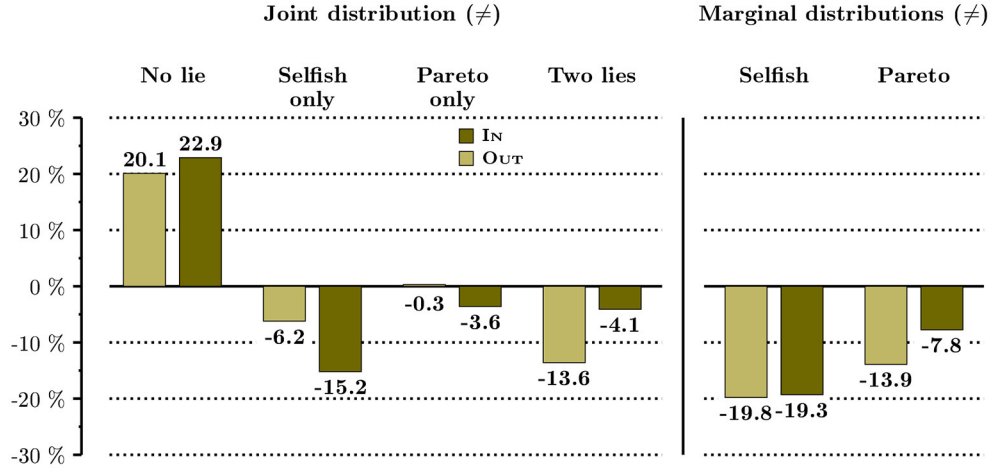


FIGURE 4 | Distributions of the marginal effect of the oath.

treatment variables and their interactions, as well as subject's age, gender, closeness to other subjects, cognitive abilities (as measured by Raven and CRT scores) and participation to previous experiments.

The estimated parameters, presented in **Table 2**, confirm that the unconditional results still hold when we introduce subjects' characteristics. First, whether future managers are interacting with a fellow manager or with a student from another field of study has no effect on behavior. In contrast, we also find the oath significantly decreases all types of lying. Results also highlight that observed heterogeneity have very little predictive power on the likelihood that subjects tell a Pareto lie only, or lie in both

TABLE 2 | Parametric estimation of the treatment effects.

	Selfish only		Pareto only		Two lies	
	Parameter estimate	*p*-value	Parameter estimate	*p*-value	Parameter estimate	*p*-value
Constant	−14.940	0.062	2.771	0.484	−0.424	0.933
Out	0.721	0.571	0.042	0.964	1.201	0.260
Oath× In	−18.696	0.000	−1.661	0.045	−2.020	0.079
Oath× Out	−4.282	0.008	−1.825	0.023	−2.314	0.011
Age	0.622	0.047	0.041	0.822	0.128	0.571
Male	1.485	0.148	0.713	0.180	0.431	0.519
Closeness, BS	−0.357	0.379	−0.305	0.113	−0.198	0.421
Closeness, not BS	−0.190	0.605	−0.209	0.313	−0.432	0.096
Raven score	0.684	0.066	0.111	0.421	0.029	0.870
CRT score	−13.854	0.000	−0.357	0.148	−0.203	0.521
Past experience	−0.339	0.397	−0.238	0.217	0.027	0.907

Multinomial logit model on the effect of individual characteristics on the likelihood to behave according to one of the four possible lying patterns in the experiment: Selfish lie only, Pareto lie only, or two lies (the reference is honesty in both instances). All explanatory variables are individual specific and do not vary at the individual level; each column reports the estimated effect (along with its p-value) of the corresponding covariate on the likelihood the outcome behavior arises.

instances. Selfish lie only stands as a notable exception: subjects who are older and performed better at the Raven test are more likely to engage into this type of lie. On the contrary, subjects who performed better at the CRT test (i.e., who override incorrect intuitive responses and engage in further reflection) are less likely to make a selfish lie only.

3.3. Does the Oath Only Affect Self-Serving Dishonesty?

To sum-up, our experiment provides clear evidence that a truth-telling oath disciplines lying behavior among future managers, but only if lying is detrimental to others. When lying rather serves both the sender's and the receiver's interest, by contrast, we observe very little to no behavioral response to the oath. The obvious question raised by those results is whether lying is perceived as dishonest when it is mutually beneficial, and if yes why we do not observe the same drastic decrease as when lying is selfish.

To answer this question, we use the self-reported level of happiness collected at the end of the survey as a proxy of the internal response of subjects to the oath (see e.g., Clark, 2015, for a discussion of the internal validity of self-reported well-being as a mesaure of individual well-being). Since the level of happiness itself is difficult to interpret, we focus on variations between responses (see Jacquemet et al., 2017a, for a similar approach) and focus in **Figure 5** on the level of self-reported happiness among senders who truthfully report the outcome of the dice ("truth") and those who lie, separately by treatment. This boxplot displays the interquartile range, i.e., the distance between the upper (75th percentile) and lower (25th percentile) quartiles. Wiskers present the 10th percentile on the bottom and the90th percentile on the top end. The bold horizontal line displays the median.

Focusing on NO-OATH subjects, we find no change in the median level of happiness according to lying behavior in both the selfish lie (the median level of happiness among liars is 5.5 vs. 6 among truth-tellers; $p = 0.280$, bootstrap KS test) and the Pareto lie (6 vs. 5.5; $p = 0.710$, bootstrap KS test) situations. Under oath, by contrast, lying comes with a sharp shift in happiness as compared to truth-telling for both kinds of lies: the median level of happiness is lower among liars as compared to truth-tellers in the context of both selfish lies (4 vs. 5; $p = 0.026$, KS bootstrap test) and Pareto lies (5 vs. 6; $p = 0.006$, KS bootstrap test). Median happiness reaches its lowest level, equal to 3.5, among subjects who engage in both a selfish and a Pareto lie ($p = 0.013$ as compared to happiness among subjects who engage in Pareto lies only; KS bootstrap test)[6]. We observe a strong internal response of subjects to the oath when they decide to lie, whatever the lying situation. This response suggests that the oath makes lying psychologically more costly. The lack of behavioral response to the oath in the Pareto lie situation suggests that the benefits of a Pareto lie still outweigh the cost of lying under oath.

4. CONCLUSION

Despite a disappearance of occupational oaths at the end of the 20th Century (Prodi, 1992), the idea of the oath has gained renewed momentum in recent years following the economic crisis and recent business scandals as a form of ethics management. In this paper, we provide the first experimental evidence of the efficacy of the oath to foster integrity of future managers. Our measure of integrity is lying behavior in two classic sender-receiver games, one mutually beneficial (win-win white lies) and one self-serving (selfish lies), in which senders are recruited from the same leading business school. In our experimental design, we vary the group to which the receiver belongs so we can test how the strength of one's professional identity (in-group or out-group) affects the behavior of our subjects.

Our baseline framework (no honesty oath) leads to several useful findings. First, in contrast with what has been observed among criminals (Cohn et al., 2015), we do not find any effect of the professional identity of managers on their lying behavior. Second, thanks to the instrumental manipulation of the group matching of subjects, our study also contributes to the burgeoning literature about in-group biases in lying behavior. The existing evidence that relies either on the minimal group paradigm (Tajfel, 1970), or on natural identities, is mixed. For example, Butler (2014) finds reduced in-group lying when identity is artificial and Maximiano and Chakravarty (2016) find a similar result with natural identities. In contrast, our results are aligned with those from the study that is closest to ours: using natural identities (based on university enrollment), Feldhaus and Mans (2014) do not find any difference in lying behavior in a sender-receiver game between in- and out-group interactions (examples of null results using the minimal identity paradigm include Benistant and Villeval, 2019; Casoria et al., 2020). While we replicate these results and confirm their robustness, we also

[6]We cannot compare this figure to the level of happiness of subjects who only engage in a selfish lie as only one subject does so in the oath treatment.

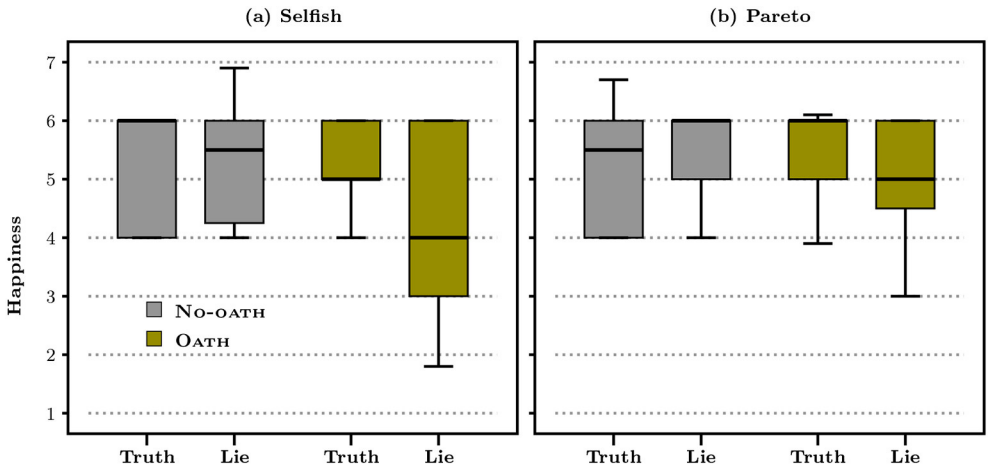


FIGURE 5 | Happiness by treatment and type of lie.

add control variables allowing to measure both the intensity of the group manipulation and the attitudes toward lying among groups. Despite a significant change in perceived closeness when interactions happen within groups, we confirm the lack of difference between in-group and out-group interactions. Importantly, this happens in a context in which attitudes toward lying are no different between groups.

For both these in-group and the out-group conditions, our main treatment variable of interest is a truth-telling oath that senders are free to sign—and which all subjects do agree to sign. Overall, our results suggest that a solemn oath like the MBA oath can increase the honesty of our future managers when the lie is for selfish reasons. The oath was less powerful on future managers, however, in reducing the frequency of "white lies" or win-win lies. This departs from previous evidence about the truth-telling oath obtained in the same setting but with students from other disciplines. Specifically, Jacquemet et al. (2019) show that (i) Pareto lies are less widespread than with future managers (60.0 vs. 79.1% herein); (ii) strongly react to the oath, with a share of Pareto lies under oath equal to 36.1% (as compared to 68.3%). Although the behavioral responses are drastically different, our results suggest that future managers do react to the oath even in the Pareto lie condition: self-reported happiness data show that the oath makes Pareto lies psychologically more costly, although not to an extent that is sufficient to undermine win-win lying behavior. An important difference between managers and the lay public is the rise in the "win-win" culture, a paradigm that promotes the alignment of interest of stakeholders, in business education and practice (see e.g., Cook, 2017, for a discussion and a historical perspective). We speculate that managers face more salient conflicting motivations when lying is mutually beneficial — an issue we leave for future research.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by GATE-Lab Review Board for ethical standards in research-Groupe d'Analyse et de Theorie Economique (GATE UMR CNRS 5824). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the French National Research Agency Grants ANR-17-EURE-001 and ANR-17-EURE-0020, and by the Excellence Initiative of Aix-Marseille University - A∗MIDEX.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.701627/full#supplementary-material

REFERENCES

Akerlof, G. A., and Kranton, R. E. (2000). Economics and Identity. *Q. J. Econ.* 115, 715–753. doi: 10.1162/003355300554881

Akerlof, G. A., and Kranton, R. E. (2010). *Identity Economics*. Princeton, NJ: Princeton University Press.

Anderson, M., and Escher, P. (2010). *The MBA* Oath: Setting a Higher Standard for Business Leaders. Penguin.

Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *J. Pers. Soc. Psychol.* 63, 596–612. doi: 10.1037/0022-3514.63.4.596

Benistant, J., and Villeval, M. C. (2019). Unethical behavior and group identity in contests. *J. Econ. Psychol.* 72, 128–155. doi: 10.1016/j.joep.2019.03.001

Benjamin, D. J., Choi, J. J., and Strickland, A. J. (2010). Social Identity and Preferences. *Am. Econ. Rev.* 100, 1913–1928. doi: 10.1257/aer.100.4.1913

Bock, O., Baetge, I., and Nicklisch, A. (2014). Hroot: hamburg registration and organization online tool. *Eur. Econ. Rev.* 71, 117–120. doi: 10.1016/j.euroecorev.2014.07.003

Butler, J. V. (2014). Trust, truth, status and identity: an experimental inquiry. *B.E. J. Theor. Econ.* 14, 293–338. doi: 10.1515/bejte-2013-0026

Carlsson, F., Kataria, M., Krupnick, A., Lampi, E., Lofgren, A., Qin, P., et al. (2013). The truth, the whole truth, and nothing but the truth-a multiple country test of an oath script. *J. Econ. Behav. Organ.* 89, 105–121. doi: 10.1016/j.jebo.2013.02.003

Casoria, F., Reuben, E., and Rott, C. (2020). *The Effect of Group Identity on Hiring Decisions With Incomplete Information*. Available at SSRN 3731536.

Cialdini, R. B. (2007). *Influence: The Psychology of Persuasion, Vol. 55*. Collins, NY: Harper Collins.

Cingl, L., and Korbel, V. (2020). External validity of a laboratory measure of cheating: evidence from Czech juvenile detention centers. *Econ. Lett.* 191:109094. doi: 10.1016/j.econlet.2020.109094

Clark, A. E. (2015). "SWB as a measure of individual well-being," in *Oxford Handbook of Well-Being and Public Policy*, eds M. Adler and M. Fleurbaey (Oxford: Oxford University Press), 518–552.

Cohn, A., Fehr, E., and Maréchal, M. (2014). Business culture and dishonesty in the banking industry. *Nature* 516, 86–89. doi: 10.1038/nature13977

Cohn, A., Maréchal, M. A., and Noll, T. (2015). Bad boys: how criminal identity salience affects rule violation. *Rev. Econ. Stud.* 82, 1289–1308. doi: 10.1093/restud/rdv025

Cook, R. (2017). Why win-win-win propositions are the future of business. Forbes, Retrieved on april 2021. Available online at: https://www.forbes.com/sites/vinettaproject/2017/06/28/why-win-win-winpropositions-are-the-future-of-business/

Dai, Z., Galeotti, F., and Villeval, M. C. (2017). Cheating in the lab predicts fraud in the field: an experiment in public transportation. *Manage Sci.* 64, 1081–1100. doi: 10.1287/mnsc.2016.2616

de Bruin, B. (2016). Pledging integrity: oaths as forms of business ethics management. *J. Bus. Ethics* 36, 23–42. doi: 10.1007/s10551-014-2504-1

Erat, S., and Gneezy, U. (2012). White lies. *Manage Sci.* 58, 723–733. doi: 10.1287/mnsc.1110.1449

Feldhaus, C., and Mans, J. (2014). *Who do You Lie to? Social Identity and the Cost of Lying*. Technical Report 76, University of Cologne, Department of Economics.

Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10, 171–178. doi: 10.1007/s10683-006-9159-4

Fischbacher, U., and Föllmi-Heusi, F. (2013). Lies in disguise. An experimental study on cheating. *J. Eur. Econ. Assoc.* 11, 525–547. doi: 10.1111/jeea.12014

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42. doi: 10.1257/089533005775196732

Gächter, S., Starmer, C., and Tufano, F. (2015). Measuring the closeness of relationships: a comprehensive evaluation of the 'inclusion of the other in the self' scale. *PLoS ONE* 10, e0129478. doi: 10.1371/journal.pone.0129478

Gough, H. G. (1979). A creative personality scale for the adjective check list. *J. Pers. Soc. Psychol.* 37, 1398–1405. doi: 10.1037/0022-3514.37.8.1398

Haley, K. J., and Fessler, D. M. T. (2005). Nobody's watching?: subtle cues affect generosity in an anonymous economic game. *Evol. Hum. Behav.* 26, 245–256. doi: 10.1016/j.evolhumbehav.2005.01.002

Hanna, R., and Wang, S.-Y. (2017). Dishonesty and selection into public service: evidence from India. *Am. Econ. J. Econ. Policy* 9, 262–290. doi: 10.1257/pol.20150029

Harris, D., Herrmann, B., Kontoleon, A., and Newton, J. (2015). Is it a norm to favour your own group? *Exp. Econ.* 18, 491–521. doi: 10.1007/s10683-014-9417-9

Jacquemet, N., James, A., Luchini, S., and Shogren, J. (2017a). Referenda under oath. *Environ. Resour. Econ.* 67, 479–504. doi: 10.1007/s10640-016-0023-5

Jacquemet, N., James, A. G., Luchini, S., Murphy, J. J., and Shogren, J. F. (2021). Do truth-telling oaths improve honesty in crowd-working? *PLoS ONE* 16:e0244958. doi: 10.1371/journal.pone.0244958

Jacquemet, N., Joule, R.-V., Luchini, S., and Shogren, J. F. (2013). Preference elicitation under Oath. *J. Environ. Econ. Manag.* 65, 110–132. doi: 10.1016/j.jeem.2012.05.004

Jacquemet, N., Luchini, S., Malézieux, A., and Shogren, J. (2020). Who'll stop lying under oath? Experimental evidence from tax evasion games. *Eur. Econ. Rev.* 20, 103369. doi: 10.1016/j.euroecorev.2020.103369

Jacquemet, N., Luchini, S., Rosaz, J., and Shogren, J. F. (2019). Truth-Telling under oath. *Manage Sci.* 65, 426–438. doi: 10.1287/mnsc.2017.2892

Jacquemet, N., Luchini, S., Shogren, J., and Zylbersztejn, A. (2017b). Coordination with communication under Oath. *Exp. Econ.* 21, 627–649. doi: 10.1007/s10683-016-9508-x

Joule, R., and Beauvois, J. (1998). *La Soumission Librement Consentie*. Paris: Presses Universitaires de France.

Koessler, A.-K., Torgler, B., Feld, L. P., and Frey, B. S. (2019). Commitment to pay taxes: Results from field and laboratory experiments. *Eur. Econ. Rev.* 115, 78–98. doi: 10.1016/j.euroecorev.2019.02.006

Maximiano, S., and Chakravarty, S. (2016). *Deception, Social Preferences, and Friendship*. SSRN Scholarly Paper ID 3589537, Rochester, NY: Social Science Research Network.

Pashler, H., Rohrer, D., and Harris, C. R. (2013). Can the goal of honesty be primed? *J. Exp. Soc. Psychol.* 49, 959–964. doi: 10.1016/j.jesp.2013.05.011

Peer, E., and Feldman, Y. (2021). Honesty pledges for the behaviorally-based regulation of dishonesty. *J. Eur. Public Policy* 28, 761–781. doi: 10.1080/13501763.2021.1912149

Potters, J., and Stoop, J. (2016). Do cheaters in the lab also cheat in the field? *Eur. Econ. Rev.* 87, 26–33. doi: 10.1016/j.euroecorev.2016.03.004

Prodi, P. (1992). *Il Sacramento Del Potere: Il Giuramento Politico Nella Storia Costituzionale Dell'Occidente, Vol. 15*. Bologna: IL Mulino.

Raven, J. (2008). "General introduction and overview: the raven progressive matrices tests: their theoretical basis and measurement model," in *Uses and Abuses of Intelligence*, Chapter 1, eds J. Raven and J. Raven (Edinburgh: Competency Motivation Project), 17–68.

Rigdon, M., Ishii, K., Watabe, M., and Kitayama, S. (2009). Minimal social cues in the dictator game. *J. Econ. Psychol.* 30, 358–367. doi: 10.1016/j.joep.2009.02.002

Shih, M., Pittinsky, T. L., and Ambady, N. (1999). Stereotype susceptibility: identity salience and shifts in quantitative performance. *Psychol. Sci.* 10, 80–83. doi: 10.1111/1467-9280.00111

Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *Econ. J.* 119, 47–60. doi: 10.1111/j.1468-0297.2008.02205.x

Tajfel, H. (1970). Experiments in intergroup discrimination. *Sc.i Am.* 223, 96–103. doi: 10.1038/scientificamerican1170-96

Turner, J. C. (1985). "Social categorization and self-concept: a social cognitive theory of group behavior," in *Advances in Group Process: Theory and Research*, ed E. J. Lawler (Greenwich: JAI Press), 77–121.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 August 2021
doi: 10.3389/fpsyg.2021.701294

No Moral Wiggle Room in an Experimental Corruption Game

Loukas Balafoutas[1]*, *Fedor Sandakov*[2] *and Tatyana Zhuravleva*[2]

[1] Department of Public Finance, University of Innsbruck, Innsbruck, Austria, [2] Higher School of Economics University, Moscow, Russia

Recent experimental evidence reveals that information is often avoided by decision makers in order to create and exploit a so-called "moral wiggle room," which reduces the psychological and moral costs associated with selfish behavior. Despite the relevance of this phenomenon for corrupt practices from both a legal and a moral point of view, it has hitherto never been examined in a corruption context. We test for information avoidance in a framed public procurement experiment, in which a public official receives bribes from two competing firms and often faces a tradeoff between maximizing bribes and citizen welfare. In a treatment where officials have the option to remain ignorant about the implications of their actions for citizens, we find practically no evidence of information avoidance. We discuss possible reasons for the absence of willful ignorance in our experiment.

Keywords: information avoidance, corruption, negative externality, experiment, Russia
JEL codes: C91; D73; D83; D91

OPEN ACCESS

Edited by:
Agne Kajackaite,
Social Science Research Center Berlin, Germany

Reviewed by:
Robert Stüber,
Social Science Research Center Berlin, Germany
Daniel Parra,
Social Science Research Center Berlin, Germany

***Correspondence:**
Loukas Balafoutas
loukas.balafoutas@uibk.ac.at

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *27 April 2021*
Accepted: *23 July 2021*
Published: *18 August 2021*

Citation:
Balafoutas L, Sandakov F and Zhuravleva T (2021) No Moral Wiggle Room in an Experimental Corruption Game. Front. Psychol. 12:701294. doi: 10.3389/fpsyg.2021.701294

INTRODUCTION

As with many types of criminal activity, individuals who are prosecuted by the law due to corruption sometimes argue that they were not aware of corrupt activity taking place, or at least that they did not knowingly participate in such activity. The possibility that someone is not aware of the harm that he or she creates is relevant from a legal, but also from a moral point of view: In particular, virtue and deontological ethics base their value judgments not on the consequences of an action, but on the action itself or on the character of the person who takes it. However, having no positive knowledge of a corrupt act does not necessarily exonerate an individual. An important and pertinent question is, *could that individual have known* of the wrongdoing in question? In 1977, the US Congress enacted the Foreign Corrupt Practices Act (FCPA), which stipulates that knowledge of a corrupt activity goes beyond actual knowledge and extends to conscious disregard, deliberate ignorance, and willful blindness. This means that individuals who willingly ignore indications of wrongdoing in their area of responsibility—despite believing that a high probability of wrongdoing exists—can face criminal liability in cases of bribery and corruption[1].

The above considerations motivate us to ask the following question in the present study: Do decision makers create moral wiggle room by choosing to remain blind to information in a corruption context, even though this information is potentially critical in distinguishing between corrupt (but privately profitable) and non-corrupt actions? A second, related question

[1]To mention one relevant example, in 2019, Former SNC-Lavalin CEO Pierre Duhaime pleaded guilty of helping a public official commit breach of trust, in a corruption scandal in connection with building a hospital in Quebec, Canada. He explicitly admitted to being willfully blind to the scandal and looking the other way, even though he was aware of corrupt actions inside his organization. Even though Mr. Duhaime was not shown to have actively engaged in the particular activity nor to have enjoyed any personal financial benefits from it, he was sentenced to 20 months in house arrest for remaining willfully blind to the scandal.

is whether corruption becomes more frequent when the possibility to engage in willful blindness is present—in other words, whether moral wiggle room is exploited. Moreover, motivated by the large costs imposed by bribery and corruption on third parties and by the open debate in the literature regarding the role of negative externalities for corrupt activities, we search for behavior consistent with moral wiggle room exploitation under different levels of negative externalities created by corruption.

