Essays on unethical behaviour

Pablo Ignacio Soto Mota

A thesis presented for the degree of

Philosophiae Doctor (Ph.D.)

Department of Economics

NHH, Norwegian School of Economics

Bergen, Norway

2023

Acknowledgements

My path during my PhD has been one with many mentors, collaborators, and friends. I want to thank everyone who influenced my growth during this time. I cannot name them all, but many were and still are very important, so an attempt is necessary.

First and foremost, I deeply thank my supervisor, Mathias Ekström. I benefited greatly from his constant attention and kind feedback during the whole research process, his patience, and advice. I always felt fortunate to have a supervisor who gave me the right balance between complete independence in pursuing research ideas and close supervision to develop them. I also want to express my gratitude to Bertil Tungodden. He has been a constant source of inspiration and encouragement since I started my PhD. I admire his energy to conduct excellent academic research while never forgetting to be kind and to see the human aspect of the job. Thank you, Mathias and Bertil, for teaching me and helping me navigate the sometimes challenging but always rewarding waters of research and academic life.

As with all good things in life, science is not done in isolation. I am fortunate to have had great co-authors in the papers presented in this dissertation. I want to thank Hallgeir Sjåstad, Mathias Ekström, and Kjetil Bjorvatn, my co-authors in the paper presented in the first chapter. Their expertise helped me to grasp the art of producing the high-quality research I aim to continue doing as an academic. I am also grateful to Luis Adrián Soto-Mota, Andrés Castañeda, and Alfonso Gulias, the doctors behind my "doctors project". I will always feel in debt for their support in helping me implement a field experiment in several hospitals during a pandemic. This experience taught me the few challenges and the many joys of multidisciplinary research in complex contexts. Likewise, I want to express my gratitude to Adrian Vargas-López, who, besides being a co-author, is a dear friend. Our constant Zoom calls for friendship and work were highlights of my PhD journey and I hope our academic paths will continue to cross.

In addition to my wonderful co-authors, the other key components in the work presented in this thesis are the Department of Economics and FAIR. Even before arriving in Norway, I felt included in a supportive, kind, and collaborative environment. I owe much to everyone at FAIR for the many talks, opportunities to present research ideas in early phases, and simple coffee breaks. In particular, I want to mention Akshay Moorty, Joel Berge, Andre Lot, Eirik Berger, and Rene Karadakic.

Also, I could not have done this without the support of Arne Nasgowitz, Aleksander Glapiak, Ricardo Gaxiola, Alessandro Pizzigoloto, and Adrián Soto-Mota. They are my best friends in the world and played a crucial role in keeping my emotional energy up during the most difficult parts of this journey. I am equally grateful to my friends in St. Paul Parish: Fides, the Sisters of Charity, and Father Erik. They are my extended family here in Norway and made me discover in this country a new home.

I checked some theses before, and, for some reason, sections like this usually finish with those who should be named first. I do not understand it, but I will not challenge the convention. I dedicate this thesis to my parents, Adrián and Maru. Since my earliest memories, they have given me unconditional and selfless support to pursue my dreams, no matter where they take me. They have always been there, providing encouragement and love. Also, I need to mention Adrian Soto-Mota once again. The reason is simple: he is simultaneously my co-author, my friend, and my brother…and he is also my doctor.

Finally, and therefore most importantly, I thank and dedicate this thesis to my wife, Suzana. You make my life complete and my days happy. Only with you, this and any other effort are possible and make sense.

Pablo.
Bergen, Norway
2023

Contents

Introduction

Unethical behaviours are costly to society. For example, theft and fraud obstruct firms' performance (Cialdini et al., 2004; Creighton et al., 2015); civil dishonesty facilitates corruption worldwide (Cohn et al., 2019; Sulitzeanu-Kenan et al., 2022); violent crimes affect the accumulation of human capital (Brown & Velásquez, 2017); and bullying hinders educational efforts (Brown & Taylor, 2008; Adam & Lawrence, 2011). Therefore, understanding what influences (un)ethical decisions is a key question to economists and the social sciences more broadly. By identifying the factors that enable and shape these behaviours, we can design more effective policies to mitigate their negative consequences.

During the last decades, the literature in Economics on unethical behaviour has broadened significantly (Jacobsen et al., 2018; Gerlach et al., 2019). Current studies assume a more nuanced understanding of human morality than the simple rational-actor models that disregard moral concerns (Fehr & Schmidt, 2006; Fischbacher & Gächter, 2010). Using field and laboratory experiments, previous studies have, for example, explored how monetary incentives, contextual factors, social norms, and institutions affect moral behaviour. The main insight in the literature is that individuals constantly face trade-offs between selfish and moral concerns, and for some individuals moral concerns trump self-interest (Kahneman et al., 1986). For example, people share money to maintain some fairness ideal (Cappelen et al., 2007); people are unwilling to lie although it would maximize their earnings (Abeler et al., 2019); and cooperate much more with others than what is predicted by the rational-actor model (Fehr & Gächter, 2000).

This dissertation adds to the current behavioural and experimental literature on unethical behaviour. In particular, the following chapters present studies that inform debates on whether promises can commit people towards honesty, the link between dishonesty and image concerns, and how groups shape the moral behaviour of individuals. While the chapters address different questions, they share a methodological approach. All three articles employ experiments to answer research questions about moral decision making. Two of them, presented in chapters 1 and 3, use online experiments, while chapter 2 is based on a field experiment. Taken together, the following chapters provide evidence that both selfish and moral concerns influence individual decision-making. Moreover, all three chapters provide results that point directly to public policies that may reduce the harmful consequences of unethical actions in society.

Chapter 1: Making a promise increases the moral cost of lying: Evidence from Norway and the United States (with Hallgeir Sjåstad, Mathias Ekström and Kjetil Bjortvatn).

Promises are ubiquitous in social life. In their simplest form, they are expressed in plain words, a handshake or a simple nod. Promises can be thought of as a commitment device, ensuring that future behaviour is in line with one's long-term interest, despite the presence of short-term temptations to deviate. Honesty is a central moral value necessary for sustained cooperation and hence a trait that likely is in the long-term interest of the individual, but constantly threatened by temptations faced in the day-to-day life. In this chapter, we answer if making a promise effectively promotes honesty by increasing the moral cost of lying.

We conducted three incentivized experiments with 7,205 participants in Norway and the US. In our three experiments, based on the "mind-game" paradigm (Jiang, 2013; Kajackaite & Gneezy, 2017), participants had monetary incentives to be dishonest about the result of a dice roll without any risk of being caught. Based on the Cognitive Dissonance Theory (Aronson, 1992; Festinger, 1957), we designed three interventions to study the effect of promises on the intrinsic costs for lying and contrasted them to a baseline condition. In one, participants made a simple non-binding, non-verifiable promise of honesty. In the second, participants received a simple trust message -"We trust you". In the third, we combine both the promise and the trust message.

Our main result supports the hypothesis of a "promise effect" in Norway and the United States. We observe an average treatment effect of 7 percentage points as compared to the control condition in both countries. This effect corresponds to a reduction in dishonesty of around 30%. This "promise effect" is consistent across various individual-level control variables. We observe no additional treatment effect from adding a trust message to the promise intervention or by itself. Finally, our study also provides suggestive evidence that the effect of promises seems to require an active choice, as promises made by default were not as effective in reducing dishonesty.

This study is the first to use the "mind-game" paradigm to examine the effect of promises on honesty. In addition, we use a large sample of participants to increase statistical power and allow tests in sub-groups of the population. Our results contribute to the literature by providing robust evidence of the effectiveness of promises in promoting honesty and exploring the potential mechanism behind them.

Chapter 2: Dishonesty at the doctor's office: what influences patients' disclosure of accurate information (with Andrés Castañeda, Alfonso Gulias and Adrián Soto-Mota).