We test for the presence of willful ignorance by decision makers in a lab experiment, which is meant to capture essential elements of a corruption setting. The decision situation in the experiment mimics a case of public procurement, where firms compete for a government contract. A *public official* purchases a service from a firm, and two competing *firms* may bribe the official in order to win the contract. There is also a *citizen* whose payoff is determined by the performance of the firm that wins the contract. We use framed and loaded instructions, in order to ensure that participants better understand the nature of the interaction and to enhance the ecological validity of our findings[2]. Experimental bribery games with participants in the role of decision makers (e.g., public officials), firms, and—sometimes also—affected third parties are very common in the literature on corruption (for a survey, see Abbink and Serra, 2012). Our experimental setting is similar to the corruption game used in Jaber-López et al. (2014), Schram et al. (2019), and García-Gallego et al. (2020). One notable difference is that the externality of corruption in our experiment is endogenous and determined through the performance difference between the two firms who compete for the government contract[3].

Corruption is as widely prevalent around the world as it is costly (Svensson, 2005). In recent years, a growing body of literature has departed from neoclassical models of crime-and-punishment calculations that model corruption as the outcome of expected payoff maximizing calculations by economic actors (such as the seminal works by Becker, 1968 or Klitgaard, 1988). In addition, given the illegal nature of the phenomenon, reliable observational data on corruption are often hard to obtain, which in turn has led to a surge in research using data from the economic lab. This recent literature has offered experimental evidence on several (behavioral) aspects relating to corruption, such as social norms and culture (Cameron et al., 2009; Barr and Serra, 2010; Salmon and Serra, 2017; Schram et al., 2019), gender (Alatas et al., 2009), monitoring and punishment (Abbink et al., 2002; Armantier and Boly, 2011; Serra, 2012; Ryvkin et al., 2017), wages and appointment procedures of public officials (Azfar and Nelson, 2007), legal immunity for bribe givers (Abbink et al., 2014), transparency (Khadjavi et al., 2017; Parra et al., 2019), audience effects and observability (Salmon and Serra, 2017; García-Gallego et al., 2020).

Also relevant to our work are previous studies that have examined experimentally the role of negative externalities for the incidence of corruption and, in particular, the hypothesis that higher externalities should lead to lower levels of corruption, ceteris paribus. Interestingly, this is not always the case (Abbink et al., 2002; Büchner et al., 2008; Barr and Serra, 2009). Recently, Guerra and Zhuravleva (2021) examined the role of negative externalities and social norms in a corruption context. The focus of that study lies not on corrupt behavior *per se*, but on the willingness of unaffected bystanders to engage in third-party punishment of corrupt activities. Guerra and Zhuravleva (2021) find that bystanders are unresponsive to the variation in the negative externality, while Guerra and Zhuravleva (2020) report that female bystanders increase punishment when the externality goes up, while male bystanders decrease it. Overall, the effect of externalities on corruption remains an open research question that our work contributes to.

The possibility that decision makers exploit moral wiggle room in order to engage in more corrupt activities has hitherto not been examined in the economic literature, but it is related to a body of research reporting that participants in experiments involving distributional decisions often willfully avoid information regarding the consequences of their actions (e.g., Konow, 2000; Dana et al., 2007; Kajackaite, 2015; Grossman and Van der Weele, 2017; Regner, 2018). Broadly speaking, the existence of moral wiggle room allows decision makers to increase their monetary income by means of more selfish actions at the cost of other individuals, without incurring too high losses in terms of social image and self-image. In our experiment, we apply this notion to a corruption game. The extent to which individuals create and exploit moral wiggle room is a question of particular importance in the case of corruption, given the high societal costs associated with corrupt activities. Moreover, while it seems clear that the social norm in settings such as the dictator game involves at least some degree of pro-social behavior (Krupka and Weber, 2013) and selfish actions can produce high moral costs in the absence of moral wiggle room, the social norm in bribery games is not necessarily as well-established. Depending on the cultural background and broader context, bribe maximizing behavior may not represent a severe norm violation in some cases, reducing the need for bribe taking individuals to engage in willful ignorance in order to preserve their self-image. If this is true, willful ignorance may be less relevant in the context of corruption.

Information avoidance and the exploitation of moral wiggle room can be viewed as part of a larger literature on motivated reasoning, which refers to the idea that individuals avoid, distort, or misinterpret information in order to maintain a certain set of beliefs, from which they draw positive utility. The only study we are aware of that examines motivated reasoning in a corruption setting is Di Tella et al. (2015), who find that dictators are more likely to believe they are interacting with a dishonest recipient

[2]The question of using neutral versus loaded instructions is debated in the experimental literature. The pioneering study of Abbink and Hennig-Schmidt (2006) does not find a significant difference between neutral and loaded framing in a bribery game. On the other hand, Ajzenman (2021) argues that people who observe that corruption is widespread are more willing to engage in corrupt behavior. Thus, using loaded instructions might have a different effect in different cultures. Our data are collected in one country, which means that the choice of loaded instructions is very unlikely to affect comparisons across treatments.

[3]The main reason for allowing firm performances to be determined endogenously by means of a real effort task has been to better capture real world settings, in which firms have a lot of discretion over the quality of their services. Another important feature of this setting is that it creates a variation in firm quality, without artificially imposing it.

when they stand to gain more by behaving selfishly themselves. The corruption game used in Di Tella et al. (2015) is, however, very different from the one in our experiment, as it does not allow for information acquisition and essentially captures a case of embezzlement by an authoritarian ruler, while our game captures cases of collusive bribery featuring bribing firms, bribe-taking officials, and inactive but affected third parties. Thus, while both studies deal with corruption in a context of motivated reasoning, they refer to very different forms of corruption and institutional settings, measure different kinds of outcomes, and approach the topic through different perspectives.

EXPERIMENTAL DESIGN

The Public Procurement Game

The experimental setting features three roles: *Public officials*, *firms*, and *citizens*. At the beginning of each session, all participants are randomly assigned one of these roles and interact in groups of four, consisting of one official, two firms, and one citizen. Participants keep their roles, but groups are re-shuffled in each round using a perfect stranger matching protocol. All groups play five rounds of the game described below, and one round is randomly selected and paid out at the end of the experiment[4]. At the beginning of the game, officials and firms receive an initial endowment of 10 ECU (where 1 ECU = 30 Ruble). Citizens receive no endowment.

Figure 1 graphically represents the stages of the game. At Stage 1, each firm carries out the real-effort task used in Weber and Schram (2017), which consists of adding numbers and is described in section The Real-Effort Task. Each firm achieves a performance, which can vary between 0 and 10. This performance determines the payment of the citizen in the following way: The performance of the firm that wins the contract is multiplied by a factor of either 1 or 2 depending on the treatment (see section Treatments), and the resulting number is the citizen's income in ECU. Hence, notice that—in contrast to most other studies on corruption—the negative externality created by corruption is endogenous in this setting.

At Stage 2, each firm observes its own performance and has the opportunity to offer a non-negative amount (bribe) to the official out of their endowment. Every time a firm offers a bribe to the official, he or she pays an irrevocable transaction cost of 1 ECU, irrespective of the size of the bribe. This is meant to capture initiation costs of the briber when he approaches a public official (Abbink et al., 2002). Then, at Stage 3 of the experiment, the official receives the information about the firms' bribes and their performances and decides which firm wins the government contract. The winning firm receives an additional 10 ECU. The official keeps the bribe of the winning firm, while the bribe of the losing firm is transferred back to that firm.

An experimental session includes 24 participants. At the final stage of the experiment, participants complete a survey that includes basic socio-demographic information and a few questions from the World Values Survey (see **Appendix D** for a list of all survey questions). In addition to payoffs from the game, each participant receives 5 ECU if he or she completes the survey at the end.

Treatments

The above description refers to the baseline treatment, which we call the *Full Information* treatment. In order to examine the presence of willful ignorance, we implement a further treatment with *Information Avoidance*. The only difference between the two treatments is at Stage 3, where the public official receives only the information about the two firms' bribes as a default option in the *Information Avoidance* treatment and has the option to also receive information about the firms' performance. This is implemented as follows: On his or her decision screen, the public official sees the two firms' bribes, and there is also a "Reveal performances" button that they can click on if they wish to. If they choose to click on that button, they see the two firms' performances on their screen.

In addition, we implement a variation in the size of the negative externality that is created by the official whenever he or she does not select the firm with the highest performance as the winner of the government contract. This is achieved by having one treatment with a *High Externality*, in which the performance of the winning firms is multiplied by a factor of 2 in order to determine the income of the citizen, and one treatment with a *Low Externality*, in which the performance of the winning firm is multiplied by a factor of 1. Hence, the experiment exploits a 2 × 2 treatment variation with four treatments in total, as shown in **Table 1**.

The Real-Effort Task

We use the real-effort task developed by Weber and Schram (2017) and used previously in corruption experiments (Schram et al., 2019; Di Zheng et al., 2020). On their computer monitor, firms see two 7 by 7 matrices filled with two-digit numbers. Their task is to find the largest number in each of the two matrices and add them up (see **Figure A1** for an illustration). After entering their answer, a new set of randomly chosen matrices appears on the next screen, irrespective of whether the number entered was correct or not. This is an individual task and each firm has 3 min to solve as many of these matrix summations as they can. Firms' total performance in the task is a proxy for firm efficiency and determines citizen welfare in our setting. The maximum possible performance is 10, i.e., if a firm solves more than 10 matrices, its final performance is reduced to 10 (this was known to all participants). On the monitor, firms can see the remaining time and also the number of attempts and correct trials. At this stage, public officials and citizens wait.

Hypotheses

While models of rational decision making predict that individuals would prefer to have more information when making decisions, the empirical literature has demonstrated that information is often avoided. A key explanation is that the lack of information serves as an excuse for selfish behavior.

[4]We let participants play five rounds of the game (instead of only one) because a tension between bribe and welfare maximization not always exists. Having five rounds practically ensures that all officials faced such a tradeoff in at least some periods. We refer to this issue again in the results section.

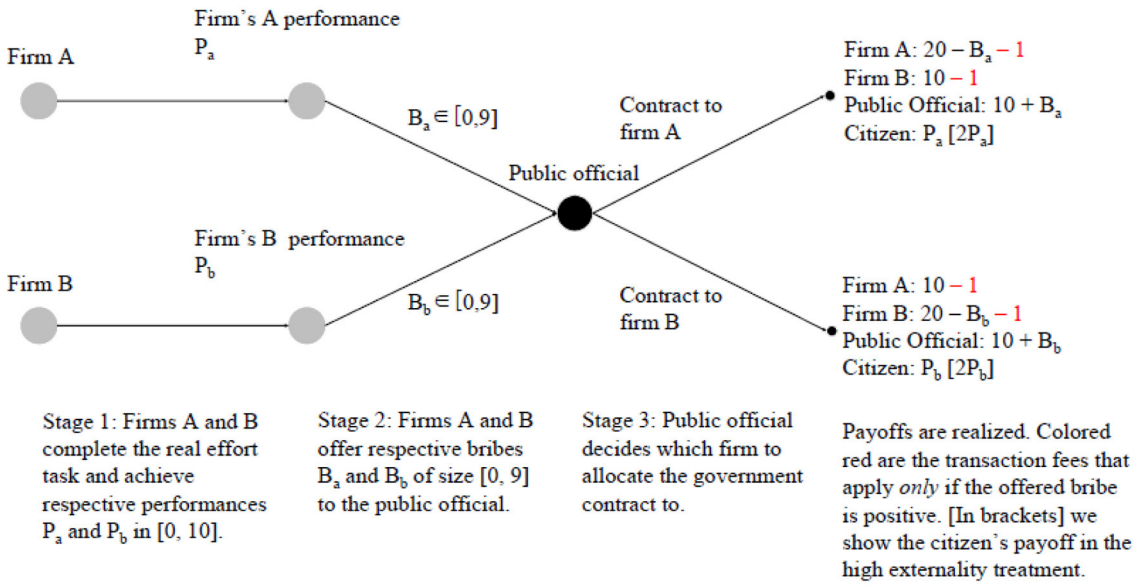


FIGURE 1 | The game tree.

TABLE 1 | Experimental treatments.

	Full information	Information avoidance
	Treatment 1:	***Treatment 2:***
High externality	Public official always knows firms' bribes and performances Citizen's payoff = Performance of winning firm, multiplied by 2	Public official always knows firms' bribes and has the option to reveal performances Citizen's payoff = Performance of winning firm, multiplied by 2
	Treatment 3:	***Treatment 4:***
Low externality	Public official always knows firms' bribes and performances Citizen's payoff = Performance of winning firm	Public official always knows firms' bribes and has the option to reveal performances Citizen's payoff = Performance of winning firm

Specifically, Grossman and Van der Weele (2017) and Serra-Garcia and Szech (2019) show theoretically that willfully chosen ignorance is a compromise between material interest and a desire to maintain self-image. Empirical work confirms this claim, although the extent of this compromise varies with the setting and the parameters of interaction. For instance, in Van der Weele (2014) the share of participants who remain ignorant in a dictator game varies between 6 and 31%, depending on cost and benefit parameters. Grossman (2014) shows that the share of dictators who remain ignorant depends crucially on whether ignorance is an act of commission or omission. Recent literature explores willful ignorance in altruistic punishment (Kriss et al., 2016; Stüber, 2019) and shows that approximately a third of participants decide to remain ignorant about selfish dictators, in order to avoid the costs of punishing them. Felgendreher (2018) finds very little evidence for willful ignorance in a very different context (purchase of ethically certified products) compared to distributional games typically played in the economic lab.

Summing up the findings in previous literature, they have generally shown that willful ignorance is common, but it is also sensitive to the conditions and consequences that come with it. It is thus important to determine the extent to which people make a trade-off between material interests and self-image in a corruption context. Our first pre-registered hypothesis is that officials will exploit opportunities to avoid information and follow the selfish strategy more often.

H1: More officials will choose to maximize bribes in the treatments with information avoidance than in those without.

Our design additionally allows us to examine the role of externalities for corrupt behavior. We expect that, as long as (at least some) officials are concerned about the well-being of citizens, doubling the size of the externality will lead to more frequent choices that maximize citizen welfare. This is motivated by the extant experimental literature showing that individuals are driven by pro-social motives, including a taste for efficiency (see, e.g., Charness and Rabin, 2002; Engel, 2011). More specifically to a corruption context, higher externalities have sometimes been shown to reduce bribery (Barr and Serra, 2009).

H2: The share of officials who choose to maximize bribes is decreasing in the level of the externality.

On the other hand, the results on the effects of negative externalities on bribing behavior are mixed, and neither Abbink et al. (2002) nor Büchner et al. (2008) find any relationship

between the level of externality and individuals' decisions in a public procurement context. We thus present and test H2 as it has been pre-registered, while keeping in mind that the literature on which it is based is rather inconclusive.

Our third hypothesis is based on H1 and H2 and essentially captures an interaction thereof. If officials deliberately avoid information as a way of justifying more selfish choices (H1), and if they tend to make fewer selfish choices in the presence of a high externality (H2), then it follows that the incentive to avoid information is weaker when the externality is higher, ceteris paribus. It should be noted, though, that in the literature there is little evidence that the loss of the other party affects the propensity of individuals to exploit moral wiggle room. For instance, in Van der Weele (2014) it is shown that the size of others' potential benefit has little effect on willful ignorance.

H3: Officials will choose to reveal more information on firms' performance when the level of the externality is higher.

Procedures

We conducted the experiment in March 2021, with 20 sessions (five for each treatment), following a pre-registration that specified the hypotheses, procedures, sample size, and data analyses[5]. This led to a sample size of 120 participants per treatment, which includes 30 public officials and thus 150 observations for officials' decisions per treatment, given that they play five rounds of the game. We note, however, that we define each official as one independent observation in the statistical analysis, given that the five decisions by an official are not independent of each other.

The experiment was run at the HSE University in Russia and all participants were students of that institution. They were recruited through the manager of each educational program at HSE University (there are about 100 programs in four campuses in Moscow, St Petersburg, Perm, and Nizhny Novgorod), who was contacted and asked to send an e-mail to all students in his or her program. More than half of the managers agreed to do so. In this e-mail, we informed students about the study, their potential payoffs and asked them to fill out a google form with a convenient time slot. Nine hundred and sixty nine students filled out the form. Then we randomized these students across treatments, respecting their time preferences, and sent the invitation to a Zoom meeting to 45 students for each session (while only 24 were needed). Overbooking was necessary, since about 25 out of 45 students showed up for the meeting at a given time. We also had some "reserve" participants to ensure full sessions if fewer than 24 students appeared[6]. Each session lasted ∼70 min in total, including reading the instructions and answering questions. The average earnings were 500 RUR (about 6.5 US dollars) per participant, which exceeds the average hourly wage in Russia.

The experiment was computerized using oTree (Chen et al., 2016) and we deployed the game online using Heroku services. The study was conducted via Zoom: Each participant received an invitation and, as soon as all participants were connected, the experimenter distributed individual links to oTree and read the instructions aloud (sample instructions are provided in **Appendices A, B, C**). In case of questions, participants could ask the experimenter directly or via the Zoom chat. Participants were asked to disable videos in Zoom in order to ensure confidentiality. To make sure that all participants understood the instructions, a computer-based quiz with four comprehension questions was conducted before starting the experiment, with direct feedback and explanations in case of an incorrect answer. About 75% of participants answered all four questions correctly on the first try[7]. The same experimenter conducted all 20 sessions, for consistency and to ensure that differences across sessions and treatments could not be attributed to experimenter-specific characteristics.

A total of 480 students participated, 34.5% of whom were male, and with a mean age of 21 years. Each subject participated in only one session. Ninety five percentage of participants were Russian by nationality and the remaining 5% came from post-Soviet republics (Belarus, Kazakhstan, Moldova, Latvia, Uzbekistan, and Ukraine). Only four out of 480 participants were married and only one had children. Most participants were undergraduate students (90%), 9% had a Bachelor degree. Thirty five percentage defined themselves as Christians, 55% as atheists, the other 10% were other denominations. **Table A1** presents descriptive statistics, for the whole sample and separately by information avoidance and full information treatment[8].

RESULTS

We report experimental results on *corruption choices* and *information choices* of public officials. To define and measure corruption, we record whether officials award the government contract to the firm with the highest performance or maximize bribes instead[9]. Information choices refer to the question of whether officials choose to reveal the information regarding the performances of the two firms when given the option.

Since we have five decisions per official (one for each round in a session), the main variable used in the data analysis, which we will be calling *share*, is the number of cases in which the official in a group took a bribe-maximizing decision during the course of the five rounds of interaction, as a share of the total number

[5]Three sessions per day were conducted on March 13, 18, and 30; two sessions per day were conducted on March 16, 17, 19, 29, and 31; and one session was conducted on March 14. All the treatments were alternated between sessions.

[6]All reserve participants were HSE students as well, and most of them were students of one of the authors. In the end, the participation of a reserve participant was needed in only four cases.

[7]We have replicated all data analyses presented in Section 3 only for those participants who answered all four understanding questions correctly. All results remain unchanged.

[8]The sample is balanced for all control variables except for gender and education. Neither gender nor education was a focus variable in our study, hence we did not use block randomization along these dimensions. To account for the small observed imbalances, we control for gender and education (along with other control variables) in the regressions of **Table 3**. This does not affect any of the results and both variables are insignificant in the regressions.

[9]For ease of exposition we will be using the terms *corruption* and *bribe maximization* interchangeably, even though we acknowledge the fact that they do not perfectly overlap conceptually.

of *relevant* observations. Relevant refers to all cases in which an official faces a tension between bribe- and welfare maximization. This tension arises when the firm with the lower performance is the one that offers the higher bribe[10]. If, for instance, an official faces such a tension in four out of five rounds and chooses to maximize bribes in two of those cases and to maximize welfare in the other two cases, the value of *share* is 0.5 (= 2/4). In addition, choices in cases of ties in performance indicate only a weakly stronger concern (for bribe- or welfare maximization) by the official. We then employ two definitions: (i) *strict*—this includes cases when the official selects a firm with a higher bribe and lower performance, as a share of all cases when both bribes and performances are different (244 out of 600 officials' decisions in total), as well as situations when bribes are different and information on performances is avoided in the information avoidance treatments (25 out of 600 officials' decisions); (ii) *weak*—this definition includes all cases that fall under the definition of *strict*, adding situations when performances are equal (63 out of 600 officials' decisions). Situations with ties in bribes are excluded (65 out of 600 officials' decisions). This results in 112 independent observations on the variable *share* following the strict definition of bribe-maximizing behavior and 118 independent observations following the weak definition[11].

The departure point in our study is the question of whether public officials remain willfully ignorant in order to exploit a moral wiggle room. Hence, we begin the presentation of results by documenting information choices of public officials, i.e., whether they reveal the information on firms' performances when given the option to do so. We record the share of officials who reveal the information on firms' performance in the two information avoidance conditions (Treatments 2 and 4). Since we have five rounds and five observations per official per group, the outcome variable is the number of rounds (between 0 and 5) in which the official chose to reveal the performances of the two competing firms in a given group. The results are striking: In total, out of 300 decisions taken in total over the entire course of the interaction, officials decided to avoid the information about firms' performances only 25 times (8.3% of cases). This rate is much lower compared to previous studies that have endowed experimental participants with the opportunity to create and exploit a moral wiggle room[12]. **Figure 2A** below shows the share of officials who revealed the information in all five rounds, those who revealed it in at least one and at most four rounds, and those who never did. We observe that the overwhelming majority of officials never avoided the information.

Figures 2B,C allow us to test H3, by comparing the share of officials who choose to reveal information across treatments. In line with the graphical impression, a Mann-Whitney U test shows that the share of rounds (out of five) in which public officials chose to reveal information does not differ significantly between the low and the high externality conditions (0.93 vs. 0.91; $z = -0.36$; $p = 0.72$; $N = 60$).

> ***Result 1.*** *In the majority of cases, public officials always reveal information about firms' performances. This is true regardless of the level of externality, leading us to reject H3.*

Our test of H1 amounts to comparing the share of officials who maximize bribes across treatments, shown in **Table 2**. This comparison reveals that the difference between the *Full Information* and *Information Avoidance* treatments goes in the direction predicted by H1, but it is very small and insignificant, with bribe maximization rates of around 50% in both cases following the strict definition and roughly 60% following the weak definition (*strict*: $z = 0.71$, $p = 0.48$, $N = 112$; *weak*: $z = 0.35$, $p = 0.73$, $N = 118$; Mann-Whitney U tests). We also consider comparisons disaggregated by the level of externality: *share* does not differ by information treatment, under the low externality (*strict*: z = 0.72, $p = 0.48$, $N = 59$; *weak*: $z = 0.17$, $p = 0.86$, $N = 60$), or under the high externality (*strict*: $z = 0.23$, $p = 0.82$, $N = 53$; *weak*: $z = 0.26$, $p = 0.80$, $N = 58$; Mann-Whitney U tests).

> ***Result 2.*** *Introducing the possibility of information avoidance has no effect on the inclination of public officials to maximize bribes over welfare. Hence, we reject H1.*

The rejection of H1 comes as no surprise, in light of the fact that public officials very rarely choose to remain willfully ignorant about the competing firms' performances. Result 1 essentially says that information avoidance is not a relevant phenomenon in the context of a bribery experiment such as the one considered here, as public officials generally do not create moral wiggle room for themselves. In line with this pattern, Result 2 states that bribe-maximizing behavior is independent of the presence of opportunities for information avoidance. We note, however, that out of the 25 cases in which public officials chose to remain willfully blind, they selected the higher bribe in 22 cases. Hence, while moral wiggle room is very scarcely created, those officials who do create it almost always exploit it.

To test H2, we compare the share of officials who maximize bribes in the two treatments with low externality (T3, T4) vs. two treatments with high externality (T1, T2). Bribe maximizing behavior is slightly more widespread under the higher negative externality: 46% of officials choose to maximize bribes when they face such a possibility (56% following the weak definition) in the low externality treatments, while in high externality treatments this share increases to 57% (67% following the weak definition). However, this difference is not statistically significant (*strict*: $z = 1.38$; $p = 0.17$; $N = 112$; *weak*: $z = 1.85$; $p = 0.06$; $N = 118$). We also consider disaggregated comparisons and test the choices of officials for each of the two information treatments separately. No significant differences are found[13].

[10] When the official has information only on bribes, selecting the highest bribe without revealing performance is still classified as a bribe-maximizing decision.

[11] *Strict*: 26 observations in Treatment 1, 27 observations in Treatment 2, 29 observations in Treatment 3, 30 observations in Treatment 4. *Weak*: 29 observations in Treatments 1 and 2, 30 observations in Treatments 3 and 4.