People value presenting a positive image of themselves in front of others. However, protecting their image sometimes leads them to lie. For example, there is abundant evidence that people lie to their doctors about behaviours that cause shame, like drinking, smoking, or having bad eating habits (Levy et al., 2018; 2019). These lies are a cause of concern among clinicians since they make it difficult to provide good diagnosis and worsens patients' health. In the second chapter, we investigate whether altering the framing of doctors' questions can improve the truthfulness of patient reports about alcohol and vegetable consumption. In addition, we examine the role of gender matching between patients and doctors in misreporting behaviour.

We conducted a field experiment in six COVID-19 triage clinics in Mexico City with 1,164 patients. In our study, doctors experimentally varied the framing of questions about alcohol and vegetable consumption. Additionally, in three of those clinics, we took advantage of the random assignment of some patients to female or male doctors to explore the effect of gender alignment on patients' reports. We compared the effect of four-question formulations designed to motivate truth-telling: (i) reminding patients about the importance of the information; (ii) giving patients a physical format; (iii) asking questions in a non-judgmental way; (iv) giving patients enough time to remember and answer.

Our findings indicate that the framing of the questions – at least the four variants we tested and that doctors use in regular medical practice – do not affect patients' responses. Thus, the question framing does not help promote patients' honesty in contexts similar to our experiment. However, though in a smaller sample, we do observe some evidence that male patients report healthier habits when interviewed by females compared to male doctors, which suggests that men are more likely to provide dishonest reports to doctors of the opposite sex.

This article contributes to economic theory and health policy. First, we show that image concerns may lead to dishonest behaviours even when it is costly to the liar, like during clinical consultations. Second, our study is the first to use a field experiment to test the effectiveness of simple, commonly-used, alternatives to framing sensitive questions in clinical interactions. Finally, we show that the simple strategies that doctors use may not be as effective as they expect in ensuring quality patient reports. However, our results highlight the importance of further investigating the effect of gender matching on the doctor-patient relationship and health outcomes.

Chapter 3: Causing harm with others: An experiment on diffusion of responsibility and social norms (with Adrian Vargas-López).

Individuals want to avoid feeling guilty when they do something that causes harm. One strategy people often take is blaming a group instead of accepting their culpability (Bartling & Fischbacher, 2012; McGloin & Piquero, 2009). For example, members of a shooting squad may perceive their actions as morally less problematic if the squad is big than if they must shoot alone, even when the result for the victim is the same. In this chapter, we answer two research questions. First, holding all else constant, are individuals in larger groups of potential accomplices more willing to collaborate in generating harm? Second, do perceived social norms affect the willingness to favour a group decision that negatively affect others?

We conducted an online experiment in which participants were pivotal in a group decision on whether to eliminate a charity donation. Voting to eliminate the donation guaranteed them a monetary incentive, but it hurt the charity. We compared groups with varying numbers of potential culprit accomplices and different levels of perceived social norms, defined as the proportion of other participants observed doing wrong. In our experiment, all participants knew they were pivotal in the group's choice, and we informed them that their behaviour would not affect or be known by others in their group. Thus, we isolate the effects of group size and social norms from other mechanisms.

Results in our experiment indicate that, consistent with the idea of "diffusion of responsibility", a larger proportion of participants accepted eliminating a charity donation for private benefit when in groups with more potential accomplices. Moreover, we find that perceived social norms significantly impact individual moral decision-making. Participants were more likely to eliminate the charity donation when they knew that a larger proportion of individuals in their group favoured that option.

Our experiment contributes to the literature explaining the differences between individual and group moral behaviour. In addition, we propose a simple theoretical framework that clarifies the interaction between mechanisms such as aversion to pivotality, replacement logic and diffusion of responsibility. Finally, our results can inform the design of interventions to promote ethical behaviours in situations where costly individual actions are necessary for obtaining a common good.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4), 1115–1153. https://doi.org/10.3982/ecta14673

Adams, F. D., & Lawrence, G. J. (2011). Bullying victims: The effects last into college. *American Secondary Education*, 4-13. Retrieved from: https://www.jstor.org/stable/23100410

Aronson, E. (1992). The Return of the Repressed: Dissonance Theory Makes a Comeback. *Psychological Inquiry*, 3(4), 303–311. https://doi.org/10.1207/s15327965pli0304_1

Bartling, B., & Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, 79(1), 67-87. https://doi.org/10.1093/restud/rdr023

Bjørnskov, C. (2021). Civic honesty and cultures of trust. *Journal of Behavioral and Experimental Economics*, 92, 101693. https://doi.org/10.1016/j.socec.2021.101693

Brown, S., & Taylor, K. (2008). Bullying, education and earnings: evidence from the National Child Development Study. *Economics of Education Review*, 27(4), 387-401. https://doi.org/10.1016/j.econedurev.2007.03.003

Brown, R., & Velásquez, A. (2017). The effect of violent crime on the human capital accumulation of young adults. *Journal of development economics*, 127, 1-12. https://doi.org/10.1016/j.jdeveco.2017.02.004

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3), 818-827. DOI: 10.1257/aer.97.3.818

Creighton, S., Johnson, G., Robertson, P., Law, J., & Szymkowiak, A. (2015). Dishonest behavior at self-service checkouts. In *HCI in Business: Second International Conference*, HCIB 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings 2 (pp. 267-278). Springer.

Cialdini, R. B., Petrova, P. K., & Goldstein, N. J. (2004). The Hidden Cost of Organizational Dishonesty. *MIT Sloan Management Review*, 45(3), 67–73. Retrieved from https://www.researchgate.net/publication/241364589

Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, 365(6448), 70-73. https://doi.org/10.1126/science.aau8712

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the economics of giving, altruism and reciprocity*, 1, 615-691. https://doi.org/10.1016/S1574-0714(06)01008-6

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford University Press.

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American economic review*, 100(1), 541-556. DOI: 10.1257/aer.100.1.541

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595), 496–499. https://doi.org/10.1038/nature17160

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological bulletin*, 145(1), 1. https://doi.org/10.1037/bul0000174

Jacobsen, C., Fosgaard, T. R., & Pascual-Ezama, D. (2018). Why Do We Lie? a Practical Guide To the Dishonesty Literature. *Journal of Economic Surveys*, 32(2), 357–387. https://doi.org/10.1111/joes.12204

Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior and Organization*, 93, 328–336. https://doi.org/10.1016/j.jebo.2013.04.003

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, 728-741. DOI: 10.1017/cbo9780511803475.019

Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, 102, 433–444. https://doi.org/10.1016/j.geb.2017.01.015

Levy, A. G., Scherer, A. M., Zikmund-Fisher, B. J., Larkin, K., Barnes, G., & Fagerlin, A. (2018). Prevalence of and Factors Associated With Patient Nondisclosure of Medically Relevant Information to Clinicians. *JAMA Network Open*, 1(8). https://doi.org/ilCilli?w252.1WcB06BX1.CmlisCnmok

Levy, A. G., Scherer, A. M., Zikmund-Fisher, B. J., Larkin, K., Barnes, G., & Fagerlin, A. (2019). Assessment of Patient Nondisclosures to Clinicians of Experiencing Imminent Threats. *JAMA Network Open*, 2(8), 1–6. doi:10.1001/jamanetworkopen.2019.9277

McGloin, J. M., & Piquero, A. R. (2009). 'I Wasn't Alone': Collective behaviour and violent delinquency. *Australian & New Zealand Journal of Criminology*, 42(3), 336-353. https://doi.org/10.1375/acri.42.3.33

Sulitzeanu-Kenan, R., Tepe, M., & Yair, O. (2022). Public-sector honesty and corruption: field evidence from 40 countries. *Journal of Public Administration Research and Theory*, 32(2), 310-325. https://doi.org/10.1093/jopart/muab033