[12] For instance, in the seminal study of Dana et al. (2007), 56% of participants chose to acquire costless information on relevant experimental parameters.

[13] *Full Information*, strict definition: z = 1.18; $p = 0.24$; $N = 55$; *Full Information*, weak definition: z = 1.15; $p = 0.25$; $N = 59$; *Information Avoidance*, strict

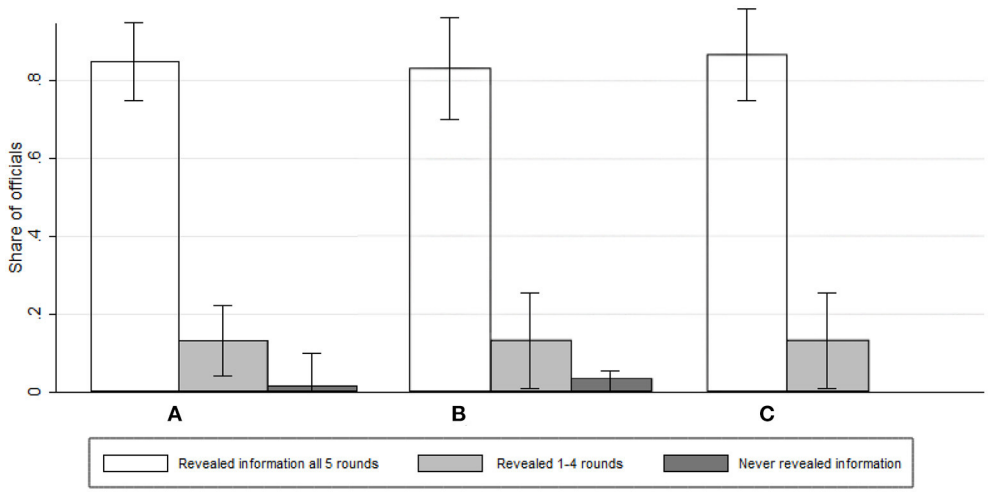


FIGURE 2 | Information avoidance by officials. **(A)** Pooled. **(B)** High externality. **(C)** Low externality.

TABLE 2 | Mean *share*, by treatment.

	Full information		Information avoidance		Overall	
	Strict	**Weak**	**Strict**	**Weak**	**Strict**	**Weak**
High externality	0.55 (0.42)	0.65 (0.38)	0.59 (0.38)	0.69 (0.31)	0.57 (0.39)	0.67 (0.34)
Low externality	0.43 (0.38)	0.55 (0.34)	0.50 (0.41)	0.56 (0.34)	0.46 (0.40)	0.56 (0.34)
Overall	0.49 (0.40)	0.60 (0.36)	0.54 (0.40)	0.63 (0.33)	0.51 (0.40)	0.61 (0.34)
N	55	59	57	59	112	118

Variable reports mean values of the variable share, as defined in the text. Standard deviations in parentheses. The number of observations (N) per condition is determined by the number of cases in which officials faced a tradeoff between bribe and citizen welfare maximization at least once.

Result 3. *When public officials face a tension between bribe- and welfare maximization, they maximize bribes in about half of such cases. The frequency of bribe-maximizing choices does not significantly vary by the negative externality imposed on the citizen, leading us to reject H2.*

In addition to the non-parametric analysis, we present in **Table 3** a series of regressions in order to offer additional insights into the various factors that affect the behavior of public officials. In one set of regressions (columns 1–4, on officials' corruption choices), the dependent variable is *share* as defined above. The independent variables are the treatment dummies; interactions between treatments; the difference in the bribes offered by the two firms, computed as the sum of absolute differences in bribes in cases where officials face a tradeoff between bribe and welfare maximization divided by the number of such cases (as a measure of the monetary incentive to maximize bribes); as well as control variables from the post-experimental survey. In another set of regressions (columns 5–6, on officials' information choices), the dependent variable is the share of cases (out of 5) in which an official revealed the information on firms' performance. These regressions include the same set of independent variables as in the first four columns, except for the *Information Avoidance* treatment dummy (since information choices are only available in that treatment). All regressions are run using Ordinary Least Squares, with standard errors clustered at the session level.

The regression analysis fully confirms Results 1–3. The level of externality affects neither the willingness of public officials to reveal information nor their propensity to maximize bribes. The option to remain ignorant about firms' performances (*Information Avoidance*) does not affect the corruption choices of public officials, and it does not interact with the externality level. As expected, a larger absolute difference in bribes—corresponding to a stronger motive for bribe maximization—is a significant predictor of an official's choice of the winning firm using the strict definition.

As a check of robustness, we estimate a set of regressions where the dependent variable is an individual official's round-by-round decision and use a random effects model to account for the interdependence of these decisions. Estimation results are given in **Table A3**. All previous results are fully confirmed: Neither the size of the externality nor the option to reveal information affects officials' willingness to maximize bribes. We observe that the absolute difference in the size of bribes (measured in a more accurate way case-by-case, compared to the sum of absolute

definition: $z = 0.73$; $p = 0.47$; $N = 57$; *Information Avoidance*, weak definition: $z = 1.37$; $p = 0.17$; $N = 59$.

TABLE 3 | Regression analysis on public officials' choices.

	Officials' corruption choice: *share*				Officials' information choice	
	Strict definition		Weak definition			
High externality	0.130	0.123	0.106	0.093	−0.020	−0.018
	(0.122)	(0.113)	(0.109)	(0.096)	(0.059)	(0.066)
Information avoidance	0.086	0.140	0.016	0.054		
	(0.123)	(0.139)	(0.098)	(0.103)		
High externality × Information avoidance	−0.056	−0.036	0.019	0.050		
	(0.155)	(0.158)	(0.138)	(0.130)		
Difference in bribes	0.032**	0.035**	0.023	0.027	−0.012	−0.004
	(0.017)	(0.020)	(0.018)	(0.022)	(0.018)	(0.022)
Control variables	No	Yes	No	Yes	No	Yes
Number of observations	112	112	118	118	60	60

*The number of observations in columns 1–4 is smaller than 120 due to the way the variable share is constructed: 8 and 2 officials (using the strict and weak definition, respectively) never faced a tradeoff between bribe and citizen welfare maximization. Standard errors, clustered at the session level, are in parentheses. ** denotes statistical significance at the 5% level.*

differences in bribes used in **Table 3**) gains both in size and statistical significance. This variable becomes significant for the officials' information choice as well, with a negative coefficient. This confirms that the difference in the size of bribes is a strong motive for corruption.

Although firms are not the focus of our study, we briefly discuss some information about their behavior. Performances in the task display sufficient variation and vary from 0 to 10 matrices, with a mean of 4.59, standard deviation of 2.15, and median of 4 (see **Figure A2**). Similarly, bribes are offered in the entire possible range from 0 to 9 ECU, with a mean of 5.25, standard deviation of 2.61 and median of 6 (see **Table A2** and **Figure A3**). Comparing across treatments, we confirm that randomization has been successful, since neither bribes nor performances differ significantly by treatment, either in the information or in the externality dimension[14]. We document a negative relationship between firm performance and bribes: The estimation of a linear regression model with bribe as the dependent variable yields a coefficient of −0.12 for performance (significant at the 5% level). This suggests that, on average, an increase of 8 points in performance leads to a one-unit reduction in the bribe.

We also compare the beliefs of firms and citizens with the actual behavior of officials in order to reveal how successful they are in predicting the incidence of corruption. While officials were making their choices in Stage 3, firms and citizens were asked the following question: *"Out of the 7 officials*[15] *in this session, how many do you think will choose a firm with a higher bribe instead of a firm with a higher performance, if they face such a tradeoff?*." Interestingly, the difference between officials' actual bribe-maximizing behavior and firms' and citizens' expectations is pronounced. On average, across treatments, the share of officials who maximize bribes is 0.54, while firms and citizens expect it to be 0.85 on average. This pattern also holds if we make comparisons separately by treatments, see **Figure 3**. We find that this perceived frequency does not differ between treatments (0.83 in Full Information vs. 0.86 in Information Avoidance, $p = 0.58$, Mann-Whitney U test). The very high reported beliefs by firms and citizens point toward the absence of a descriptive norm against bribe taking, and it can help explain why officials do not create and exploit a moral wiggle room in our experiment. We return to this point in the discussion section.

DISCUSSION AND CONCLUSION

Motivated by the legal and moral implications of willful blindness in settings of corruption, this study has examined the question of whether the widely documented phenomenon of information avoidance in economic experiments can be detected in a public procurement game. Our data deliver a negative answer: The majority (85%) of decision makers in the role of public officials obtain all relevant information in every round of the game when given the option to do so. Given this pattern, it is no surprise that we also document no differences in bribe taking behavior between the treatments with and without information avoidance opportunities. In addition, our study contributes to the open question on the role of negative third-party externalities on corruption: We find that bribe taking as a means of maximizing own payoff over citizen (and total) welfare is independent of the size of the negative externality.

The rejection of H1 and H3 is a consequence of the fact that public officials do not exploit opportunities for willful ignorance. The rejection of H2 is somewhat more surprising, although it fits well within the context of mixed findings in previous literature. At the same time, the rejection of all hypotheses may also be related to the small sample size of our study and to the

[14] Bribes: $z = 0.59$, $p = 0.56$ comparing *Full Information* vs. *Information Avoidance*, and $z = -0.91$, $p = 0.36$, *High Externality* vs. *Low Externality*. Performances: $z = -1.46$, $p = 0.14$ comparing *Full Information* vs. *Information Avoidance*, and $z = -1.42$, $p = 0.16$, *High Externality* vs. *Low Externality*. All tests reported here are Mann-Whitney U tests, treating average bribes or performances over all rounds within a group as one independent observation.

[15] We had 6 officials in each session but we asked about 7 in this question, due to an error. To compensate for this error, we computed the share for beliefs dividing the reported number by 7. This share can be compared against the share of corrupt choices by officials, dividing the number of bribe maximizing choices by the number of cases when officials faced a tradeoff.

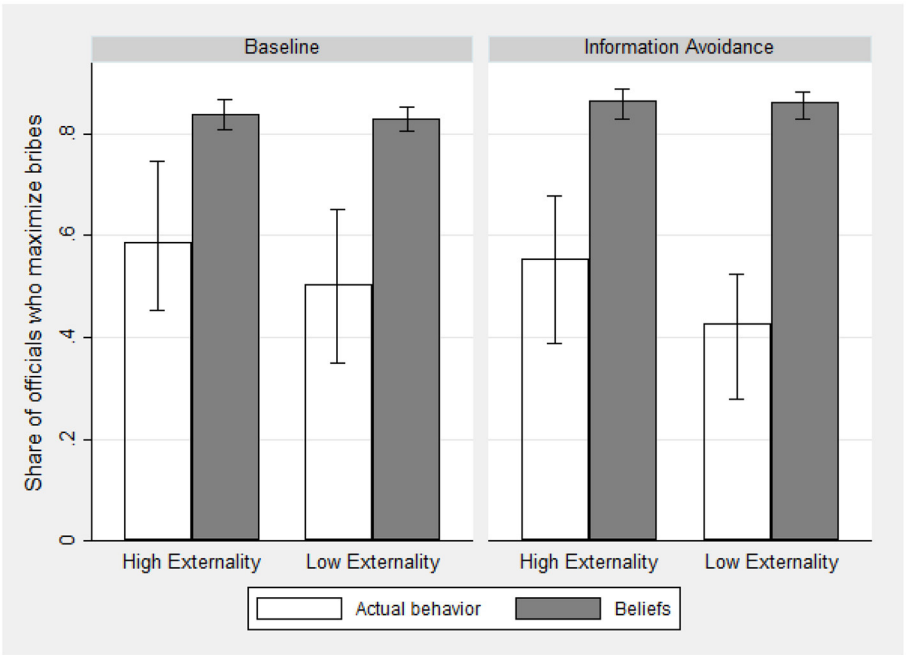


FIGURE 3 | Bribe maximization: Beliefs vs. actual behavior, by treatment.

way we define independent observations. While we have 590 (560) relevant observations on the behavior of officials following the weak (strict) definition of bribe maximizing behavior, our conservative testing procedure as defined in the pre-registration is based on only one-fifth of these figures. Minimum detectable effect sizes using this conservative procedure are as follows (referring here only to the weak definition in the interest of brevity): 0.19 for H1, 0.18 for H2, and 0.15 for H3, hence about one half of the observed standard deviations. This means that we cannot rule out the possibility that differences do exist across treatments in one or the other dimension, but they are smaller than the above figures and therefore not detectable in our study. This important caveat must be kept in mind and calls for replications of our results and additional evidence on the topic.

Why do participants in our experiment not avoid information? While we cannot give a definitive answer to this important question based on our dataset, we offer some thoughts on it. First, as noted in the introduction, ours is the first experimental study on information avoidance in a corruption setting. The fact that individuals have often been shown to avoid information in a self-serving manner does not mean that they will do so in every context. Corruption in the form of bribe payments (as implemented in our experiment) is a sensitive topic, widely discussed in politics and the media, and with far-reaching implications for society. When asked to place themselves in this situation (through the structure of the game and the loaded and framed instructions), experimental participants may find it important to have all available information at their disposal before they decide on a course of action. Their desire to make an informed decision for themselves and for the three other members of their micro-society may weigh in more than the motivation to maximize their own income without running the danger of compromising their (self-)image[16].

Another possibility is that these findings are culture-specific. Our experiment was conducted in Russia, a country where corruption is very widespread. For instance, Mironov and Zhuravskaya (2016) reveals corruption in Russia by measuring the amount of cash channeled illegally out of firms around the time of regional elections and relating it to the probability that the firms obtained procurement contracts from the government. Zhuravleva (2015, 2021) shows that Russian households with workers in the public sector receive lower earnings than households with members employed in the private sector but enjoy the same level of consumption, and justifies this unexplained consumption-income gap by unreported income in the public sector. In 2020, Russia ranked only 129th out of 180 countries worldwide in the Corruption Perceptions Index published by Transparency International. In the 7th wave of the World Values Survey (WVS, see Haerpfer et al., 2020), respondents in Russia perceive corruption as very pervasive and the likelihood of being held accountable for corrupt practices as low. This question is also available for our sample. As it turns out, our participants are actually even more pessimistic than WVS respondents: In the question *"How would you place your views*

[16]It must be noted that our game differs from most of the previous literature in one additional dimension (besides considering a different context and game): participants interact, and thus officials decide whether to remain ignorant, over five rounds. It might be harder to uphold a positive self-image when (knowingly) remaining ignorant for several times. If so, this feature of the design can help explain the very low incidence of information avoidance.

on corruption in your country," mean responses are 7.35 among WVS respondents and 8.95 in our sample (on a scale from 1 to 10, with higher values indicating higher perceived corruption). For comparison, the mean response in Germany (available during the same WVS wave and ranked at ninth place in the Corruption Perceptions Index) is 5.41.

In this context of high perceived and actual incidence of corruption, the moral costs of engaging in it are most likely substantially lower than in countries where bureaucrats are seen as more honest (see Balafoutas, 2011, for a theoretical model on public beliefs about corruption and how they shape the psychological costs for corrupt bureaucrats). Indeed, we have already shown in the previous section that the large majority of firms and citizens expect officials to maximize bribes when facing a tradeoff between bribes and citizen welfare. As a result, moral wiggle room is not as valuable, and much less often exploited. In terms of policy-related insights, this suggests that claims of ignorance often encountered in cases of corruption are quite unlikely to be true, and may be more often than not used as cheap talk or as an excuse by corrupt public officials. This would imply that such claims must be treated with particular skepticism by investigating authorities. Following these considerations, we believe that the replication of our study in countries with a lower incidence of corruption and strong anti-corruption norms would be a very interesting endeavor.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Board for Ethical Questions in Science of the University of Innsbruck. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

We gratefully acknowledge financial support from the University of Innsbruck, the basic research program of the HSE University, and the Austrian Science Fund (FWF) through special research area grant SFB F63. This study has received ethical approval from the Institutional Review Board of the University of Innsbruck (no. 08/2021). The pre-registration for the experiment is available under the following link: https://aspredicted.org/blind.php?x=h2au6m.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.701294/full#supplementary-material

REFERENCES

Abbink, K., Dasgupta, U., Gangadharan, L., and Jain, T. (2014). Letting the briber go free: an experiment on mitigating harassment bribes. *J. Public Econ.* 111, 17–28. doi: 10.1016/j.jpubeco.2013.12.012

Abbink, K., and Hennig-Schmidt, H. (2006). Neutral versus loaded instructions in a bribery experiment. *Exp. Econ.* 9, 103–121. doi: 10.1007/s10683-006-5385-z

Abbink, K., Irlenbusch, B., and Renner, E. (2002). An experimental bribery game. *J. Law Econ. Organ.* 18, 428–454. doi: 10.1093/jleo/18.2.428

Abbink, K., Serra, D. and Wantchekon, L (eds.) (2012). "Anticorruption policies: lessons from the lab," in *New Advances in Experimental Research on Corruption*. Vol. 15 Bingly: Emerald Group Publishing Limited.

Ajzenman, N. (2021). The power of example: corruption spurs corruption. *Am. Econ. J. Appl. Econ.* 13, 230–257. doi: 10.1257/app.20180612

Alatas, V., Cameron, L., Chaudhuri, A., Erkal, N., and Gangadharan, L. (2009). Gender, culture, and corruption: Insights from an experimental analysis. *South. Econ. J.* 75, 663–680.

Armantier, O., and Boly, A. (2011). A controlled field experiment on corruption. *Eur. Econ. Rev.* 55, 1072–1082. doi: 10.1016/j.euroecorev.2011.04.007

Azfar, O., and Nelson, W. R. (2007). Transparency, wages, and the separation of powers: an experimental analysis of corruption. *Public Choice* 130, 471–493. doi: 10.1007/s11127-006-9101-5

Balafoutas, L. (2011). Public beliefs and corruption in a repeated psychological game. *J. Econ. Behav. Organ.* 78, 51–59. doi: 10.1016/j.jebo.2010.12.007

Barr, A., and Serra, D. (2009). The effects of externalities and framing on bribery in a petty corruption experiment. *Exp. Econ.* 12, 488–503. doi: 10.1007/s10683-009-9225-9

Barr, A., and Serra, D. (2010). Corruption and culture: an experimental analysis. *J. Public Econ.* 94, 862–869. doi: 10.1016/j.jpubeco.2010.07.006

Becker, G. S. (1968). *Crime and Punishment: An Economic Approach*. The Economic Dimensions of crime. London: Palgrave Macmillan, 13–68.

Büchner, S., Freytag, A., González, L. G., and Güth, W. (2008). Bribery and public procurement: an experimental study. *Public Choice* 137, 103–117. doi: 10.1007/s11127-008-9315-9

Cameron, L., Chaudhuri, A., Erkal, N., and Gangadharan, L. (2009). Propensities to engage in and punish corrupt behavior: experimental evidence from Australia, India, Indonesia and Singapore. *J. Public Econ.* 93, 843–851. doi: 10.1016/j.jpubeco.2009.03.004

Charness, G., and Rabin, M. (2002). Understanding social preferences with simple tests. *Q. J. Econ.* 117, 817–869. doi: 10.1162/003355302760193904

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree - an open-source platform for laboratory, online, and field experiments. *J. Behav. Exp. Finance* 9, 88–97. doi: 10.1016/j.jbef.2015.12.001

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econ. Theory* 33, 67–80. doi: 10.1007/s00199-006-0153-z

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: avoiding altruism by distorting beliefs about others' altruism. *Am. Econ. Rev.* 105, 3416–3442. doi: 10.1257/aer.20141409

Di Zheng, J., Schram, A., and Dogan, G. (2020). Friend or foe? Social ties in bribery and corruption. *Exp. Econ.* 1–29. doi: 10.1007/s10683-020-09683-7

Engel, C. (2011). Dictator games: a meta study. *Exp. Econ.* 14, 583–610. doi: 10.1007/s10683-011-9283-7

Felgendreher, S. (2018). Do consumers choose to stay ignorant? The role of information in the purchase of ethically certified products. *Working Papers in Economics*. University of Gothenburg.

García-Gallego, A., Georgantzis, N., Jaber-López, T., and Michailidou, G. (2020). Audience effects and other-regarding preferences against

corruption: experimental evidence. *J. Econ. Behav. Organ*. 180, 159–173. doi: 10.1016/j.jebo.2020.09.025

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Manage. Sci*. 60, 2659–2665. doi: 10.1287/mnsc.2014.1989

Grossman, Z., and Van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *J. Eur. Econ. Assoc*. 15, 173–217. doi: 10.1093/jeea/jvw001

Guerra, A., and Zhuravleva, T. (2020). *Do Women Always Behave as Corruption Cleaners*? Available online at: https://ssrn.com/abstract=3601696 (accessed August 2, 2021).

Guerra, A., and Zhuravleva, T. (2021). Do bystanders react to bribery?. *J. Econ. Behav. Organ*. 185, 442–462. doi: 10.1016/j.jebo.2021.03.008

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J. M., et al. (eds.). (2020). *World Values Survey: Round Seven – Country-Pooled Datafile*. Madrid; Vienna: JD Systems Institute and WVSA Secretariat.

Jaber-López, T., García-Gallego, A., Perakakis, P., and Georgantzis, N. (2014). Physiological and behavioral patterns of corruption. *Front. Behav. Neurosci*. 8:434. doi: 10.3389/fnbeh.2014.00434

Kajackaite, A. (2015). If I close my eyes, nobody will get hurt: the effect of ignorance on performance in a real-effort experiment. *J. Econ. Behav. Organ*. 116, 518–524. doi: 10.1016/j.jebo.2015.05.020

Khadjavi, M., Lange, A., and Nicklisch, A. (2017). How transparency may corrupt – experimental evidence from asymmetric public goods games. *J. Econ. Behav. Organ*. 142, 468–481. doi: 10.1016/j.jebo.2017.07.035

Klitgaard, R. (1988). *Controlling Corruption*. Oakland, CA: University of California press.

Konow, J. (2000). Fair shares: accountability and cognitive dissonance in allocation decisions. *Am. Econ. Rev*. 90, 1072–1091. doi: 10.1257/aer.90.4.1072

Kriss, P. H., Weber, R. A., and Xiao, E. (2016). Turning a blind eye, but not the other cheek: on the robustness of costly punishment. *J. Econ. Behav. Organ*. 128, 159–177. doi: 10.1016/j.jebo.2016.05.017

Krupka, E. L., and Weber, R. A. (2013). Identifying social norms using coordination games: why does dictator game sharing vary?. *J. Eur. Econ. Assoc*. 11, 495–524. doi: 10.1111/jeea.12006

Mironov, M., and Zhuravskaya, E. (2016). Corruption in procurement and the political cycle in tunneling: evidence from financial transactions data. *Am. Econ. J. Econ. Policy* 8, 287–321. doi: 10.1257/pol.20140188

Parra, D., Muñoz-Herrera, M., and Palacio, L. (2019). The limits of transparency as a means of reducing corruption (No. SP II 2019-401). *WZB Discussion Paper*.

Regner, T. (2018). Reciprocity under moral wiggle room: is it a preference or a constraint?. *Exp. Econ*. 21, 779–792. doi: 10.1007/s10683-017-9551-2

Ryvkin, D., Serra, D., and Tremewan, J. (2017). I paid a bribe: an experiment on information sharing and extortionary corruption. *Eur. Econ. Rev*. 94, 1–22. doi: 10.1016/j.euroecorev.2017.02.003

Salmon, T. C., and Serra, D. (2017). Corruption, social judgment and culture: an experiment. *J. Econ. Behav. Organ*. 142, 64–78. doi: 10.1016/j.jebo.2017.06.004

Schram, A., Zheng, J. D., and Zhuravleva, T. (2019). Contagious corruption: cross-country comparisons. *EUI Working Paper MWP*. 2019. No. 06. doi: 10.2139/ssrn.3487972

Serra, D. (2012). Combining top-down and bottom-up accountability: evidence from a bribery experiment. *J. Law Econ. Org*. 28, 569–587. doi: 10.1093/jleo/ewr010

Serra-Garcia, M., and Szech, N. (2019). The (in) elasticity of moral ignorance. *CESIFO Working Paper*.

Stüber, R. (2019). The benefit of the doubt: willful ignorance and altruistic punishment. *Exp. Econ*. 23, 848–872. doi: 10.1007/s10683-019-09633-y

Svensson, J. (2005). Eight questions about corruption. *J. Econ. Perspect*. 19, 19–42. doi: 10.1257/089533005774357860

Van der Weele, J. J. (2014). Inconvenient truths: determinants of strategic ignorance in moral *dilemmas*. doi: 10.2139/ssrn.2247288

Weber, M., and Schram, A. (2017). The non-equivalence of labour market taxes: a real-effort experiment. *Econ. J*. 127, 2187–2215. doi: 10.1111/ecoj.12365

Zhuravleva, T. (2015). Does the Russian government pay a "fair" wage: review of studies. *Voprosy Econ*. 11, 62–85. doi: 10.32609/0042-8736-2015-11-62-85

Zhuravleva, T. (2021). Is the difference in consumption and income an indication of petty corruption?. *J. New Econ. Assoc*. 49, 115–136. doi: 10.31737/2221-2264-2021-49-1-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 August 2021
doi: 10.3389/fpsyg.2021.684735

Investigating Dishonesty-Does Context Matter?

*Aline Waeber**

Institute of Insurance Economics, University of St. Gallen, St. Gallen, Switzerland

This paper introduces frame-specific randomization devices to vary the situational context of an online lying experiment. Participants are asked to report outcomes of random draws from two different sources of uncertainty—decimals of the value of a stock index or a neutrally framed random number generator. The findings show that the frame-specific randomization device is not prone to the social norm effects documented in the literature. Because different environments can evoke different norms, I replicate the experiment in the more constrained setting of a traditional physical laboratory revealing no systematic differences in behavior. Furthermore, I am not able to show that participants who take longer to report are more honest and this is specific to the physical laboratory environment. Finally, the findings reveal gender differences in honesty depending on the environment—males are more honest when they participate in the laboratory as opposed to online.

Keywords: lying, honesty, moral behavior, framing, context-dependence

OPEN ACCESS

Edited by:
Nora Szech,
Karlsruhe Institute of Technology (KIT), Germany

Reviewed by:
Roland Pfister,
Julius Maximilian University of Würzburg, Germany
Valerio Capraro,
Middlesex University, United Kingdom

***Correspondence:**
Aline Waeber
aline.waeber@unisg.ch

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: *23 March 2021*
Accepted: *15 June 2021*
Published: *23 August 2021*

Citation:
Waeber A (2021) Investigating Dishonesty-Does Context Matter? Front. Psychol. 12:684735. doi: 10.3389/fpsyg.2021.684735

1. INTRODUCTION

This research mainly centers around two different research questions addressed in two experimental studies. In the first study, I estimate whether financial market saliency triggers dishonest behavior in an online experiment. More specifically, the experimental design allows to test whether participants are more honest when they are introduced to a financial market context as opposed to a neutral context. I use frame-specific randomization devices to vary the situational context of the game (i.e., stock market or neutral context). Although most of standard economic theory implicitly assumes that people act as if preferences are stable, there is abundant evidence (Tversky and Kahneman, 1981; Dufwenberg et al., 2011) showing that subtle differences in the way a situation is framed can cause changes in preferences[1]. To develop a deeper understanding of the causes of potential differences in behavior when the source of uncertainty is a stock market index, I elicit individual beliefs about dishonest behavior of others. In the second study, I aim to find out whether the environment (i.e., physical laboratory or online) has an effect on dishonest behavior. In both studies, the dependent variable is the reported draw defined on the interval between 0 and 9. I additionally capture the variation in behavior that is induced not only by the previously mentioned independent variables (i.e., financial market setting/environment), but also decision times.

Previous research suggests that different environments evoke different norms of behavior. The stock market environment may be linked to contexts in which competitive or exploitative norms prevail (Liberman et al., 2004; Cohn et al., 2014). This means that the stock market context may trigger a stronger desire to be greedy. Participants in the stock market context could thus feel as it is easier to justify dishonest behavior to increase payoffs if he/she believes the norm in that specific

[1] Capraro and Perc (2021) provide an exhaustive review of moral and norm framing effects.

environment is to make as much money as possible. In contrast, the neutral setting should not evoke any strong connotations (Cohn et al., 2014). I replicate the experiment in the more constrained setting of the traditional physical laboratory. It is possible that participants feel more socially distant from others in an online environment[2]. This might reduce participants' need to adhere to social norms of behavior. Another source of variation in dishonest behavior (though endogenous) is the time it takes to make a decision. I thus explore differences in decision times depending on the environment (i.e., laboratory or online). A recent meta-analysis finds that honesty is deliberative (Köbis et al., 2019). I thus expect decision times and dishonesty to be negatively correlated.

The results from the online experiment show no significant differences in dishonest behavior between the two enviroments—stock market and neutral. This indicates that this specific source of uncertainty is not prone to the social norm effects documented in the literature. The findings confirm previous studies that do not find significant differences in a student sample between a financial and a neutral context (Cappelen et al., 2013; Huber and Huber, 2020). More specifically, the frame-specific device does not shift participants' beliefs about the prevailing honesty norm. Furthermore, there are no significant differences in dishonest behavior conditional on the environment (i.e., physical lab or online). Looking at decision times, I find that participants who take longer to report are more honest. However, this is only true for subjects in the physical laboratory—a possible sign of self-reflection of self-image violating behavior. Finally, the results suggest gender differences in honest behavior depending on the environment. Even the slightest cues of being observed seem to affect male but not female reporting behavior.

Related Literature. The study first and foremost relates to the literature on framing effects in social preference games. Framing generally refers to the observation that a decision problem can be presented in different ways, for example, in positive or negative connotations or "frames"[3]. Of particular importance are studies contrasting conditions in which the description of the relevant task evokes norms related to competitive vs. cooperative norms. Earlier work reveals that people cooperate more in a prisoner's dilemma when it is called the *Community Game* than when it is called the *Wall Street Game* (Kay and Ross, 2003; Liberman et al., 2004; Ellingsen et al., 2012). However, social framing effects in the prisoner's dilemma vanish when the game is played sequentially. This suggests that social cues primarily work by changing participants' beliefs about other people in the interaction rather than participants' preferences (Fehr and Schmidt, 2006; Ellingsen et al., 2012). Dreber et al. (2013) investigate whether social framing effects are also present in dictator games. They find that dictators are not sensitive to different frames. Contrary to this, Chang et al. (2019) do find an effect in a politically framed dictator game. They vary whether participants are shown neutrally framed or tax-framed dictator games. The aim is to render a U.S. political identity salient (i.e., Democrat or Republican) and to evoke the associated norm for that identity. They show that framing causes participants to apply different norms to the situation which affects their behavior. Andreoni (1995) finds significant differences in contributions when a public goods game is framed as *giving to a public good* as opposed to *taking from a public good*. In Krupka and Weber (2013), when the dictator game is framed as taking from another's endowment (i.e., a bully game) as opposed to giving away a portion of one's own endowment (i.e., a standard dictator game), bullies claimed less than did dictators. Thus, these findings indicate that changes in norms induce changes in behavior in otherwise identical economic games. Similarly, Capraro and Vanzo (2019) show that the words used to describe the available actions can affect people's decisions in extreme dictator games. However, in their study, the *take* frame does not give rise to a rate of pro-sociality significantly higher than the *give* frame.

Regarding the effect of financial market saliency on dishonest behavior, the evidence is mixed. Research from priming studies finds that simply priming subjects with the concept of money evokes more selfish behavior (Vohs et al., 2008; Vohs, 2015). In a subsequent study, Cohn et al. (2014) find that when financial professionals are reminded of their professional identity, they become more dishonest than their colleagues who are asked to think about leisure activities. The authors argue that "the prevailing business culture in the banking industry weakens and undermines the honesty norm." However, more recent studies challenge these findings. For example, Rahwan et al. (2019) failed to replicate the results of more dishonest behavior among bankers across several populations. Rahwan et al. (2019) argue that differences in honesty could be attributed to heterogeneity in national banking norms, especially heterogeneity in the general population's relative expectation of bankers.[4]. Other studies point out that using a neutral prime for the control group (instead of *leisure activities*) might change results (Stöckl, 2015; Vranka and Houdek, 2015). Framing their experiment in a financial context, Huber and Huber (2020) show that financial professionals act more honestly in a financial context as opposed to a neutral context. However, this difference in behavior cannot be confirmed within a sample of students. The authors identify reputational concerns as one of the drivers of financial professionals' behavior. Similarly, Cappelen et al. (2013) find that students do not lie significantly less when they are in a market context. The above-mentioned studies vary the name attached to a game, while I vary the situational context of the game using a frame-specific device.

My work is furthermore related to the literature focusing on the psychological costs of dishonesty. The recent experimental literature has shown that individuals are often willing to forego

[2]Earlier studies indicate that a lower degree of social distance between parties increases prosocial behavior (Hoffman et al., 1996; Bohnet and Frey, 1999; Charness et al., 2007; Charness and Gneezy, 2008).

[3]The framing effect concept was coined by Tversky and Kahneman (1981). This study focuses on framing effects related to the labeling of the situational context of the experiment.

[4]In the jurisdiction of the original study (Cohn et al., 2014), the banking industry has a very bad reputation at the time of the experiment. They are perceived to be less honest than doctors, to be less honest than the general population and to behave about as dishonestly as prison inmates. This does not apply to other jurisdictions in Rahwan et al. (2019).

financial benefits to behave honestly (Gneezy, 2005; Mazar et al., 2008; Erat and Gneezy, 2012; Cappelen et al., 2013; Fischbacher and Föllmi-Heusi, 2013; Abeler et al., 2014; Gneezy et al., 2018). The literature on intrinsic costs of lying suggests that people have internal standards for honesty which influence their self-concept (see Mazar et al., 2008). These internal standards are shaped by the norms and values of a society (Henrich et al., 2001). People thus do not only consider the expected monetary gains from lying, the probability of being caught, and the potential punishment but also how the act of lying might make them perceive themselves. This means that people do lie when it pays, but only to the extent that their perception of themselves as an honest person is not violated. Analyzing dishonesty in low stake scenarios, Barron (2019) shows that a substantial fraction of subjects lie downwards (i.e., giving up money to signal honesty). These subjects care about appearing good in more lucrative interactions[5].

Fraud and unethical behavior are recurring issues in markets, which are costly for all market participants. Dishonesty poses a severe negative externality to markets, which can ultimately cause market failure. If everyone behaves honestly, eveyone benefits because high costs arise in doing business otherwise. An example of everyday deception is insurance fraud. The FBI estimates the total cost of insurance fraud in the U.S. (non-health insurance) to be more than USD 40 billion per year, which increases premiums for the average U.S. family between USD 400 and USD 700 annually (FBI, 2020). Similar acts of dishonesty can be observed in tax reporting. A recent Internal Revenue Service (IRS) study estimates the tax gap (i.e., the difference between what the IRS estimates taxpayers should pay and what they actually pay) at USD 441 billion per year for the 2011–2013 timeframe (Internal Revenue Service, 2019).

2. METHODS-EXPERIMENTAL DESIGN

I present a novel experimental design to measure dishonest behavior in an online setting. The experimental task is a one-shot individual decision-making situation. I rely on a between-subjects design in which the treatments are distinguished by how the particular decision situation is framed (i.e., random number, financial market).

The experiment follows the fundamental idea of other experimental setups to infer dishonest behavior (e.g., Fischbacher and Föllmi-Heusi, 2013) by asking participants to report a randomly generated number[6]. I collect reports of unobserved payout-determining random draws from two novel non-physical and verifiably random sources of uncertainty. I let participants report the outcomes of random draws from either decimals of a stock index price (T_{FM}) or a random number generator (T_{RN}). As mentioned earlier, participants took part in the study not in the laboratory but at home in the main part of this study.

2.1. Treatment Variations

I implement two treatments in a between-subjects design. Under both conditions, participants report outcomes using an online form. In treatment 1, the payoff is determined by the second decimal place of either the Swiss Market Index (SMI) or the DAX Performance Index (DAX) at a particular point in time; the reported value equals the payoff in CHF. Participants in the experiment are asked to lookup the value of their respective index of choice on a Google Widget, showing either the SMI or the DAX (see **Figure 1** for an example of the SMI). In treatment 2, I measure dishonest behavior using a neutrally framed randomization device. The payoff is determined by looking up a random number (between zero and nine) on a Google Widget—the reported value equals the payoff in CHF. Because the payoff participants receive for participation depends on the reported value, there is a clear incentive to report higher numbers. I emphasize that I do not know about participants' choice of index/random number, lending credibility to the unobservability of the source of uncertainty, for which it is important to avoid reputation and strategic concerns.[7] As opposed to previous studies (e.g., Cohn et al., 2014), a subject's payoff is not dependent on others' choices[8]. By the design, I cannot detect dishonesty at the individual level, but, because I know the actual distribution of values[9] I can infer dishonesty for different subpopulations. The full set of experimental instructions can be found in the **Appendix**[10].

The non-strategic nature of the experiment makes it rather easy to establish different environments in which I can hold constant important features of the decision task, while varying context in a way that can influence norms.

Subsequent to the main experiment, I examine the role of dispositional greed in explaining potential differences in dishonest behavior between the two groups. I focus on the Dispositional Greed Scale DGS (Seuntjens et al., 2015) to measure individual differences in people's propensity to be greedy. All items (e.g., "As soon as I have acquired something, I start to think about the next thing I want.") were rated using a five-point Likert-scale, ranging from strongly disagree (1) to strongly agree (5). It

[5]Pfister et al. (2019) find that dishonest behavior does not only entail aforementioned intrinsic costs but that they also come with cognitive costs that emerge right before and while a person deliberately violates a rule.

[6]Fischbacher and Föllmi-Heusi (2013) had participants roll a die *in private* and report their roll. Participants were paid CHF 1, 2, 3, 4, and 5 for the corresponding outcome and CHF 0 for an outcome of 6.

[7]To test whether fear of detection plays a role, I elicit subjects' risk attitudes to test whether there is a relationship between risk aversion and dishonest behavior. If the fear of detection is a relevant issue in the design, I expect that more risk averse subjects are less dishonest. I find that subjects' risk aversion does not significantly decrease reported values ($p = 0.519$). It can thus be concluded that punishment concerns do not play a role in the experimental design.

[8]As pointed out by Stöckl (2015), due to the competitive aspect of the experiment in Cohn et al. (2014), subjects actually play a strategic game.

[9]The distribution of second decimals in the range of [0,9] of the DAX and SMI stock index is not only known approximately but exactly, as we look up the data in 5-min intervals at a later point in time in the respective Google Widgets. It is important to emphasize that we are not able to observe the same data as participants during the time of the experiment. I show in **Figure A1** of the Appendix that the values are indeed equally distributed.

[10]On a cautionary note, I should mention that while the two different randomizer layouts reduce participants' beliefs of experimenter-induced influence of outcomes, it should not be ignored that such differences in the specific layouts may give rise to systematic variation between conditions.

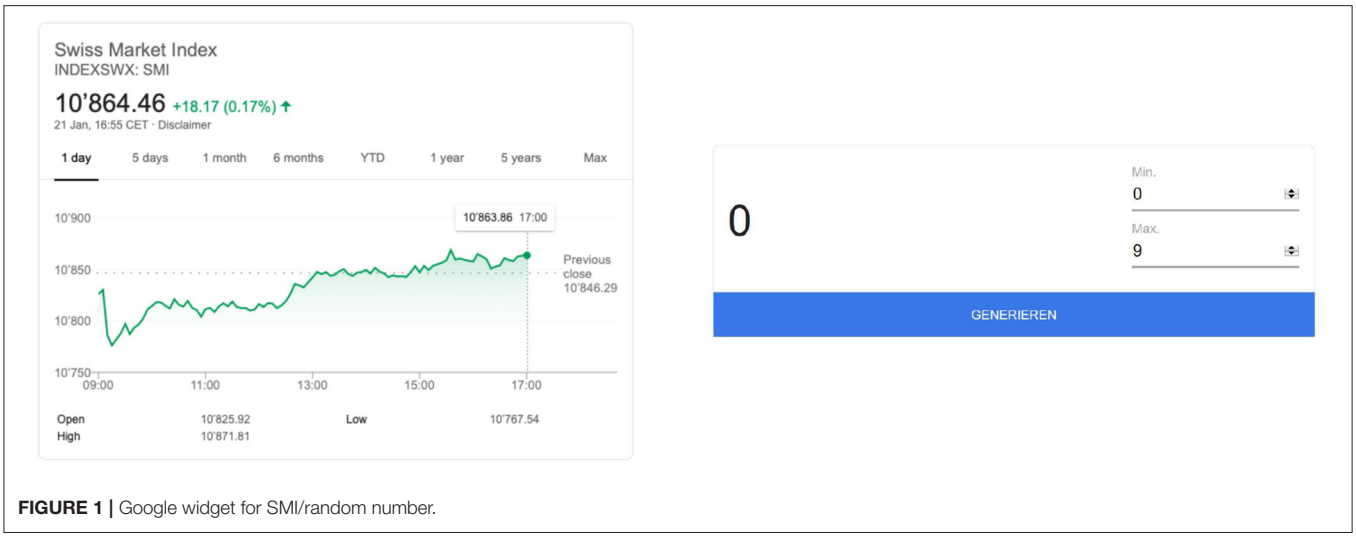


FIGURE 1 | Google widget for SMI/random number.

has been shown that greedy people take more and contribute less in economic games (Seuntjens et al., 2015) and are more willing to accept bribes and engage in unethical behavior (Seuntjens et al., 2019). If financial markets are linked to norms that encourage greedy behavior, I expect that experimental measures of honesty will differ in these two treatments. I additionally elicit subject's risk attitudes using the survey questions developed by Dohmen et al. (2011).

2.2. Procedures

I recruited participants from the participant pool of the behavioral lab at the University of St.Gallen. This allows us to attentively control the pool of participants, which mitigates experimenter control problems. Additionally, the design is conceptually rather simple, which should reduce concerns about participants' mental performance being worse in the online setting compared to a laboratory setting[11]. Because the design requires that I conduct sessions during the trading hours of the SMI and DAX stock indices, participants were asked to select a time slot before taking part in the study. The link to the study was sent out in a separate e-mail shortly before the session started. Due to the nature of the experiments, some participants could access the experiment from their home, while others could do so in a noisy environment. I thus asked participants to make sure that they are in a quiet place without any distractions when starting the experiment.

In the experiment, I informed participants that the data is anonymized and treated confidentially. The context of the experiment was framed as a survey on health-related and and risk-related questions, for which participants are being paid. Participants first received instructions of the experiment via the experimental software oTree (Chen et al., 2016). The experiment then proceeds to the game, and participants were assigned to one of the conditions, assuring equal distribution of treatments within one experimental session.

Participants were compensated with a fixed participation fee of CHF 6 plus an additional payoff that varied with each participant and was conditional on a random draw (i.e., ranging between CHF 0 and CHF 9). Payments were sent to the participants' bank accounts the evening of the day of participation. To strengthen the credibility of the payment procedure, I asked subjects to enter their bank account information that is (or will be) associated with their PayPal account in the description of the study as well as in the experimental instructions. I asked participants for their bank information on a separate website connected to a separate database that I cannot link with the experimental data. This also reduces the possibility that some subjects will participate more than once. The average duration of an experimental session was about 9 min.

TABLE 1 | Summary statistics by group.

Variable	Levels	n	Min	$\bar{x}$	Max
Age (in years)	T_{RN}	67	18	24.37	51
	T_{FM}	68	19	23.57	30
$p = 0.67$	all	135	18	23.97	51
Gender	T_{RN}	67	0	0.48	1
	T_{FM}	68	0	0.50	1
$p = 0.80$	all	135	0	0.49	1
Income	T_{RN}	67	0	0.43	1
	T_{FM}	68	0	0.59	1
$p = 0.07$	all	135	0	0.51	1
Income source	T_{RN}	67	0	0.55	1
	T_{FM}	68	0	0.56	1
$p = 0.94$	all	135	0	0.56	1
Risk aversion	T_{RN}	67	2	6.04	10
	T_{FM}	68	1	5.82	10
$p = 0.37$	all	135	1	5.93	10

[11] Anderhub et al. (2001) show that subjects are less attentive in an online experiment compared to a class experiment. In contrast, Bosch-Domenech et al. (2002) find that the results of a guessing game are similar in both settings.

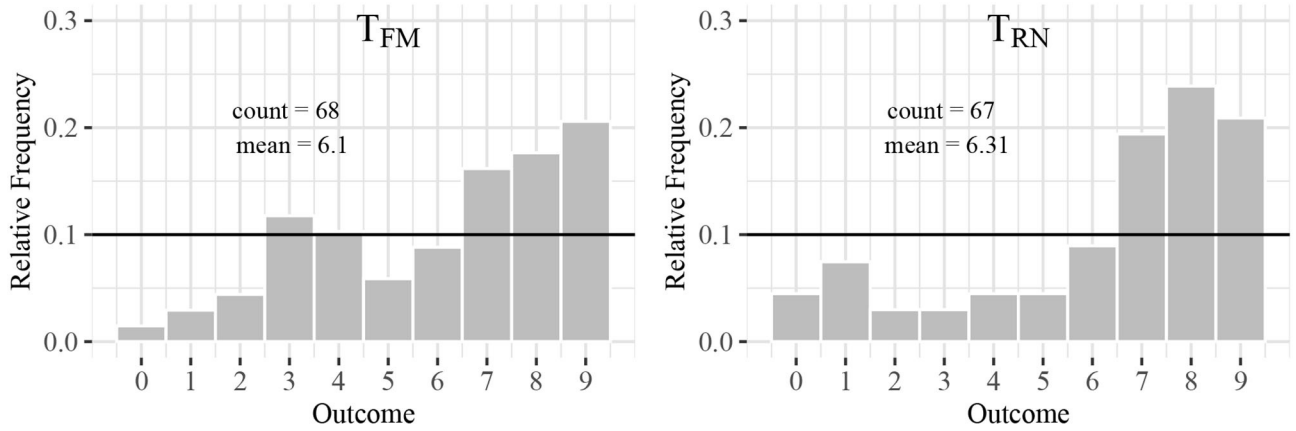


FIGURE 2 | Reports of random draws. The figures show reported random draws conditional on treatment assignment.

TABLE 2 | Results of OLS and probit regressions.

	Value reported (1)	High report (2)
T_{FM}	−0.094	−0.089
	(0.452)	(0.085)
Gender	−0.930**	−0.124
	(0.451)	(0.084)
Constant	5.064***	
	(1.570)	
Controls	Yes	Yes
Observations	135	135
R^2	0.046	
Adjusted R^2	0.016	

*The table shows (1) OLS estimates of the treatment effects on the reported random draw defined on the interval between 0 and 9 and (2) probit estimates of the treatment effects of reporting very high values (i.e., > 6). The reference category is T_{RN}. Additional independent variables include age in years, dummies for being female, different study major. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.*

3. RESULTS

3.1. Summary Statistics

The participants in the study were 135 students at the University of St. Gallen and the Fachhochschule St. Gallen. In terms of gender, the sample is quite balanced. The sample includes 69 (0.51) men and 66 (0.49) women with an average age of 23.97 years, ranging from 19 to 51. **Table 1** provides summary statistics[12].

Regarding the categorial variables, about 51 percent of respondents reported to have between CHF 500 and CHF 1,499 at their disposal per month. When asked about their sources of income, more than 50 percent of respondents indicated their family, 30 percent referenced their job, and 10 percent reported a scholarship as their main source of income. Sixty-seven percent of participants ranked their health status as excellent or very good. **Table A1** in the Appendix provides further details on the subjects' demographics across different samples.

3.2. Main Results

Figure 2 shows the distribution of reported outcomes conditional on treatment assignment, and **Table 2** shows estimates of treatment effects on reported outcomes (OLS) and reporting very high outcomes (Probit). In general, numbers above (below) six are significantly more (less) frequently reported than their expected true share of 10 percent (p <0.001). This suggests that some participants reported higher numbers than the one they had actually seen. I can thus confirm the findings on dishonest behavior from previous studies (Fischbacher and Föllmi-Heusi, 2013). Contrary to the expectations, participants in treatment T_{FM} do not cheat more frequently than participants in treatment T_{RN} (Kolmogorov-Smirnov test: p = 0.896). I further observe that controlling for additional individual characteristics does not have an effect on the significance of the differences between the two treatments. This confirms previous studies, which find no significant differences in student samples between a financial and a neutral context (Huber and Huber, 2020). Similarly, Cappelen et al. (2013) do not find a significant effect when priming students to think about markets.[13]. I consider additional heterogeneous treatment effects. As previous research shows (Capraro, 2018; Gerlach et al., 2019), I find that women are more honest on average (p = 0.041).