Chapter 1: Making a promise increases the moral cost of lying. Evidence from Norway and the United States

Hallgeir Sjåstad
Norwegian School of Economics-FAIR, Norway

Mathias Ekström
Norwegian School of Economics-FAIR, Norway

Pablo Soto-Mota
Norwegian School of Economics-FAIR, Norway

Kjetil Bjorvatn
Norwegian School of Economics-FAIR, Norway

Abstract

Honesty is a central moral value constantly threatened by temptations faced in the day-to-day life. Promises, on the other hand, are a common practice to commit future behaviour to one's long-term goals. But do promises work? In particular, can making a promise effectively promote honesty by increasing the moral cost of lying? We explored this research question in three high-powered experiments using the Mind Game paradigm, in which dishonesty is economically incentivized but impossible to detect at the individual level, recruiting representative samples from Norway and the US (N=7,205). On average, the results show that about 1/4 of participants are willing to lie to win a bonus reward ($100). Crucially, dishonesty was significantly reduced, on average by 30 percent, when participants were confronted with a promise treatment asking them to make an informal and inconsequential promise to report accurate information. We further show that the promise effect is generalizable across two different countries and a long list of individual traits, and independent of experimentally induced feelings of trust. However, promises reduced dishonesty only when they required active choice and not when implemented via passive default. Given the recent debate about the replicability of dishonesty research and the effectiveness of "nudging" interventions, our results provide high-quality evidence of the broad relevance of promises, documenting a moderate but robust effect on subsequent dishonesty in two different countries.

Keywords: Dishonesty, Moral costs, Mind game, Oaths

JEL: C91, D01, D91

1.1 Introduction

Promises play an important role in human communication and social life. The practice of exchanging promises about future behaviour has been observed across different cultures for millennia. In contrast to contracts, which restrict participants to specific actions, a promise is a non-binding communication device to signal the commitment to act according to a standard of conduct (Jacquemet et al., 2019; Koessler, 2022). Another characteristic of promises is that, though they can be expressed as oaths using formal language, a promise can be confirmed by simple words, a handshake, or even a nod. From the business world to courtrooms, promises are frequently encouraged and sometimes required. In fact, after the 2008 crisis, there has been an increasing interest in promises and oath-taking by managers, business schools, and financial institutions (Rabesandratana, 2022; de Bruin, 2016).

However, do promises work? In other words, is there any causal effect of making a promise about honesty on subsequent moral behaviour? One pessimistic view is that promises could be understood as a form of superstition or as cheap talk: costless messages that may change the listener's beliefs without affecting the sender's actions (Farrell & Rabin, 1996). Another view emphasize that people are only inclined to follow their promises when they are made public such that reputational concerns can shape their decisions in the present (Sjåstad, 2019; Vonasch & Sjåstad, 2021). In this study, however, we investigate whether it could be that promises also operate at a deeper psychological level, amplifying the *internal* moral cost of lying and thereby reduce dishonesty?

To answer this question, we conducted three incentivized experiments with more than 7,000 participants from Norway and the US. Our study is the first to use the "mind-game" paradigm (Jiang, 2013; Kajackaite & Gneezy, 2017) to isolate the potential effect of promises on the internal cost of lying. In our experiments, participants are put in a situation with financial incentives to be dishonest but no risk of being caught. We contrast a baseline condition to one in which participants make a simple non-binding and non-verifiable promise of honesty. Moreover, since trust has been a key predictor of honesty worldwide (Bjørnskov, 2021; Tannenbaum et al., 2022), we test whether increased perceived trust, experimentally manipulated through a simple message, moderates the effect of promises on lying.

As our primary result, we find robust evidence that making a promise reduces dishonesty, by increasing the internal moral cost of lying. We find consistent support for this promise effect in two different societies, Norway and the United States. We observe an average treatment effect of 7 percentage points as compared to the control condition in both countries. This effect corresponds to a reduction in dishonesty of around 30%. We further show that this "promise effect" generalize