[12]For ease of interpretation, I recoded the categorial variables into binary variables (i.e., dummy coding). Gender is a dummy variable that takes on the value 1 when the participant is female. Income is a dummy variable that takes on the value of 1 if the participant has between CHF 500 to 1,499 at his/her disposal per month. Income source is a dummy taking on the value of 1 if their main source of income is their family.

[13]In the base treatment, participants are asked to write about the city of Bergen, Norway. In the market treatment, they are asked to write about when they had benefited from buying or selling a good or service.

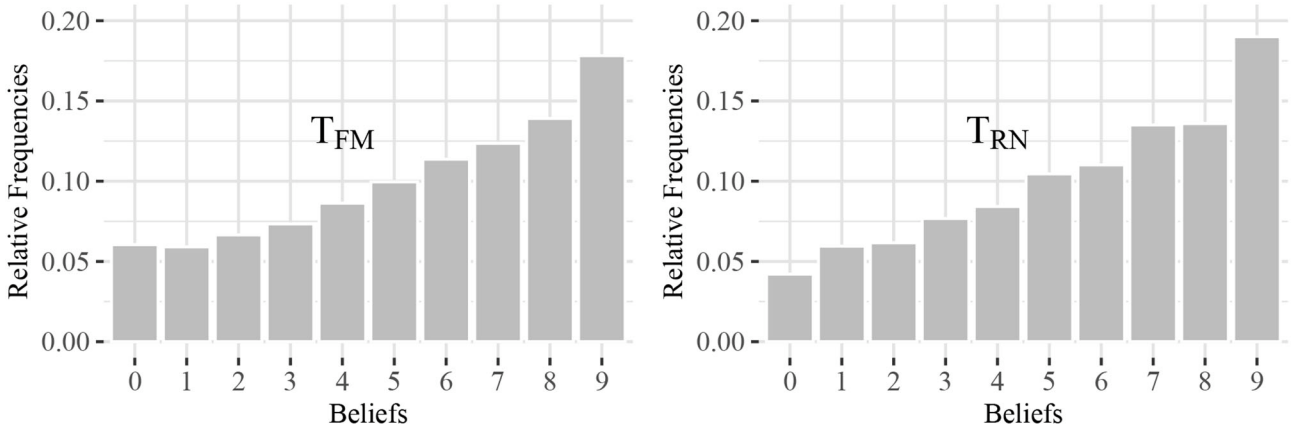


FIGURE 3 | Beliefs about reported payoffs in reference experiment. The figures show the beliefs about reported payoffs in a reference experiment in terms of the share of participants that reported a payoff.

I further examine the role of dispositional greed in explaining potential differences in dishonest behavior between the two groups. It is possible that the financial market setting evokes greedy behavior. Earlier research shows that greed is associated with fraudulent behavior (Seuntjens et al., 2019). However, I cannot find a significant impact of dispositional greed on dishonest behavior ($p = 0.841$).

3.3. Elicitation of Descriptive Norms

A large body of research shows that dishonest behavior also depends on the social norms implied by the dishonesty of others or by beliefs about what constitutes honest behavior (Fischbacher and Föllmi-Heusi, 2013; Cohn et al., 2014; Kocher et al., 2018). To identify norms separately from behavior, I use the norm elicitation method by Krupka and Weber (2013). In particular, I aim to test whether different expectations exist toward dishonest behavior when the source of uncertainty is a stock market index. I thus aim to test whether different expectations exist toward dishonest behavior when the source of uncertainty is a stock market index. The focus in this research is on descriptive norms.

I conduct an additional experiment with a new set of subjects. In the experiment, participants must guess other participants' reporting behavior. More specifically, I prompt participants to predict the behavior of other participants in a previously run experiment (i.e., the "reference experiment"). On the first page of the experiment, I explain the setting of the reference experiment. Participants then guess what percentage of participants reported a specific payoff. They were paid depending on the accuracy of their predictions (Fischbacher and Föllmi-Heusi, 2013). Participants could earn CHF 9 if they guess all shares correctly. For every percentage point deviation from the correct share, I reduce participants' payoff by CHF 0.1. The minimum payoff in the belief elicitation task is CHF 1. Participants received a show-up fee of CHF 6 that was added to the earnings from the experiment. I recruited participants from the participant pool of the behavioral lab at the University of St.Gallen and excluded all subjects with previous experience in similar experiments. In total, 95 participants took part in this experiment (48 had to guess the behavior in T_{FM} and 47 in T_{RN}). **Figure 3** shows participants' beliefs about the behavior of others in terms of honesty. The data shows that beliefs increase in the reported number. Subjects report a belief that a fraction of more than 10 percent reported the highest number. Thus, they believe that similar decision makers act dishonestly. I observe that the distributions of beliefs correspond fairly closely to the distributions of the actual reporting behavior. This shows that participants act in accordance with the perceived norm. I do not find a significant influence of the frame-specific device on beliefs. A Kolmogorov-Smirnov test indicates that the two distributions are not significantly different from each other ($p = 0.880$). More specifically, the frame-specific device does not shift participants' beliefs about the prevailing honesty norm.

3.4. Laboratory Evidence

I extend the study to the traditional laboratory (using the same treatment conditions T_{RN} and T_{FM}) in order to understand whether the environment (i.e., traditional lab experiment or online experiment) has an effect on the decision to be honest. To make online experiments comparable with laboratory experiments, investigating potential differences in results is of crucial importance. The environment of subjects in a laboratory is quite different from the environment of subjects taking part in the study using a Web browser at home.[14]. It is possible that participants feel more socially distant from others in an online environment. This might reduce participants' need to adhere to behavior norms. In a typical laboratory setting, participants can see each other and possibly even talk to each other. I therefore replicate the experiment in the more constrained setting of the traditional physical laboratory.

[14]Previous studies that compare laboratory to Internet data mostly use a very different subject pool in the online experiment. I, however, conduct an online experiment with subjects comparable to the subjects in the laboratory, as I recruit subjects from the same participant pool.

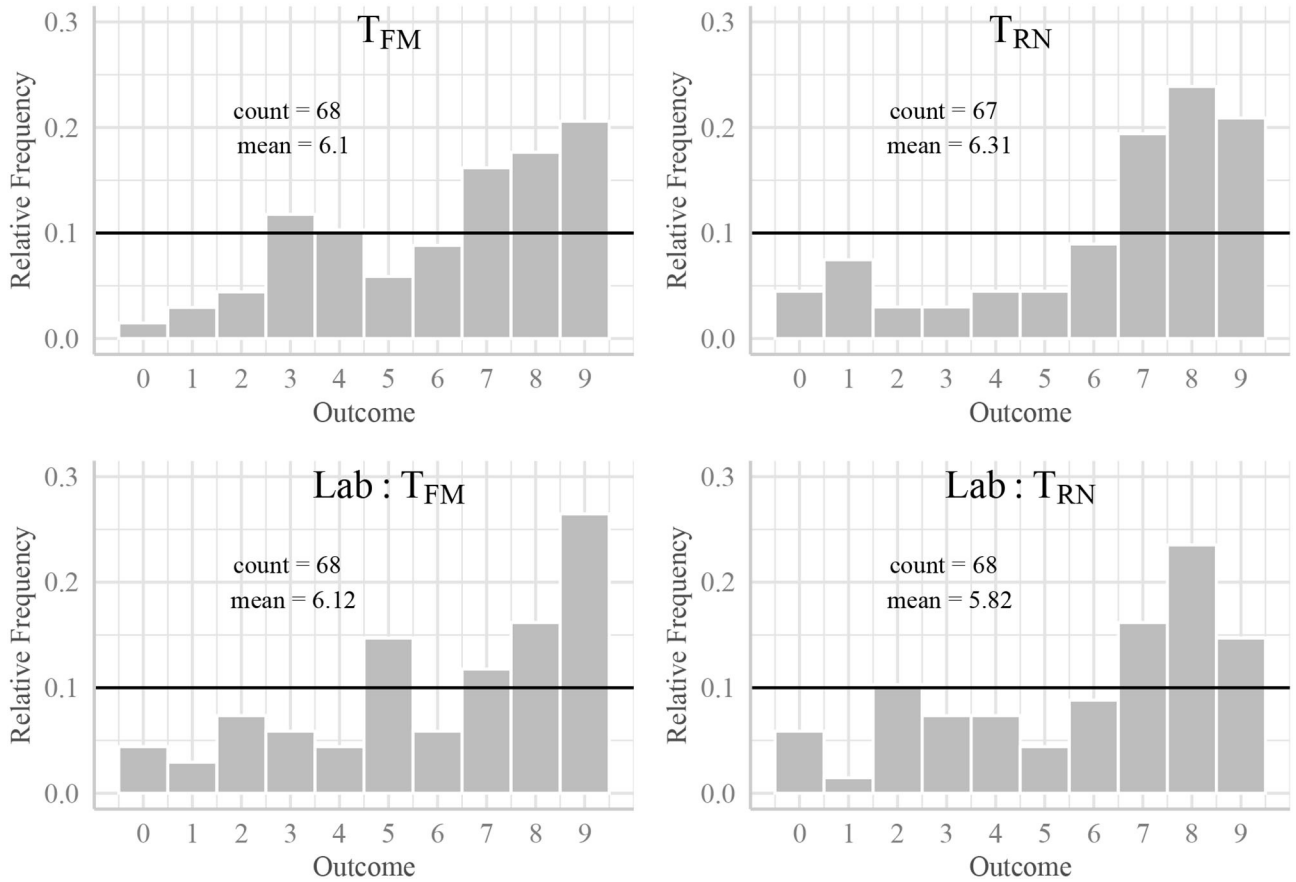


FIGURE 4 | Reports of random draws. The figures show reported random draws conditional on treatment assignment.

I recruited participants from the participant pool of the behavioral lab at the University of St.Gallen using the same instructions (i.e., random number or financial market), the same incentive-compatible design, and the same decision interface. This ensures the credibility of comparability of the two groups. During the experiment, each participant sat at a randomly assigned, separated PC terminal. No form of communication was allowed during the experiment. I conducted all sessions at the behavioral lab in St. Gallen. I excluded all subjects with previous experience in the honesty task. The participants in the study were 135 students at the University of St. Gallen and the Fachhochschule St. Gallen. The sample appears balanced across treatment conditions (see **Table A2** in the Appendix). This is expected due to the randomized assignment to treatment. To make payments in the lab as salient as in the online setting, payments were sent to the participants' bank accounts the evening of the day of participation.

Figure 4 shows the distribution of reported outcomes conditional on treatment assignment. As a means of comparison, I show both results from the online and the laboratory experiment. Supporting earlier results from the online setting, I find that participants in the lab are not more dishonest in treatment T_{FM}. To put it differently, mean outcomes do not significantly differ depending on the environment (KS test $p = 0.734$).

In a next step, I explore differences in decision times depending on the environment (i.e., laboratory or online). A large body of literature suggests that deception is cognitively more demanding than responding honestly, and, thus, honesty is considered as behavioral default (Foerster et al., 2013). This conclusion was supported by more recent studies which find that time pressure promotes honesty (Capraro, 2017; Capraro et al., 2019). Other research, however, reported the opposite (Shalvi et al., 2012). More precisely, Shalvi et al. (2012) show that lying is an initial, automatic tendency that is overcome only if sufficient time to deliberate is available and if dishonest behavior cannot be justified. This is supported by earlier findings in neuropsychological research showing that the right dorsolateral prefrontal cortex, a brain area involved in executive control, is associated with overriding selfish impulses in economic decisions (Knoch et al., 2006) and that this area, together with two other brain areas associated with self-control, is activated when individuals make an effort to forgo lying (Greene and Paxton, 2009).

To control for differences in decision times, I observe the time difference (in seconds) between the instruction to look

TABLE 3 | Results of OLS regressions.

	Reporting time
Seconds	0.365
	(0.457)
Lab	0.007
	(0.013)
Seconds:Lab	−0.034**
	(0.017)
Constant	5.863***
	(1.375)
Controls	Yes
Observations	270
R^2	0.028
Adjusted R^2	0.010

*The table shows OLS estimates of reporting time (i.e., time spent on the reporting page) as well as its interaction with a dummy variable indicating whether subjects conducted the study in the laboratory or at home on the reported random draw defined on the interval between 0 and 9. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.*

up the stock index value and the actual reporting of results (this is not visible to participants). I find that the average decision time in the laboratory is significantly higher than the average decision time online ($p = 0.042$). This confirms previous results (Anderhub et al., 2001; Hergueux and Jacquemet, 2015). I include decision times in the regressions presented in **Table 3**, both as an additional control variable and as an interaction term with a dummy variable indicating whether the experiment was conducted in the lab or online. This allows capturing the variation in honest behavior that is induced by the environment and decision times. The results confirm what (Shalvi et al., 2012) had indicated—participants who take longer to report are more honest. However, this is only true for laboratory subjects. As depicted in **Table 3**, a 1-s increase in the time to report changes the report by –0.034 ($p = 0.042$) for participants in the laboratory. This could potentially be explained by the fact that the latter group of participants take more time to think about others' behavior (i.e., what is socially acceptable).

My conclusion that honesty is deliberative should, however, be interpreted with caution. A recent replication study (Van der Cruyssen et al., 2020) was not able to yield support for the original study of Shalvi et al. (2012). Having said this, my results are in line with a recent meta-analysis, indicating that, honesty is deliberative, but only when no concrete other is harmed (Köbis et al., 2019).

Finally, I look at potential gender differences in terms of reporting behavior in the laboratory experiment by including a dummy variable for gender into the regression. I additionally include an interaction between the gender dummy and a dummy indicating whether subjects conducted the study in the laboratory or online. This interaction term allows testing whether either gender is more sensitive to the environment. **Table A3** in the Appendix presents the results. The results illustrate that the coefficient of the gender dummy is significantly different depending on the environment—men are significantly more honest when they conduct the experiment in the laboratory as opposed to online ($p = 0.021$). Thus, lesser social distance affects truth-telling behavior of men. Another explanation may be subjects' reputation. Even though the experimental design allows to credibly eliminate any reputation concerns, it is possible that participants feel observed by other students (and the experimenter) when they are sitting in the lab. Even the slightest cues of being observed seem to affect male but not female reporting behavior.

4. CONCLUSION

In this study, I investigate dishonest behavior using frame-specific randomization devices to vary the situational context of the game (i.e., stock market or neutral context). The results show no significant differences in dishonest behavior between the two groups. This indicates that this specific source of uncertainty is not prone to the social norm effects documented in the literature. The findings confirm previous studies (Cappelen et al., 2013; Huber and Huber, 2020) and extend them by varying the setting.

As different environments can render different social norms salient, I replicate the experiment in the more constrained setting of the traditional physical laboratory. I cannot confirm significant differences in dishonest behavior depending on the environment. Additional estimations capture the variation in honest behavior that is induced not only by the environment, but also decision times. I find that participants who take longer to report are more honest. However, this is only true for subjects in the physical laboratory. Depending on the experimental setting, the inclusion of controls for differences in decision times among online subjects can be important for future studies. Finally, the results suggest that even the slightest cues of being observed affects truth-telling behavior of male but not female participants.

The present study has some limitations. First, the frame-specific device showing a stock market index may not have been strong enough to activate the norms related to financial markets. Second, due to the nature of the online experiment, some participants may have accessed the experiment from a quiet place, while others may have done so in a noisy environment. However, the design is conceptually rather simple, which should reduce concerns that subjects are less attentive in the online environment.

Lastly, this paper also makes a methodological contribution. The experimental approach to measure dishonest behavior outside of the lab can be applied broadly in decentral experimental setups as well as surveys. Non-physical and verifiable sources of uncertainty are key to extending the valid measurement of dishonest behavior to broader settings such as online experimentation. The non-strategic nature of the experiment makes it rather easy to establish different environments in which one can hold constant important features of the decision task—the payoffs, the description of the way the task works, and so on—, while varying context in a way that can influence social norms.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The study was reviewed and approved by the Ethics Committee of the University of St. Gallen, Switzerland (HSG-EC-20210715-A). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.684735/full#supplementary-material

REFERENCES

Abeler, J., Becker, A., and Falk, A. (2014). Representative evidence on lying costs. *J. Pub. Econ.* 113, 96–104. doi: 10.1016/j.jpubeco.2014.01.005

Anderhub, V., Müller, R., and Schmidt, C. (2001). Design and evaluation of an economic experiment via the internet. *J. Econ. Behav. Organ.* 46, 227–247. doi: 10.1016/S0167-2681(01)00195-0

Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Qu. J. Econ.* 110, 1–21.

Barron, K. (2019). *Lying to Appear Honest*. WZB Discussion Paper No. SP II.

Bohnet, I., and Frey, B. S. (1999). The sound of silence in prisoner's dilemma and dictator games. *J. Econ. Behav. Organ.* 38, 43–57.

Bosch-Domènech, A., Montalvo, J. G., Nagel, R., and Satorra, A. (2002). One, two, (three), infinity, …: newspaper and lab beauty-contest experiments. *Am. Econ. Rev.* 92, 1687–1701. doi: 10.1257/000282802762024737

Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2013). When do we lie? *J. Econ. Behav. Organ.* 93, 258–265. doi: 10.1016/j.jebo.2013.03.037

Capraro, V. (2017). Does the truth come naturally? Time pressure increases honesty in one-shot deception games. *Econ. Lett.* 158, 54–57. doi: 10.1016/j.econlet.2017.06.015

Capraro, V. (2018). Gender differences in lying in sender-receiver games: a meta-analysis. *Judg. Decis. Making* 13, 345–355. doi: 10.2139/ssrn.2930944

Capraro, V., and Perc, M. (2021). Mathematical foundations of moral preferences. *J. R. Soc. Interf.* 18. doi: 10.1098/rsif.2020.0880

Capraro, V., Schulz, J., and Rand, D. G. (2019). Time pressure and honesty in a deception game. *J. Behav. Exper. Econ.* 79, 93–99. doi: 10.1016/j.socec.2019.01.007

Capraro, V., and Vanzo, A. (2019). The power of moral words: Loaded language generates framing effects in the extreme dictator game. *Judg. Decis. Making* 14, 309–317. doi: 10.2139/ssrn.3186134

Chang, D., Chen, R., and Krupka, E. (2019). Rhetoric matters: a social norms explanation for the anomaly of framing. *Games Econ. Behav.* 116, 158–178. doi: 10.1016/j.geb.2019.04.011

Charness, G., and Gneezy, U. (2008). What's in a name? Anonymity and social distance in dictator and ultimatum games. *J. Econ. Behav. Organ.* 68, 29–35. doi: 10.1016/j.jebo.2008.03.001

Charness, G., Haruvy, E., and Sonsino, D. (2007). Social distance and reciprocity: an Internet experiment. *J. Econ. Behav. Organ.* 63, 88–103. doi: 10.1016/j.jebo.2005.04.021

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *J. Behav. Exper. Finance* 9, 88–97. doi: 10.1016/j.jbef.2015.12.001

Cohn, A., Fehr, E., and Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature* 516, 86–89. doi: 10.1038/nature13977

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: measurement, determinants, and behavioral consequences. *J. Eur. Econ. Assoc.* 9, 522–550. doi: 10.1111/j.1542-4774.2011.01015.x

Dreber, A., Ellingsen, T., Johannesson, M., and Rand, D. G. (2013). Do people care about social context? Framing effects in dictator games. *Exper. Econ.* 16, 349–371. doi: 10.1007/s10683-012-9341-9

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games Econ. Behav.* 73, 459–478. doi: 10.1016/j.geb.2011.02.003

Ellingsen, T., Johannesson, M., Mollerstrom, J., and Munkhammar, S. (2012). Social framing effects: preferences or beliefs? *Games Econ. Behav.* 76, 117–130. doi: 10.1016/j.geb.2012.05.007

Erat, S., and Gneezy, U. (2012). White lies. *Manag. Sci.* 58, 723–733. doi: 10.1287/mnsc.1110.1449

FBI (2020). *Insurance Fraud*. Availabel online at: https://www.fbi.gov/stats-services/publications/insurance-fraud

Fehr, E., and Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook Econ. Giving Altruism Reciprocity* 1, 615–691. doi: 10.1016/S1574-0714(06)01008-6

Fischbacher, U., and Föllmi-Heusi, F. (2013). Lies in disguise-an experimental study on cheating. *J. Eur. Econ. Assoc.* 11, 525–547. doi: 10.1111/jeea.12014

Foerster, A., Pfister, R., Schmidts, C., Dignath, D., and Kunde, W. (2013). Honesty saves time (and justifications). *Front. Psychol.* 4:473. doi: 10.3389/fpsyg.2013.00473

Gerlach, P., Teodorescu, K., and Hertwig, R. (2019). The truth about lies: a meta-analysis on dishonest behavior. *Psychol. Bull.* 145, 1–44. doi: 10.1037/bul0000174

Gneezy, U. (2005). Deception: the role of consequences. *Am. Econ. Rev.* 95, 384–393. doi: 10.1257/0002828053828662

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *Am. Econ. Rev.* 108, 419–453. doi: 10.1257/aer.20161553

Greene, J. D., and Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12506–12511. doi: 10.1073/pnas.0900152106

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* 91, 73–78. doi: 10.1257/aer.91.2.73

Hergueux, J., and Jacquemet, N. (2015). Social preferences in the online laboratory: a randomized experiment. *Exper. Econ.* 18, 251–283. doi: 10.1007/s10683-014-9400-5

Hoffman, E., McCabe, K., and Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *Am. Econ. Assoc.* 86, 653–660.

Huber, C., and Huber, J. (2020). Bad bankers no more? Truth-telling and (dis)honesty in the finance industry. *J. Econ. Behav. Organ.* 180, 472–493. doi: 10.1016/j.jebo.2020.10.020

Internal Revenue Service (2019). *Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2011–2013*. Publication 1415.

Kay, A. C., and Ross, L. (2003). The perceptual push: the interplay of implicit cues and explicit situational construals on behavioral intentions in the prisoner's dilemma. *J. Exper. Soc. Psychol.* 39, 634–643. doi: 10.1016/S0022-1031(03)00057-X

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, T., and Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314, 829–932. doi: 10.1126/science.1129156

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2019). Intuitive honesty versus dishonesty: meta-analytic evidence. *Perspect. Psychol. Sci.* 14, 778–796. doi: 10.1177/1745691619851778

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? we lie! why? Experimental evidence on a dishonesty shift in groups. *Manag. Sci.* 64, 3995–4008. doi: 10.1287/mnsc.2017.2800

Krupka, E. L., and Weber, R. A. (2013). Identifying social norms using coordination games: why does dictator game sharing vary? *J. Eur. Econ. Assoc.* 11, 495–524. doi: 10.1111/jeea.12006

Liberman, V., Samuels, S. M., and Ross, L. (2004). The name of the game: predictive power of reputations versus situational labels in determining Prisoner's Dilemma game moves. *Person. Soc. Psychol. Bull.* 30, 1175–1185. doi: 10.1177/0146167204264004

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: a theory of self-concept maintenance. *J. Market. Res.* 45, 633–644. doi: 10.1509/jmkr.45.6.633

Pfister, R., Wirth, R., Weller, L., Foerster, A., and Schwarz, K. A. (2019). Taking shortcuts: cognitive conflict during motivated rule-breaking. *J. Econ. Psychol.* 71, 138–147. doi: 10.1016/j.joep.2018.06.005

Rahwan, Z., Yoeli, E., and Fasolo, B. (2019). Heterogeneity in banker culture and its influence on dishonesty. *Nature* 575, 345–349.

Seuntjens, T. G., Zeelenberg, M., Van De Ven, N., and Breugelmans, S. M. (2015). Dispositional greed. *J. Person. Soc. Psychol.* 108, 917–133. doi: 10.1037/pspp0000031

Seuntjens, T. G., Zeelenberg, M., Van De Ven, N., and Breugelmans, S. M. (2019). Greedy bastards: testing the relationship between wanting more and unethical behavior. *Person. Indiv. Diff.* 138, 147–156. doi: 10.1016/j.paid.2018.09.027

Shalvi, S., Eldar, O., and Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychol. Sci.* 23, 1264–1270. doi: 10.1177/0956797612443835

Stöckl, T. (2015). Dishonest or professional behavior? Can we tell? A comment on: Cohn et al. 2014, Nature 516, 86-89, "Business culture and dishonesty in the banking industry". *J. Behav. Exper. Finance* 8, 64–67. doi: 10.1016/j.jbef.2015.10.003

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458.

Van der Cruyssen, I., D'hondt, J., Meijer, E., and Verschuere, B. (2020). Does honesty require time? Two preregistered direct replications of experiment 2 of Shalvi, Eldar, and Bereby-Meyer (2012). *Psychol. Sci.* 31, 460–467. doi: 10.1177/0956797620903716

Vohs, K. D. (2015). Money priming can change people's thoughts, feelings, motivations, and behaviors. *J. Exper. Psychol. Gen.* 144, e86–e93. doi: 10.1037/xge0000091

Vohs, K. D., Mead, N. L., and Goode, M. R. (2008). Merely activating the concept of money changes personal and interpersonal behavior. *Curr. Direct. Psychol. Sci.* 17, 208–212. doi: 10.1111/j.1467-8721.2008.00576.x

Vranka, M. A., and Houdek, P. (2015). Many faces of bankers' identity: How (not) to study dishonesty. *Front. Psychol.* 6:302. doi: 10.3389/fpsyg.2015.00302

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 September 2021
doi: 10.3389/fpsyg.2021.675776

People Judge Discrimination Against Women More Harshly Than Discrimination Against Men – Does Statistical Fairness Discrimination Explain Why?

Eberhard Feess[1], *Jan Feld*[1,2]* *and Shakked Noy*[1]

[1] School of Economics and Finance, Victoria Business School, Victoria University of Wellington, Wellington, New Zealand,
[2] Institute of Labor Economics, Bonn, Germany

OPEN ACCESS

Edited by:
Nora Szech,
Karlsruhe Institute of Technology (KIT), Germany

Reviewed by:
Rosemary Hopcroft, University of North Carolina at Charlotte, United States
Cory Clark, Florida State University, United States
Michael Kurschilgen, Technical University of Munich, Germany

***Correspondence:**
Jan Feld
jan.feld@vuw.ac.nz

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

Received: 03 March 2021
Accepted: 23 July 2021
Published: 20 September 2021

Citation:
Feess E, Feld J and Noy S (2021) People Judge Discrimination Against Women More Harshly Than Discrimination Against Men – Does Statistical Fairness Discrimination Explain Why? Front. Psychol. 12:675776. doi: 10.3389/fpsyg.2021.675776

Previous research has shown that people care less about men than about women who are left behind. We show that this finding extends to the domain of labor market discrimination: In identical scenarios, people judge discrimination against women more morally bad than discrimination against men. This result holds in a representative sample of the US population and in a larger but not representative sample of Amazon Mechanical Turk (Mturk) respondents. We test if this gender gap is driven by statistical fairness discrimination, a process in which people use the gender of the victim to draw inferences about other characteristics which matter for their fairness judgments. We test this explanation with a survey experiment in which we explicitly hold information about the victim of discrimination constant. Our results provide only mixed support for the statistical fairness discrimination explanation. In our representative sample, we see no meaningful or significant effect of the information treatments. By contrast, in our Mturk sample, we see that providing additional information partly reduces the effect of the victim's gender on judgment of the discriminator. While people may engage in statistical fairness discrimination, this process is unlikely to be an exhaustive explanation for why discrimination against women is judged as worse.

Keywords: gender, discrimination, statistical fairness discrimination, employment discrimination (gender), moral judgments

INTRODUCTION

Labor market outcomes of women are still worse than those of men. The gender wage gap is decreasing over time, but the ratio between full-time median salaries for women and men still varies between about 90% in Continental Europe and around 80% in the United States and the United Kingdom (Olivetti and Petrongolo, 2016; Ortiz-Ospina and Roser, 2018). Discrepancies in salaries are associated with differences in occupations, but are also found within the same occupation (see the overview by Kunze, 2018). In addition, field studies suggest that women are assigned to less challenging tasks than their male colleagues (De Pater et al., 2010; Bertrand, 2011; Chan and Anteby, 2016; Babcock et al., 2017), and are disadvantaged with regards to career opportunities (Allen et al., 2016) and dismissal decisions (Gupta et al., 2020).

Recent literature analyzes many potential channels for these differences, including differences in capital accumulation, preferences for professions, job descriptions, and competitiveness (see the overviews in Azmat and Petrongolo, 2014; Niederle, 2017). The literature also points to the role of stereotypes and discrimination (see the overview in Blau and Kahn, 2017). Laboratory experiments explore discrimination in controlled environments, and find that women with the same performance are less likely to be hired for male-stereotyped tasks (Reuben et al., 2014; Bohnet et al., 2016; Coffman et al., 2018). According to the European Working Conditions Survey, 3.2% of women report having been subjected to discrimination at work on the basis of their sex in the preceding 12 months, compared to only 1.1% of men (Eurofund, 2018).

Notably, however, women have also overtaken men in some domains. Across most OECD countries, women are now more likely to graduate from university (OECD, 2020). There are growing concerns about the job prospects of low-skilled men, who have seen a significant reduction in real income in the US, and many of whom have left the labor force (Autor and Wasserman, 2013; Binder and Bound, 2019). There is also a growing literature documenting a gender bias against men in several fields (Booth and Leigh, 2010; Breda and Hillion, 2016; Bohren et al., 2019).

Against this backdrop, we investigate how the gender of the victim of discrimination affects people's moral judgment about the discriminator. We answer this question using a survey experiment with two samples. Our main sample consists of 478 respondents who are representative of the US population in terms of gender, age, race, education and political orientation. Our replication sample consists of 1,169 US based respondents recruited from Amazon Mechanical Turk (Mturk).[1] For this study, we define gender discrimination as hiring someone of one gender *despite knowing that an applicant of the opposite gender is more qualified and more productive*. In our base treatment, each respondent is shown two scenarios (in random order) in which a manager discriminates, and is asked to evaluate managers' decisions on a scale that ranges from 0 "very morally wrong" to 100 "very morally right." In one of these scenarios, the manager discriminates against a woman, and in the other against a man.

We use these judgments to estimate the effects of the victim's gender on the moral evaluation of the actions of the discriminator in two ways. The first is the (within-subject) *pro-women attitude*, which we define as the difference in judgment of a manager who discriminates against a woman compared to the judgment of a manager who discriminates against a man *for a particular respondent*. Using this measure, we find that respondents judge discrimination against women on average 5.5 points more morally wrong. This measure is based on respondents' answers to two scenarios - presented right after each other – in which only the gender of the victim differs. By judging these two related scenarios, however, respondents may feel the need to be consistent, which would reduce the measured pro-women attitudes. For example, having just judged discrimination against a woman as very morally bad may compel respondents to also judge discrimination against a man as very bad. By contrast, in many real world applications respondents only judge one case of discrimination at a time. We therefore also report a *between-subject pro-women attitude*, which is the difference in judgment of discrimination against a woman compared to discrimination against a man, based on respondents' judgment of the scenario they saw first. The between-subject pro-women attitude is 11.8 points, which is substantially larger. We replicate both of these results with our Mturk sample.

We further investigate potential reasons for the pro-women attitude. This investigation is inspired by a recent study by Cappelen et al. (2019), who ask whether people are less concerned about men falling behind than about women falling behind. In their experiment, observers can redistribute money from women who win to losing men and vice versa. Observers are less likely to redistribute money to low-performing men, suggesting that they are indeed more accepting of men falling behind. Interestingly, this gender gap disappears when losers and winners are determined by chance instead of their performance in a real effort task. The authors interpret this result as evidence for statistical fairness discrimination, that is, that people use gender as a signal for unobserved characteristics which matter for their fairness judgments. In the experiment, people who engage in statistical fairness discrimination may be less likely to help men because they believe that men who have fallen behind have worked less hard. This interpretation is consistent with earlier findings by Cappelen et al. (2007) that many people believe productivity differences justify wage differentials if and only if they reflect different effort.

We explore the role of statistical fairness discrimination in explaining the average pro-women attitude with an embedded survey experiment. We ask randomly selected respondents to judge additional scenarios that are either very similar to the base scenarios (control group) or explicitly state that the job is in an industry without gender discrimination, that the man and woman who applied for the job worked equally hard in their career, and that both applicants would suffer equally from not getting the job (treatment group).

The results of this survey experiment show only mixed support for the statistical fairness discrimination hypothesis. In our main sample with Qualtrics respondents, we see no evidence that the additional information changes respondents' pro-women attitude. The average pro-women attitudes in the treatment and control groups are very similar, and none of the differences are statistically significant. By contrast, in our replication sample with Mturk respondents, we do see that providing additional information significantly reduces the pro-women attitude of respondents who exhibited a positive pro-women attitude in the base scenarios. However, even in scenarios in which we hold applicants' effort, suffering, and exposure to discrimination constant, pro-women respondents still show a statistically significant pro-women attitude. While statistical fairness discrimination may play a role in explaining differences in judgments about discriminated women and men, it is unlikely to be the whole story.

The concept of statistical fairness discrimination as defined by Cappelen et al. (2019) builds on the more general

[1] We pre-registered our study at socialscienceregistry.org (Feess et al., 2021).

and widely discussed concept of statistical discrimination (Phelps, 1972; Arrow, 1973). In both concepts, observable characteristics of people are used to infer unobservable, but relevant, characteristics. Traditional statistical discrimination may enhance efficiency, but it also violates widespread fairness norms. A famous and highly controversial example is the finding by Knowles et al. (2001) that police checking for illegal drugs are more likely to search cars of Black than White drivers, which the authors argue equilibrates the detection probabilities for the two groups at the margin. Many forms of statistical discrimination are prohibited. In many countries it is illegal to use race, sex, age or disability as criteria for decisions on hiring or promotion, even if these characteristics predict performance. While traditional statistical discrimination involves a tension between efficiency and fairness, statistical fairness discrimination is based on fairness considerations itself. It builds on the idea people use group characteristics to draw inferences about unobserved characteristics (e.g., deservingness of help) that matter for their fairness judgments.

The concept of statistical fairness discrimination may be useful for a better understanding of a wide range of differences in people's social preferences (see the overview by Eckel and Grossman, 2008). Many papers find that subjects care more about women in social dilemma situations (FeldmanHall et al., 2016), that defendants killing women are far more likely to be sentenced to death than defendants killing men (Shatz and Shatz, 2012), and that subjects give more to women in dictator games (Engel, 2011). While these observations may just reflect people caring more about the wellbeing of women (see the literature review in Eagly and Mladinic, 1994), they might also reflect statistical fairness discrimination in the sense that, for example, women are seen as more vulnerable.

Our finding that people are less concerned about discrimination against men than women relates to a paper by Block et al. (2019). Their paper first shows that people are more concerned about the underrepresentation of women in male-dominated careers (Science, Technology, Engineering, and Math) than about men in female-dominated careers (Healthcare, Early Education, and Domestic roles). They derive three main insights on the reasons for this difference: First, people believe that underrepresentation only deserves countervailing measures if it is based on discrimination rather than on preferences. As men are perceived as being not interested in female-dominated careers, there seems to be no reason to worry about their underrepresentation. Second, female-dominated careers are viewed as less prestigious, so that underrepresentation of males is not interpreted as a disadvantage. Third, differences in salaries hardly matter for peoples' different concerns.

In another set of studies, Winegard et al. (2018) document that liberals' judgments favor groups they perceive to be disadvantaged, like women and Black people. Their approach is similar to ours in that they compare judgments of identical situations in which one key demographic characteristics differs (see also Stewart-Williams et al., 2021). For example, they show liberals trust otherwise identical scientific studies more if the results are favorable for disadvantaged groups (women and Blacks) than privileged groups (men and Whites). These differences in judgments are consistently predicted by Equalitarianism – the belief that differences between demographic groups are not driven by biological factors but by prejudice, and that society can and should make all groups equal.

In another related paper, Haaland and Roth (2021) use a representative sample of the US household population to analyze beliefs about racial discrimination, and investigate how these beliefs are correlated with the view on affirmative action. They also find that providing different kinds of information influences people's perceptions of discrimination. Interestingly, the authors show that, while providing accurate information changes the beliefs on the actual degree of discrimination, it has only little impact on the view on affirmative action.

Overall, our paper makes three main contributions. First, we add to a body of research showing that many people show more concern for disadvantaged groups than advantaged groups. While these studies are typically done with convenient samples (e.g., from Mturk) we show that this conclusion also holds in a representative sample. Second, we carefully investigate to what extent these differences in judgment are driven by statistical fairness discrimination – an explanation which has only received limited attention in the literature. Our embedded survey experiment lends some, but not very strong support for this explanation. Third, our comparison of the within-subject and between-subject results reveals that people's judgments are influenced by the tension between finding discrimination against women worse and the normative view that the gender of the victim should not affect their judgments.

SAMPLE

Our main analysis is based on a sample of 478 respondents recruited by Qualtrics. Respondents participated between 4 June 2020 and 30 June 2020. This main sample is representative of the population of US adults in terms of sex, age, education, and political orientation. Qualtrics achieved this representativeness by recruiting respondents whose characteristics match population statistics taken from the 2018 American Community Survey (for sex, age in bins[2], education, and race, see United States Census Bureau, 2021) and a May 1–13, 2020 Gallup survey (for political orientation, see Gallup, 2021). Representativeness targets were reached for all of these characteristics, except that the mean age in our sample is 3 years under the mean age of over-18 Americans, mainly due to undersampling of people in the over-65 age bin.

Table 1 shows summary statistics for our main estimation sample (based only on Qualtrics respondents). Respondents are on average 46 years old; 51% are female, 74% are White and 12% are Black; 38% have a high school degree or less and 12% have a graduate degree. The political leaning

[2]The age bins and sampling targets were: 18–24 years (12.4%), 25–34 years (17.9%), 35–44 years (16.35%), 45–54 years (17.1%), 60–64 years (7.9%), 65+ years (19.8%).

TABLE 1 | Summary statistics (N = 478).

	(1)	(2)	(3)	(4)	(5)
	Mean	***SD***	**Min**	**Max**	**Target mean**
Age	46.18	17.37	18	100	49.68
Female	0.51	0.50	0	1	0.51
Race					
White	0.74	0.44	0	1	0.74
Black	0.12	0.33	0	1	0.12
Asian or Pacific Islander	0.06	0.23	0	1	0.06
Amer. Indian or Alaska Native	0.01	0.10	0	1	0.01
Other	0.07	0.25	0	1	0.07
Educational attainment					
High school or less	0.38	0.49	0	1	0.40
Some college	0.23	0.42	0	1	0.23
Associates degree	0.08	0.28	0	1	0.08
Bachelor's degree	0.18	0.39	0	1	0.18
Graduate degree	0.12	0.32	0	1	0.11
Political orientation					
Democrat	0.31	0.46	0	1	0.31
Republican	0.28	0.45	0	1	0.28
Independent	0.36	0.48	0	1	0.37
Other	0.04	0.20	0	1	0.04

These summary statistics are based on our Qualtrics sample. Column (5) shows the target means for each variable which are based on a Gallup survey for political orientations and the 2018 American Community Survey for all other variables.

of the respondents is measured with the following Gallup question: "Generally speaking, do you think of yourself as a Republican, Democrat, Independent, or what?". 31% of respondents identify as Democrats, 28% as Republican, and 36% as Independents. The data and Stata do-files to create **Table 1** and all other results shown in our paper are available following this link: osf.io/2eq43.

Our Qualtrics sample is not representative of the US adult population along all dimensions. For example, some people in the US could not have made it into our sample because they do not speak English, or they do not have access to the internet. We therefore do not interpret our results as unbiased estimates of the relevant population parameters. However, having representativeness along several key dimensions gives us confidence that the direction of our point estimates will also hold in the general population.

To be able to test the robustness of our results, we additionally collected data from respondents recruited on Mturk. These respondents filled in a shorter version of the questionnaire, which included the same scenarios as in the Qualtrics survey but excluded some questions about beliefs and demographics. By shortening the questionnaire, we could stretch our research budget and increase the total number of Mturk respondents in our estimation sample to 1,169. This estimation sample excludes 13 respondents who indicated that they were less than 18 years old. Mturk respondents filled in a shorter version of the survey between 30 July 2020 and 22 August 2020. This sample is not representative of the US population. Respondents in our Mturk sample are on average 37 years old, 50% female and more educated than our main sample (see **Supplementary Appendix Table 1** for more summary statistics).

HOW DOES THE GENDER OF THE VICTIM AFFECT PEOPLE'S MORAL JUDGMENT ABOUT THE DISCRIMINATOR?

The Survey

Figure 1 shows the structure of the survey. In this section, we will describe the questions relating to the first part of our analysis. We will describe the questions relating to the second part of our analysis in Section "Predictions." The complete survey text is available in **Supplementary Appendix B**.

Judging Discrimination in Two Base Scenarios

In the first part of the survey, both Qualtrics and Mturk respondents were asked to judge discrimination in two scenarios ("Base questions"). These scenarios consist of situations in which a manager has to decide between giving the job to a man or a woman. In one scenario, the manager discriminates against the woman and in the other scenario the manager discriminates against the man. More specifically, the discrimination-against-the-woman scenario states that "*[t]aking into account all characteristics of the two applicants (qualifications, experience, personality, etc.), the manager knows that the* ***woman is slightly more qualified*** *and hiring her would bring slightly higher profits for the company. After considering everything,* ***the manager hires the man."*** (see screenshot of the whole scenario text in **Figure 2**). The discrimination-against-the-man scenario was identical except for the man being slightly more qualified and the manager hiring the woman. The order of those two scenarios was randomly assigned. For each scenario, respondents were asked to judge the manager's decision on a scale that ranges from 0 "very morally wrong" to 100 "very morally right".

For ease of interpretation, we center and reverse the judgments scores shown in this figure so that they range from −50 (very morally right) over 0 (neutral) to +50 (very morally wrong).

Follow-Up Questions to Clarify Judgments

After judging discrimination in the two base scenarios, a randomly selected 50% of respondents were asked to confirm their judgments from the first two scenarios. More specifically, the questionnaire showed a different follow-up question for each of the following three types of respondents: (1) Those who judged discrimination against women more negatively, (2) those who judged discrimination against women and men equally bad, and (3) those who judged discrimination against men more negatively. Each of these types of respondents was given the possibility to confirm their initial judgments. For example, respondents who judged discrimination against women more negatively saw the following text: *"Your evaluations of the manager's decisions in these two scenarios suggest that: You find it worse (from a moral perspective) if a manager hires a less qualified man over a more qualified woman (compared to the other way around)."* Respondents could then clarify their evaluations of the

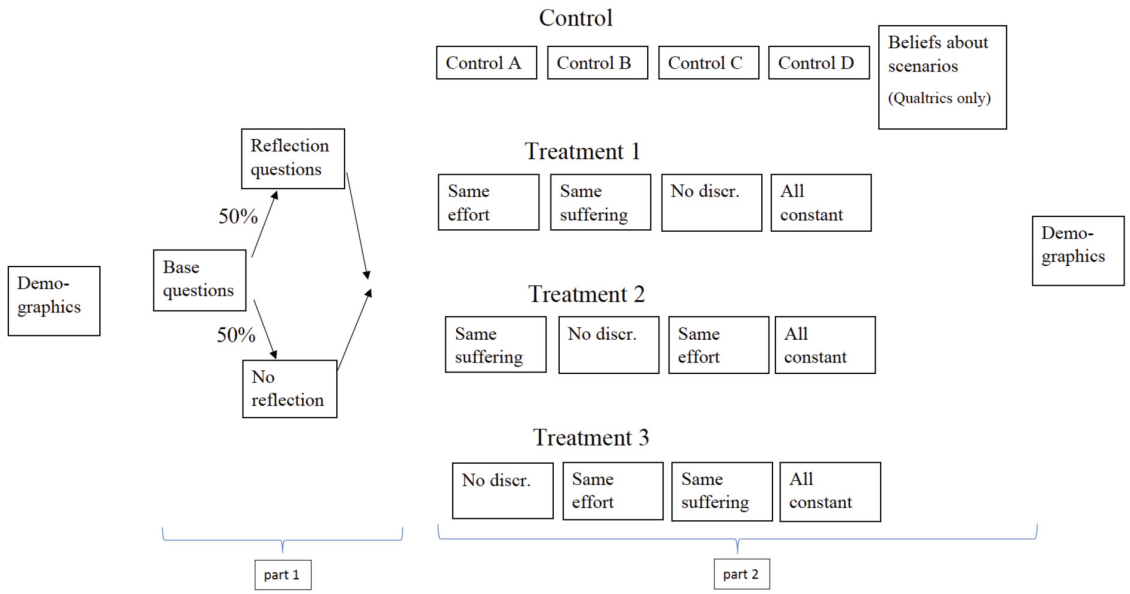


FIGURE 1 | Structure of survey.

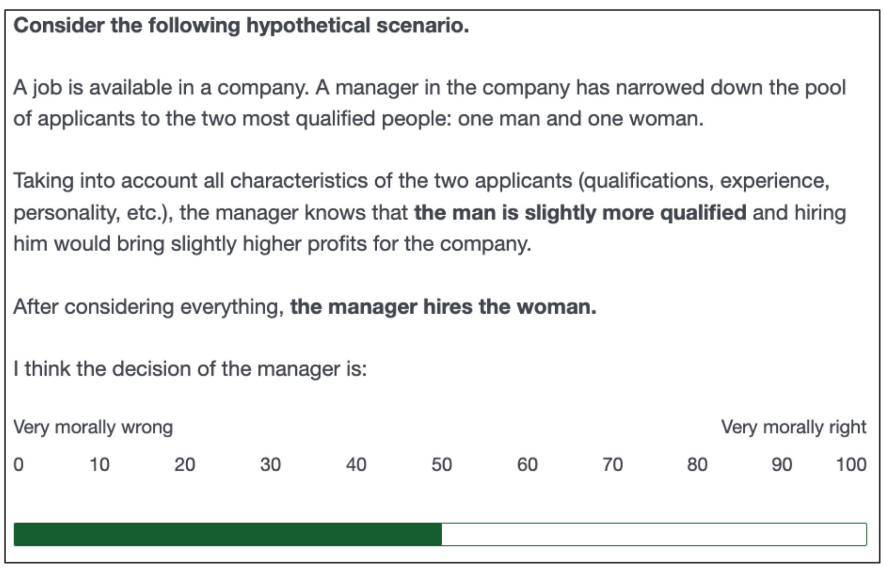


FIGURE 2 | Screenshot of discrimination-against-women base scenario.

first two scenarios by choosing one of the following three answer options: "*Yes, this is correct,*" "*No, I find both equally bad (or good),*" or "*No, I find it worse if the manager hires a less qualified woman over a more qualified man.*"

Measures of the Effect of the Victim's Gender on Judgments About Discrimination

We use two methods to measure the effect of the victim's gender on the judgment of the discriminator. Our first measure consists of the judgment of a manager who discriminates against a man minus the judgment of a manager who discriminates against a woman *within each respondent*. We will refer to this difference as within-subject pro-women attitude, or with the shorthand "pro-women attitude." For example, a respondent who judges discrimination against a woman with a score of 10 (somewhat morally wrong) and discrimination against a man with a score of 0 (neutral) has a pro-women attitude of 10 points. Negative values of this measure show pro-men attitudes. Besides computing the average pro-women attitude, we will also show their distribution. Based on the pro-women attitude in the base scenario, we classify

respondents as "pro-women" if their pro-women attitude is larger than 1 point, as "neutral" if their pro-women attitude is between –1 and +1 points, and as "pro-men" if their pro-women attitude is below –1 point.

Our second measure of the effect of the victim's gender relies on the judgment of the first scenario respondents saw (where the gender of the victim was randomized across subjects), which allows us to calculate the *between-subject pro-women attitude*. This measure is equal to the average judgment of managers who discriminate against a man minus the average judgment of managers who discriminate against a woman, *across different subjects*. Naturally, we cannot calculate this measure for individual respondents.

The difference between these two measures allows us to infer to what extent respondents themselves believe that the gender of the victim should not affect their judgment. The key to identifying this belief is that respondents can adjust their second judgment to be consistent with their first judgment. For example, assume that respondents find on average discrimination against women intuitively worse, but also believe that the victim of the gender should not affect their judgments. In this case the between-subject pro-women attitude will reveal the intuition of the respondents: in the first scenarios they see, respondents would judge a discriminator more harshly if the victim is a woman. However, if respondents also hold the normative view that the gender of the victim should not matter, they would adjust their second judgment to be consistent with their first judgment. If this adjustment is complete, we would observe no pro-women attitude with our within-subject measure. However, such adjustment would be revealed by order effects. Respondents who see scenarios describing discrimination against a woman first should judge the manager in *both* scenarios more harshly: in the first scenario because they feel discrimination against woman is particularly bad, and in the second because they feel compelled to judge discrimination similarly harshly if the victim is a man. Using such differences between within and between subject judgments is a tool commonly used in psychological studies to draw inferences about conflicting motivations (e.g. Uhlmann et al., 2009; Winegard et al., 2018).

TABLE 2 | Estimating the between-subject pro-women attitude and order effects (Qualtrics sample).

	(1)	(2)	(3)	(4)
	Woman discr. (1st judgment)	**Man discr. (2nd judgment)**	**Man discr. (1st judgment)**	**Woman discr. (2nd judgment)**
Av. judgment in each	13.3	9.0	1.5	8.0
Av. judgment in both (Col 1, 2 and Col 3, 4)	11.1		4.7	
Order effect	6.4			
Between-subject pro-women attitude	11.8			

All values refer to the reversed and centered judgment score. Higher values indicate finding discrimination more morally bad.

Results

Between-Subject and Within-Subject Pro-women Attitudes

When considering all judgments in the base scenarios, we see that respondents judge discrimination against a woman as 5.5 points more morally bad than discrimination against a man (10.5 points vs. 5.0 points). A one-sample *t*-test confirms that the average of this within sample pro-women attitude differs significantly from zero (p-value: < 0.001).

Table 2 shows the average judgment of the manager in scenarios in which a woman was discriminated and in which a man was discriminated, separately for respondents who randomly saw a discriminated woman (Columns 1 and 2) and man first (Columns 3 and 4). Focusing on the first judgments (Columns 1 and 3), we see a substantial between-subject pro-woman attitude. Respondents evaluate discrimination against a woman 11.8 points more morally wrong than discrimination against a man (13.3 points vs. 1.5 points). A two-sample *t*-test shows that these judgments are significantly different from each other (p-value < 0.001). When judging discrimination in isolation, respondents judge managers who discriminate against a woman substantially more harshly.

Table 2 also reveals that there are order effects. Despite seeing two identical scenarios, respondents judge the behavior of the manager in both scenarios on average 6.4 points more morally wrong if they first saw the scenario with the discriminated woman (11.1 points vs. 4.7 points). Following Uhlmann et al. (2009) and Winegard et al. (2018), we interpret this order effect as evidence that on average respondents themselves think that the gender of the victim should matter less than the between-subject pro-women attitude reveals. The within-subject pro-woman attitude therefore only shows the part of the pro-woman attitude which respondents are comfortable revealing (either to themselves or the researcher).

Figure 3 shows the distribution of the within-subject pro-women attitude. Based on this measure, we classify 38% of respondents as pro-women, 38% of respondents as neutral, and 24% of respondents as pro-men. Furthermore, pro-women respondents feel more strongly than pro-men respondents. On average pro-women respondents judge discrimination against women 22.4 points more morally bad, while pro-men respondents judge discrimination against men only 12.3 points more morally bad.

The distribution of the pro-women attitude shown in **Figure 2** partly reflects measurement error because not all respondents can precisely state their views using sliders in an online questionnaire. If such measurement error is random — and we believe that is most plausible — it should not affect the average pro-women attitude in our sample. However, it would increase the variance of our measure of the pro-women attitude. Random measurement error would also cause us to wrongly classify some respondents' views. Take, for example, a respondent who finds discrimination against women and men equally bad. Having just judged a

scenario in which a woman is discriminated, this respondent may not remember the exact position of the slider on the previous page and by chance judge discrimination against a man as worse. We would wrongly judge such a respondent as being pro-men.

Measurement error is not a concern when considering respondents' self-classifications. Based on answers to the follow-up question, the share of self-classified neutral respondents increases to 43%; leaving us with 34% self-classified pro-women and 22% self-classified pro-men respondents. Besides measurement error, the higher share of neutral respondents may also be triggered by the chance to reflect on their previous judgments.

Figure 4 shows different average values of pro-women attitude along the lines of gender, education, income, and political orientation. Women's average pro-women attitude is 6.4 points and men's average pro-women attitude is 4.4 points; the average pro-women attitude by level of education ranges from 3.2 points for respondents with a Bachelor's degree to 7.3 points for respondents with an Associate's degree; respondents earning below $50,000 have a very similar pro-women attitude to respondents who earn $50,000 + per year (5.8 points vs. 5.1 points). As expected, Democrats have 7.5 points stronger pro-women attitude than Republicans, but even the Republicans' pro-women attitude is positive (4.3 points). However, *F*-tests reveal that none of the aforementioned differences is statistically significant.

Replication With Mturk Sample

While the magnitudes differ, all our key results replicate with our non-representative Mturk sample. In this sample, respondents show on average a statistically significant pro-women attitude of 4.1 points (one sample *t*-test, p-value < 0.001). We classify 43% of respondents as pro-women, 30% as neutral, and 27% as pro-men. When giving respondents the chance to clarify their view, the percentage of neutral respondents increases to 37%; leaving 36% pro-women, and 27% pro-men respondents. The between-subject pro-women attitude is a statistically significant 6.2 points (two sample *t*-test, p-value < 0.001), which is substantially larger than the within-subject pro-women attitude of 4.1 points.

DOES STATISTICAL FAIRNESS DISCRIMINATION DRIVE RESPONDENTS' PRO-WOMEN ATTITUDE?

In our base scenarios, we stated that the manager hires a woman instead of a more productive man or vice versa. We neither gave reasons for the productivity difference nor mentioned explicitly that the two applicants are otherwise in identical situations. A plausible explanation for the pro-women attitude is hence that respondents have engaged in statistical fairness discrimination: Respondents may use the gender of the victim of discrimination as a signal for other unobserved characteristics of the situation which affect their judgment of the discriminator.

The Survey Experiment and Questions About Beliefs

We investigate the role of beliefs about unobserved characteristics using an embedded survey experiment. After judging the base scenarios, each respondent saw four additional pairs of scenarios that were again identical except for the victim's gender. Each pair of scenarios was shown on the same page allowing respondents to easily compare their judgments about managers who discriminate against a woman and managers who discriminate against a man.

Half of respondents were randomly assigned to the control group. These respondents saw scenarios that only differed from the base scenarios in the location of the job (urban area, suburban area, rural area, major city). We added this arguably irrelevant piece of information to avoid showing scenarios identical to the base scenario. The other half of respondents were randomly assigned to one of three treatment arms. Respondents in each treatment arm saw scenarios describing jobs in the same locations as the control group. Besides seeing the same locations, respondents in each treatment arm saw the same additional texts. The "same effort" text stated that the woman and man under consideration worked equally hard to get the job; the "same suffering" text stated that they would suffer equally from not getting the job; the "no discrimination" text stated that the job is in an industry without gender discrimination, and the "all constant" treatment combines the previous three texts. **Table 3** shows the exact wording of all treatment texts.

Figure 5 shows the structure of the experiment. For each of the treatment arms, the first three information treatments consisted of the "same effort," "same suffering," and "no discrimination" texts. The order of these texts differed between treatment arms to prevent order effects from driving our results. Each of these three information treatments appears in the first set of scenarios in one treatment arm, in the second set of scenarios in another treatment arm, and in the third set of scenarios in another treatment arm. All three treatment arms saw the "all constant" treatment last. We conducted the survey experiment with the Qualtrics and Mturk samples.

Respondents in the control group first completed the survey experiment and were then asked for their beliefs about the women and men in the previous scenarios. In particular, we asked to which extent the women or men in the previous scenarios (1) would have suffered more from not getting the job, (2) worked harder to get where they are in their career, (3) are generally more hard-working (in their career and other aspects of their life), and (4) would be more discriminated against in the labor market. We elicited these beliefs only in the Qualtrics sample.

Predictions

If statistical fairness discrimination drives respondents' within-subject pro-women attitude, we should see two patterns in the data. First, we should see that the information texts should lead to more gender-neutral judgments about discrimination compared to the control group. Thus, holding suffering and effort of both candidates constant as well as stating that the job is in an industry without gender discrimination should reduce the pro-women

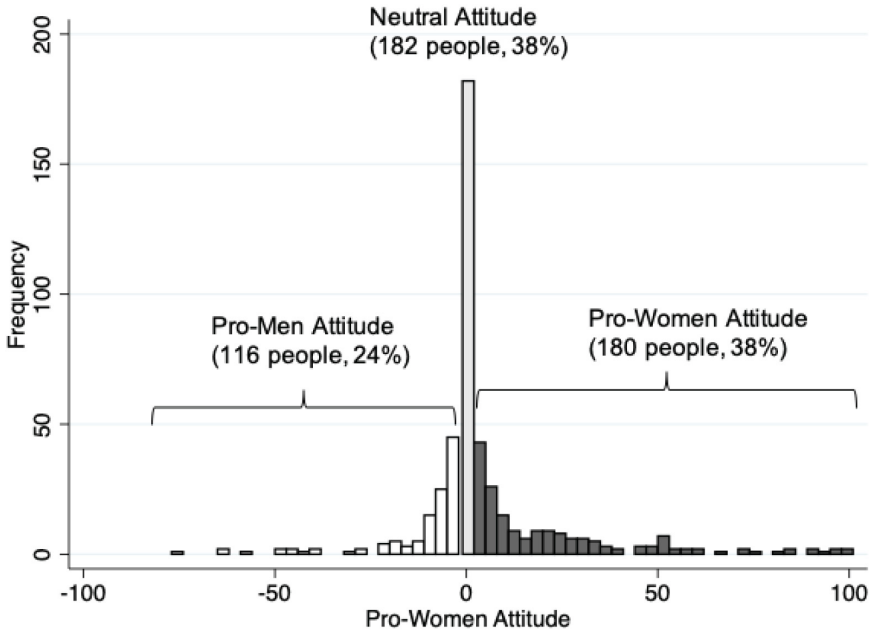


FIGURE 3 | Distribution of pro-women attitude. This figure shows the pro-women attitude of 478 respondents in the Qualtrics sample. The pro-women attitude of each respondent is the moral judgment about a manager who discriminates against a woman (on a –50 to +50 scale where higher values indicate more disapproval) minus the moral judgment about a manager who discriminates against a man. Positive values mean respondents judge discrimination against a woman as more morally bad.

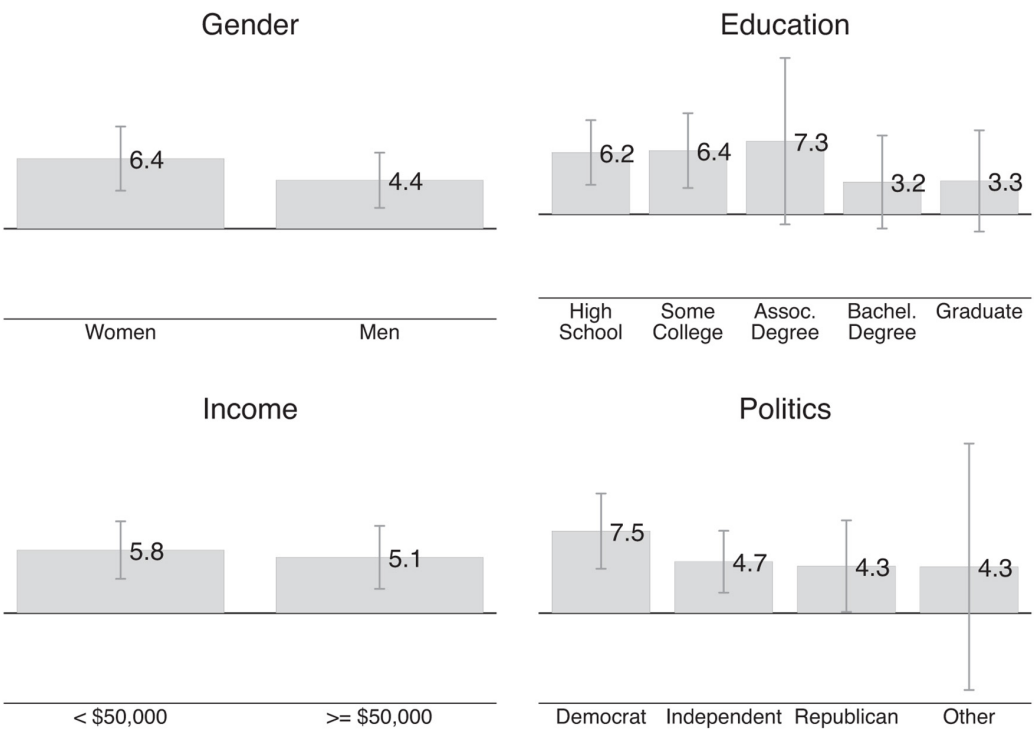


FIGURE 4 | Demographic predictors of pro-women attitude. This figure shows the average pro-women attitude for different groups in our Qualtrics sample. The pro-women attitude of each respondent is the moral judgment about a manager who discriminates against a woman (on a –50 to +50 scale where higher values indicate more disapproval) minus the moral judgment about a manager who discriminates against a man. Positive values mean that respondents judge discrimination against a woman as more morally bad.

TABLE 3 | Treatment text for survey experiment.

Treatment name	Treatment text
Same effort	**The man and the woman have worked equally hard in their career.** For example, both regularly studied on the weekends while their friends were out partying.
Same suffering	**The man and the woman would suffer equally much from not getting the job.** For example, both are currently unemployed, but have enough savings so that they could go without getting a paycheck for another 4 weeks. Also, both would find it equally hard to get a new job. Neither of them has to support a family.
No discrimination	**The job is in an industry where there is no gender discrimination.** A number of studies have convincingly shown that in this industry neither men nor women face discrimination in hiring decisions, nor do they face any other unfair treatment by coworkers or supervisors because of their gender.
All constant	**The man and the woman would suffer equally much from not getting the job, the man and the woman have worked equally hard in their career, and the job is in an industry with no gender discrimination.**

*We use **bold** here as we did in the actual survey.*

attitude of pro-women respondents and increase the pro-women attitude of pro-men respondents (i.e., reduce their *pro-men* attitude). The pro-women attitude of neutral respondents should not be affected.

Second, respondents that we classified based on their answers in the first part of the survey as pro-women and pro-men should believe that the women and men described in the scenarios differ along characteristics that make discrimination against their favored gender more objectionable. In particular, we would expect that pro-women respondents believe that the women described in the control scenarios would have worked harder in their career and in general, would suffer more from not getting the job, and would have suffered more discrimination. Pro-men respondents should hold beliefs in opposite directions.

Analysis of Survey Experiment

We analyze the results of the survey experiment by comparing the mean within-subject pro-women attitude between respondents in the control and treatment scenarios *separately for* pro-women, neutral, and pro-men respondents. More specifically, we estimate three separate regressions, one for each group of respondents. In each regression, the dependent variable is the within-subject pro-women attitude in the survey experiment. Independent variables are four treatment indicators (one for each information treatment), leaving the control group as our comparison group. The coefficients of the treatment indicators show the mean differences between the pro-women attitude in a given treatment compared to the control group. For example, the "same effort" coefficient shows the difference between the mean pro-women attitude in the control group (across all four scenarios) and the mean pro-women attitude in the scenarios which contained the "same effort" text (the first scenario in Treatment 1, the third scenario in Treatment 2, and the second scenario in Treatment 3). For all three regressions, we cluster our standard errors at the individual level. This way of clustering accounts for the fact that we observe multiple pro-women attitudes for each respondent.

We report our results by showing the mean pro-women attitude in the control group and each treatment group, again separately for all three groups of respondents. These means directly relate to our regression coefficients. For the control group, the mean pro-women attitude is equal to the constant. For the treatment groups, the means are equal to the constant plus the respective treatment coefficient.

Results

Effects of Information Treatments on Pro-women Attitude

Figure 6 shows the results of the survey experiment for the Qualtrics sample. The gray bars show the average between-subject pro-women attitude in the control group and for the various information treatments; separately for respondents who we classified, based on their responses to the base scenarios, as pro-women, neutral, or pro-men.

Overall, we see no meaningful or statistically significant effect for any of our three groups of respondents. Pro-women respondents show an average pro-women attitude of 6 points in the control scenarios and an almost identical pro-women attitude of 6.1 points in scenarios which held effort and suffering of the male and female applicant constant as well as describing a job in an industry without gender discrimination (All constant treatment). While point estimates differ slightly, we also see no evidence of an effect in any of the other information treatments: holding suffering, effort or discrimination individually constant has no meaningful effect on pro-women attitude. Three of the four point estimates (same suffering, no discrimination, all constant) even suggest that the information treatments increased respondents' pro-women attitude. These results go against our predictions.

For neutral respondents, we also see no impact of any of the information treatments. They show an average pro-women attitude of 0.7 points in the control scenarios, which is almost identical to the pro-women attitude of 0.6 points in the scenarios which held effort, suffering and discrimination constant. None of the average pro-women attitudes are significantly different from zero. These results are in line with our predictions. Holding reasons for finding discrimination against one gender more objectionable constant does not affect the pro-women attitude of respondents who already judged discrimination against women and men as equally bad.

For pro-men respondents, we also see no significant changes in their pro-women attitude in response to any of the information treatments. This result is driven by the control group. Respondents with an initial pro-men attitude in the base scenarios who were randomly assigned to the control group now find discrimination against women slightly more bad: They show a positive pro-women attitude of 1.7 points. Neither this pro-women attitude nor any of the pro-women attitudes in the treatment scenarios are significantly different from zero. These results are again inconsistent with our predictions.

What could be driving the increase of the pro-women attitude (i.e., reduction of pro-men attitude) of respondents who we initially classified as pro-men? Part of this increase is likely be driven by regression to the mean. Some respondents

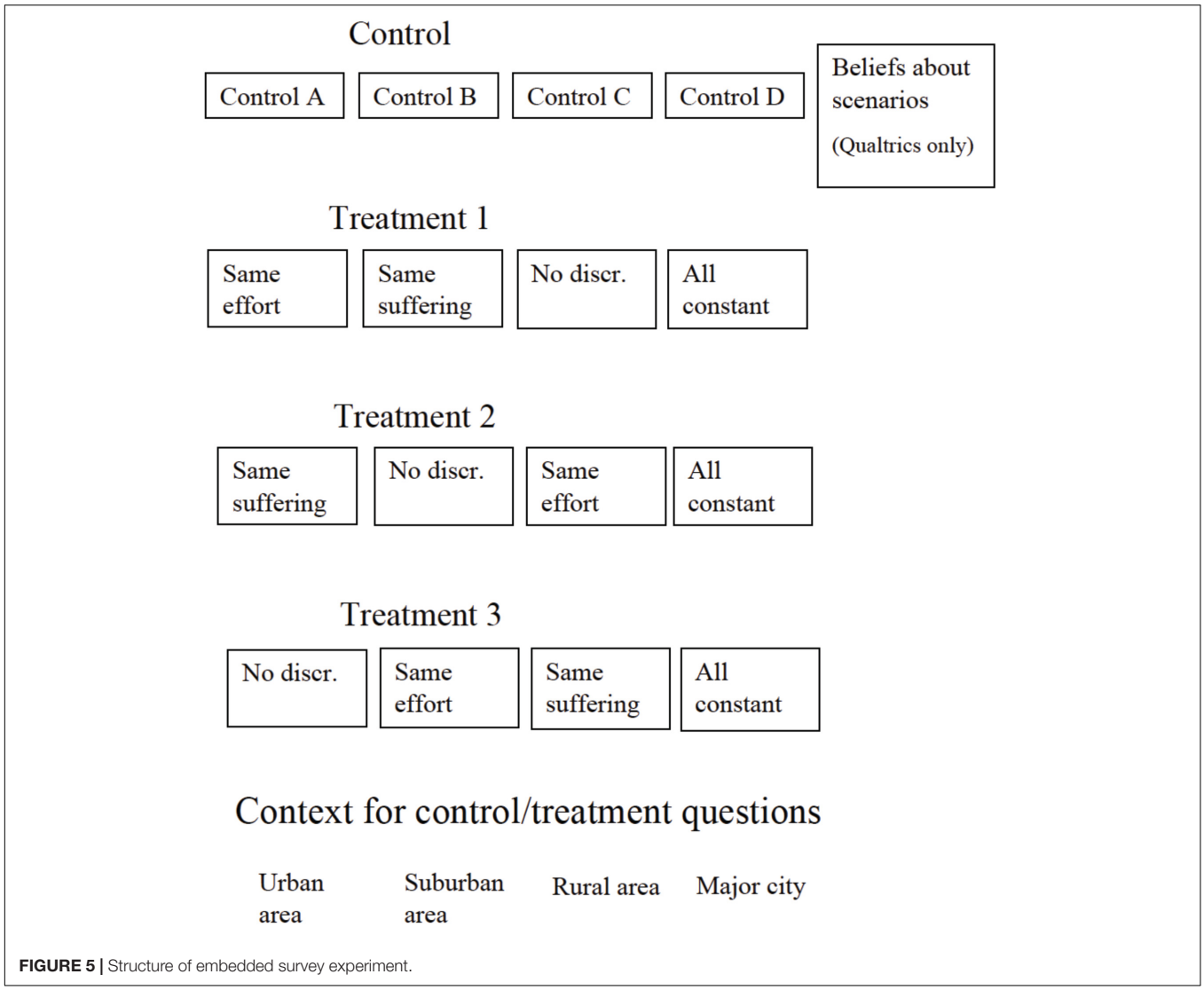


FIGURE 5 | Structure of embedded survey experiment.

classified as pro-men may in fact be neutral or pro-women but have, by chance, moved the slider to indicate that they find discrimination against men more morally wrong. When evaluating similar subsequent scenarios, those respondents may have on average reverted back to their true value of pro-women attitude.[3] It may also be that these respondents feel increasingly uncomfortable of revealing their pro-men attitudes to us as the researchers. Whatever the reasons for this reduction are, pro-men respondents' pro-men attitude from the base scenario is not stable. In subsequent scenarios in the control and treatment groups, their average pro-women attitudes are not statistically distinguishable from zero.

Besides using three separate samples, we also estimate the effects of the information treatments in one fully interacted model. We regress pro-women attitude in the survey experiment on one dummy variable for each information treatment, respondents' pro-women attitude in the base scenarios, and four interaction terms of the information treatment times respondents' pro-women attitude in the base scenarios (e.g., same-effort-X-pro-women-base). This model allows the effect of the information treatment to depend on respondents' initial pro-women attitude in a more fine-grained way. While with using three separate samples we allowed for the effect of the information treatment to differ between each group of respondents, a model with interaction terms allows for the effect to be larger within each group as well. If respondents engage in statistical fairness discrimination, we would expect the coefficients of the interaction terms to be negative to capture that the effects of the information treatments are more negative for respondents who show a larger pro-women attitude in the base scenarios.

Column (1) of **Supplementary Appendix Table 2** shows that none of the main effects of the information treatments nor any of the interaction terms are statistically significant. Furthermore,

[3]Regression to the mean can also explain the decrease of pro-women attitude of the pro-women respondents from 22.4 points in the base scenario to 6 points in the survey experiment.

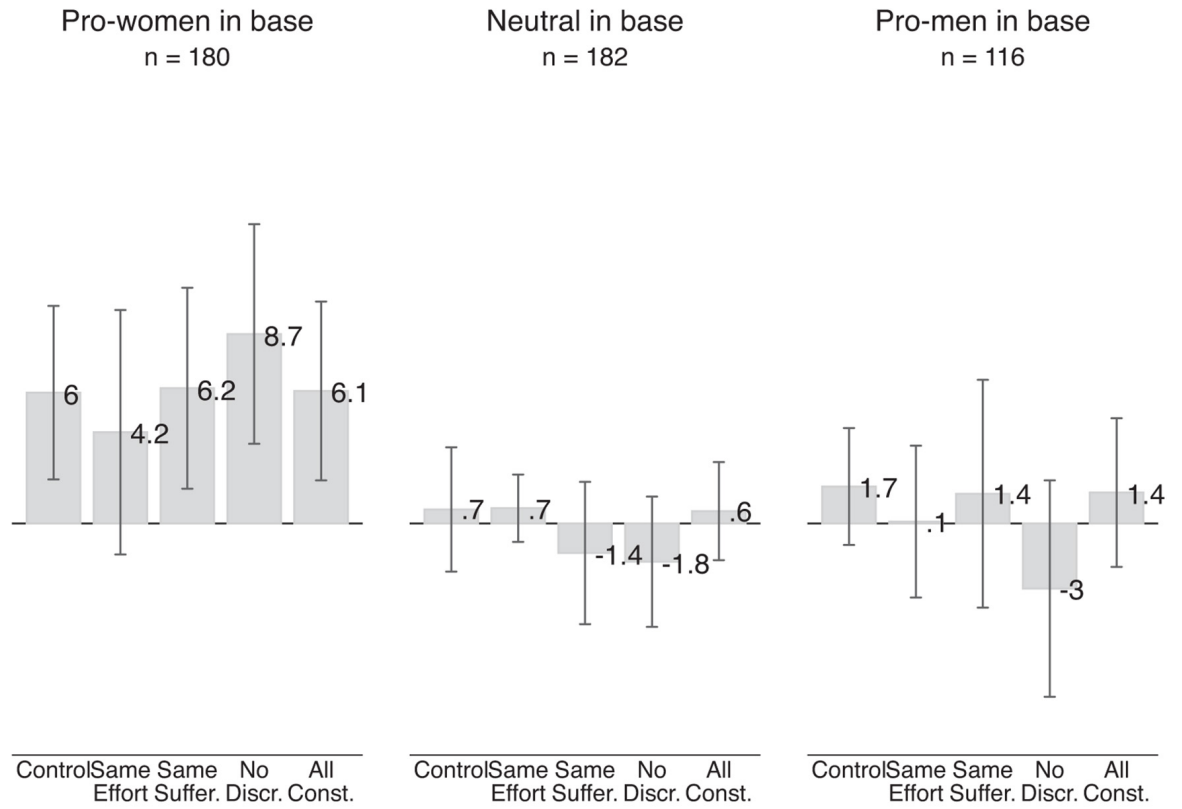


FIGURE 6 | Effect of information treatments on pro-women attitude (Qualtrics sample). The gray bars show the average pro-women attitude in a given treatment. The left-most bar is based on 688 judgments made by 86 respondents who were pro-women in the base scenarios and were randomly assigned to the control group, judging four pairs of scenarios involving two judgments each. Each of the next four bars is based on 188 judgments made by 94 respondents who were pro-women in the base scenarios and were randomly assigned to any of the three treatment groups. Similarly, the left-most bar in the middle section is based on 744 judgments by 93 respondents who were neutral in the base scenario, and each of the next four bars is based on 178 judgments made by 89 respondents. The left-most bar in the right-most section is based on 480 judgments made by 60 respondents who were pro-men in the base scenario, and each of the next four bars is based on 112 judgments by 56 respondents. Vertical lines show 95 percent confidence intervals based on standard errors clustered at the individual level.

the *F*-test for joint significance does not reject the null hypothesis that all four included interaction terms are equal to zero (*p*-value: 0.1360). Also with this way of estimating the effect of the information treatments, we see no evidence for statistical fairness discrimination.

Beliefs About Discriminated Women and Men

The gray bars in **Figure 7** show the beliefs of the control group's respondents about the women and men in the scenarios separately for pro-women, neutral, and pro-men respondents. The numbers reported in the figure are based on scales that range from −50 points to +50 points, where 0 indicates gender neutrality and higher values that the statements shown in the figure apply more to women.

The beliefs about gender differences are weak and often not statistically significant. However, the direction of the point estimates is broadly consistent with our predictions: Pro-women respondents believe that the women in the scenarios would suffer more from not getting the job, worked harder in their career (but not in general), and would have suffered more discrimination. Pro-men respondents believe that men would suffer more from not getting the job, and worked harder in their career and in general. However, they also believe that women suffered more discrimination.

Replication With Mturk Sample

While we did not elicit their beliefs, we did run the survey experiment with Mturk respondents. **Figure 8** shows the results of this survey experiment.

In contrast to our results with the Qualtrics sample, we do see significant treatment effects for pro-women respondents. In the control scenarios, the average pro-women attitude of respondents who we classified as pro-women based on their answer to the base scenarios is 9.2 points. In the treatment scenarios, the pro-women attitude is between 3.1 points and 5.0 points lower. These differences are statistically significant at the 5 percent level for the "same effort" scenarios, "all constant" scenarios, and significant at the 1 percent level for the "no discrimination" scenarios.

Our results for neutral and pro-men respondents are similar to the results in the Qualtrics sample. We see no significant effect of the information treatment for either group of respondents. The

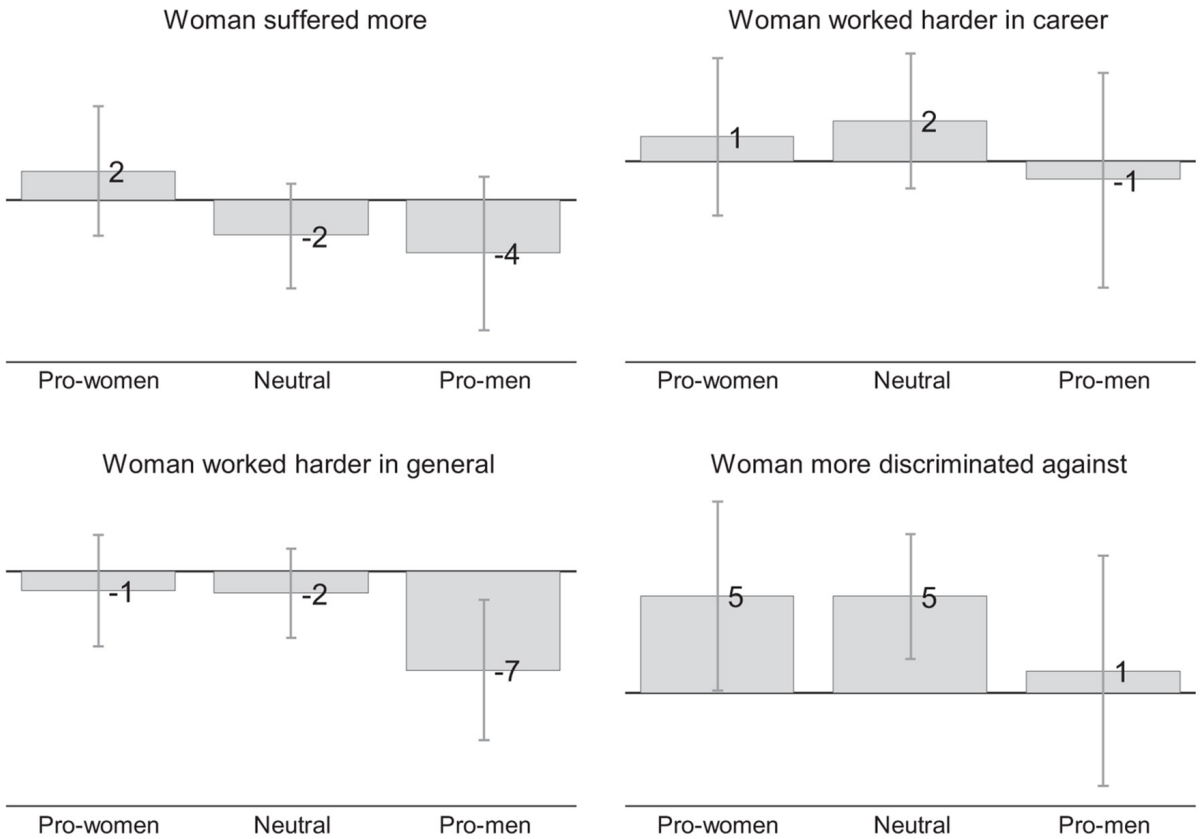


FIGURE 7 | Beliefs about women and men in control scenarios (Qualtrics). This figure is based on responses of Qualtrics respondents who were randomly assigned to the control group. The numbers shown in the figure show belief scores which can range from –50 to +50 where 0 indicates gender neutrality and positive values mean that the statement in the subfigure heading applies more to women. For example, positive values in the left-most figure mean that respondents believe that the women showed in the previous scenarios in the survey experiment would suffer more from not getting the job while negative values mean the opposite. Vertical lines show 95 percent confidence intervals.

absence of the treatment effect for pro-men respondents is again driven by an increase in the pro-women attitude (i.e., a reduction in pro-men attitude) in the control group.

When we estimate the effect of the information treatments using one fully interacted model we see that the effect of the "all constant" treatment is significantly more negative for respondents who showed a larger pro-women attitude in the base scenario (see Column 2 of **Supplementary Appendix Table 2**). Also with this empirical approach we find some evidence that Mturk respondents engage in statistical fairness discrimination.

Summary of Results and Discussion

The results of the survey experiment only present mixed evidence for the statistical fairness discrimination explanation. In our main sample, we see that holding constant additional information on characteristics that may explain gender differences in deservingness does not significantly affect respondents' pro-women attitude. In our replication sample we find effects that are statistically significant and go in the expected direction for pro-women respondents but not for pro-men respondents.

The difference in the effect of the information treatment between the two samples might be driven by differences in underlying beliefs. In the Qualtrics sample, respondents' beliefs showed that gender of the victim may not have been a useful signal for inferring applicants' deservingness. While pro-women respondents in this sample believe that the woman (compared to the man) described in the scenarios would suffer more from not getting the job, worked harder in their career, and would have suffered more discrimination, the magnitude of these differences is small. It is therefore not surprising that explicitly holding those factors constant did not have much of an effect on respondents' pro-women attitude. In the Mturk sample, the meaningful treatment effects might have been driven by gender being a stronger signal for candidate's deservingness. For example, the significant effect of the "same effort" treatment might be driven by pro-women respondents in the Mturk sample believing that the women described in the scenarios worked much harder in their career.

Differences in beliefs between samples and contexts could also explain why our results differ from those reported by Cappelen et al. (2019). In their context, subjects may believe that men who lose have simply not worked hard enough and are therefore less deserving of benefiting from redistribution. The experimental manipulation of determining winners and losers by chance may

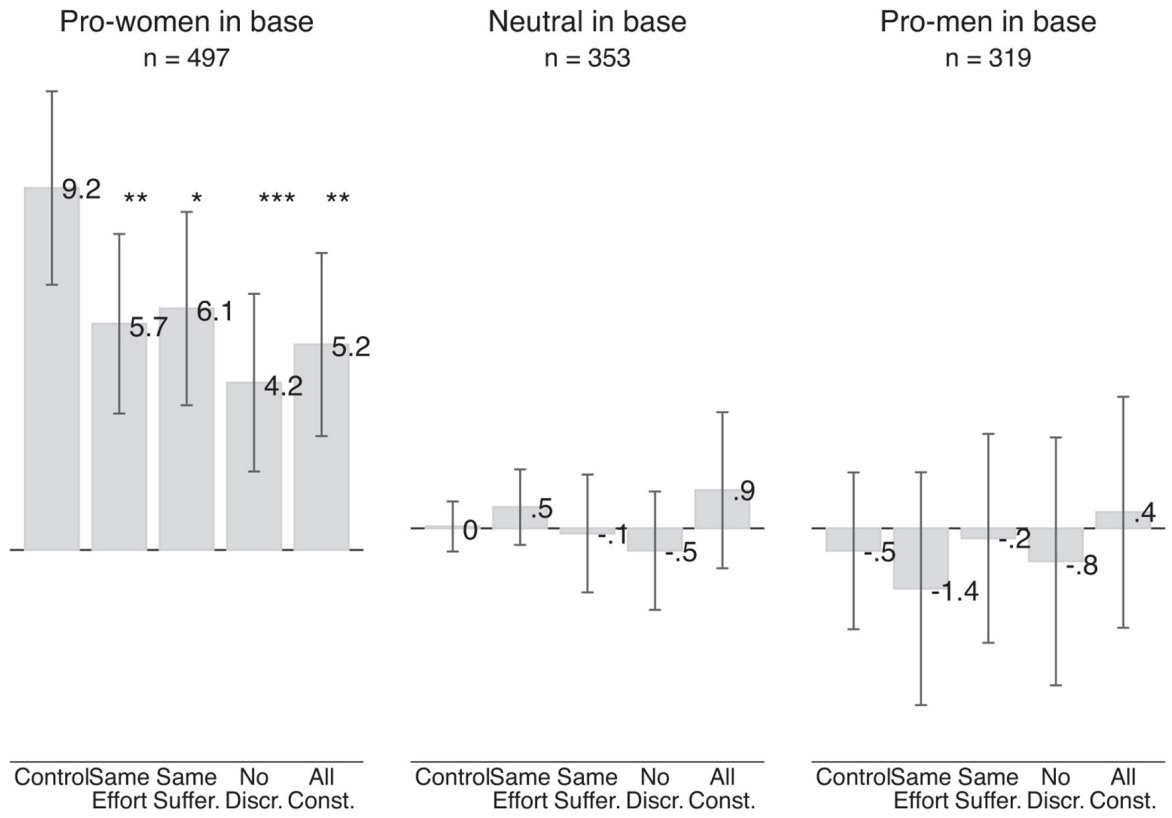


FIGURE 8 | Effect of information treatments on pro-women attitude (Mturk sample). The gray bars show the average pro-women attitude in a given treatment. Each bar in the control group shows the average pro-women attitude for pro-women, neutral and pro-men respondents, respectively. The left-most bar is based on 2,144 judgments made by 268 respondents who were pro-women in the base scenarios and were randomly assigned to the control group, judging four pairs of scenarios involving two judgments each. Each of the next four bars is based on 458 judgments made by 229 respondents who were pro-women in the base scenarios and were randomly assigned to any of the three treatment groups. Similarly, the left-most bar in the middle section is based on 340 judgments by 170 respondents who were neutral in the base scenario, and each of the next four bars is based on 366 judgments made by 183 respondents. The left-most bar in the right-most section is based on 340 judgments made by 170 respondents who were pro-men in the base scenario, and the next four are based on 298 judgments by 149 respondents. Vertical lines show 95 percent confidence intervals based on standard errors clustered at the individual level. Statistical significance of the difference between control group and scenarios with different information treatments is denoted by ***p-value < 0.01, **p-value < 0.05, and *p-value < 0.10.

have affected people's decision by ruling out this reason for treating men and women differently.

While we believe that statistical fairness discrimination matters, using gender to draw inferences about deservingness is unlikely to be the only reason for differences in judgments about discriminated women and men. In both of our samples, we still see significant levels of pro-women attitude in scenarios for which we have explicitly held suffering, effort *and* discrimination constant (Qualtrics sample: 6.1 point, Mturk sample 5.2 points). While we cannot rule out that the remaining pro-women attitude is completely driven by beliefs about other unobserved characteristics of the victim, we do not think this is plausible. Instead, we find it more likely that respondents judge discriminators according to factors other than the victim's deservingness, such as the inferred intentions of the manager. For example, respondents may have assumed that a manager who discriminates against women may have bad intentions (e.g., sexism) whereas a manager who discriminates against men may have good intentions (e.g., increasing gender equality).

CONCLUSION

We have shown that even in apparently identical scenarios people judge discrimination against women less harshly than discrimination against men. We have further investigated to what extent this gender gap is driven by what Cappelen et al. (2019) have termed "statistical fairness discrimination." Our results only lend mixed support for this mechanism: the use of gender as a signal for the victim of discrimination's deservingness is unlikely to account for the whole pro-women attitude. However, the victim of the gender may have been used as a signal for other relevant characteristics such as the intention of the discriminator.

All our results and conclusions are based on variations of the same generic scenario. We hope that future research establishes to what extent people show a pro-women attitude in other scenarios as well. Some factors that might affect the pro-women attitude are the gender of the manager (which we did not specify), whether the job is in a predominantly male or female industry, and the social status of the job.

The concept of statistical fairness discrimination also inspires promising avenues for future research. Could this practice, for example, explain why people give less harsh sentences for women than men who committed similar crimes (Shatz and Shatz, 2012)? If it does, what are the differences in beliefs that are driving this? To answer these and related questions, researchers could follow our approach of measuring beliefs about unobserved gender differences and randomly holding additional information constant.

DATA AVAILABILITY STATEMENT

The data and Stata do-files to replicate our results are made available at osf.io/2eq43.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Victoria University of Wellington Human Ethics Committee. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

EF, JF, and SN designed the experiment and wrote the manuscript. SN programmed the survey, collected and cleaned the data. JF did the analysis. All the authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We are grateful for comments from participants of the New Zealand e-Seminar Series.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.675776/full#supplementary-material

REFERENCES

Allen, T. D., French, K. A., and Poteet, M. L. (2016). Women and career advancement: issues and opportunities. *Organ. Dynam.* 45, 206–216. doi: 10.1016/j.orgdyn.2016.07.006

Arrow, K. J. (1973). "The theory of discrimination," in *Discrimination in Labor Markets*, eds O. Ashenfelter and A. Rees (Princeton: Princeton University Press).

Autor, D., and Wasserman, M. (2013). *Wayward Sons: the Emerging Gender Gap in Labor Markets and Education*. Washington, D.C.: Third Way.

Azmat, G., and Petrongolo, B. (2014). Gender and the labor market: what have we learned from field and lab experiments? *Labour Econ.* 30, 32–40. doi: 10.1016/j.labeco.2014.06.005

Babcock, L., Recalde, M. P., Vesterlund, L., and Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *Am. Econ. Rev.* 107, 714–747. doi: 10.1257/aer.20141734

Bertrand, M. (2011). "New perspectives on gender," in *Handbook of Labor Economics*, eds C. David and O. Ashenfelter (Amsterdam: Elsevier).

Binder, A. J., and Bound, J. (2019). The declining labor market prospects of less-educated men. *J. Econ. Perspect.* 33, 163–190. doi: 10.1257/jep.33.2.163

Blau, F. D., and Kahn, L. M. (2017). The gender wage gap: extent, trends, and explanations. *J. Econ. Literat.* 55, 789–865.

Block, K., Croft, A., De Souza, L., and Schmader, T. (2019). Do people care if men don't care about caring? the asymmetry in support for changing gender roles. *J. Exp. Soc. Psychol.* 83, 112–131. doi: 10.1016/j.jesp.2019.03.013

Bohnet, I., Van Geen, A., and Bazerman, M. (2016). When performance trumps gender bias: joint vs. separate evaluation. *Manag. Sci.* 62, 1225–1234. doi: 10.1287/mnsc.2015.2186

Bohren, J. A., Imas, A., and Rosenberg, M. (2019). The dynamics of discrimination: theory and evidence. *Am. Econ. Rev.* 109, 3395–3436. doi: 10.1257/aer.20171829

Booth, A., and Leigh, A. (2010). Do employers discriminate by gender? a field experiment in female-dominated occupations. *Econ. Lett.* 107, 236–238. doi: 10.1016/j.econlet.2010.01.034

Breda, T., and Hillion, M. (2016). Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science* 353, 474–478. doi: 10.1126/science.aaf4372

Cappelen, A. W., Falch, R., and Tungodden, B. (2019). *The Boy Crisis: Experimental Evidence on the Acceptance of Males Falling Behind*. Norway: Institutt for samfunnsøkonomi. NHH Dept of Economics Discussion Paper No. 06/2019.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø, and Tungodden, B. (2007). The pluralism of fairness ideals: an experimental approach. *Am. Econ. Rev.* 97, 818–827. doi: 10.1257/aer.97.3.818

Chan, C. K., and Anteby, M. (2016). Task segregation as a mechanism for within-job inequality: women and men of the transportation security administration. *Administrative Sci. Q.* 61, 184–216. doi: 10.1177/0001839215611447

Coffman, K. B., Exley, C. L., and Niederle, M. (2018). *When Gender Discrimination is Not About Gender*. Boston: Harvard Business School. doi: 10.1177/0001839215611447

De Pater, I. E., Van Vianen, A. E., and Bechtoldt, M. N. (2010). Gender differences in job challenge: a matter of task allocation. *Gender Work Organ.* 17, 433–453.

Eagly, A. H., and Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *Eur. Rev. Soc. Psychol.* 5, 1–35. doi: 10.1080/14792779543000002

Eckel, C. C., and Grossman, P. J. (2008). "Differences in the economic decisions of men and women: experimental evidence," in *Handbook of Experimental Economics Results*, eds C. R. Plott and V. L. Smith (Amsterdam: Elsevier).

Engel, C. (2011). Dictator games: a meta study. *Exp. Econ.* 14, 583–610. doi: 10.1007/s10683-011-9283-7

Eurofund (2018). *Discrimination Against Men at Work: Experiences in Five Countries*. Luxembourg: Publications Office of the European Union.

Feess, E., Jan, F., and Shakked, N. (2021). *Attitudes towards hiring decisions. AEA RCT Registry*. Available online at: https://doi.org/10.1257/rct.5064-3.0 (accessed January 21, 2021).

FeldmanHall, O., Dalgleish, T., Evans, D., Navrady, L., Tedeschi, E., and Mobbs, D. (2016). Moral chivalry: gender and harm sensitivity predict costly altruism. *Soc. Psychol. Personal. Sci.* 7, 542–551. doi: 10.1177/1948550616647448

Gallup (2021). *Party Affiliation*. Washington, DC: Gallup.

Gupta, V. K., Mortal, S. C., Silveri, S., Sun, M., and Turban, D. B. (2020). You're fired! gender disparities in CEO dismissal. *J. Manag.* 46, 560–582. doi: 10.1177/0149206318810415

Haaland, I., and Roth, C. (2021). *Beliefs About Racial Discrimination and Support for Pro-Black Policies*. Coventry: University of Warwick, Department of Economics. CESifo Working Paper.

Knowles, J., Persico, N., and Todd, P. (2001). Racial bias in motor vehicle searches: theory and evidence. *J. Political Economy.* 109, 203–229. doi: 10.1086/318603

Kunze, A. (2018). *The Gender Wage Gap in Developed Countries. Oxford Handbook Women Economy*, Vol. 369. Available online at: https://books.google.co.nz/books?hl=en&lr=&id=7GdaDwAAQBAJ&oi=fnd&pg=PA369&dq=Kunze,+A.+(2018).+The+gender+wage+gap+in+developed+countries.+Oxford+Handb.+Women+Economy+369.&ots=4xowYz0Fi9&sig=D5419h_3D4GJWAhhGlWy0ktDfTY#v=onepage&q=Kunze%2C%20A.%20(2018).%20The%20gender%20wage%20gap%20in%20developed%20countries.%20Oxford%20Handb.%20Women%20Economy%20369.&f=false

Niederle, M. (2017). A gender agenda: a progress report on competitiveness. *Am. Econ. Rev.* 107, 115–119. doi: 10.1257/aer.p20171066

OECD (2020). *Education at a Glance 2020*. Paris: OECD.

Olivetti, C., and Petrongolo, B. (2016). The evolution of gender gaps in industrialized countries. *Annual Rev. Econ.* 8, 405–434. doi: 10.1146/annurev-economics-080614-115329

Ortiz-Ospina, E., and Roser, M. (2018). *Economic inequality by gender*.28October2019 Available online at: from https://ourworldindata.org/economic-inequality-by-gender (accessed 28 October, 2019)

Phelps, E. S. (1972). The statistical theory of racism and sexism. *Am. Econ. Rev.* 62, 659–661.

Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in science. *Proc. Natl. Acad. Sci. U S A.* 111, 4403–4408. doi: 10.1073/pnas.1314788111

Shatz, S. F., and Shatz, N. R. (2012). Chivalry is not dead: murder, gender, and the death penalty. *Berkeley J. Gender L Just.* 27:64.

Stewart-Williams, S., Chang, C. Y. M., Wong, X. L., Blackburn, J. D., and Thomas, A. G. (2021). Reactions to male-favouring versus female-favouring sex differences: a pre-registered experiment and Southeast Asian replication. *Br. J. Psychol.* 112, 389–411. doi: 10.1111/bjop.12463

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., and Ditto, P. H. (2009). The motivated use of moral principles. *Judgment Dec. Mak.* 4, 479–491.

United States Census Bureau (2021). *American Community Survey*. Suitland, MD: United States Census Bureau.

Winegard, B., Clark, C., Hasty, C. R., and Baumeister, R. (2018). *Equalitarianism: A Source of Liberal Bias*. Available online at: https://ssrn.com/abstract=3175680 (accessed May 8, 2018).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership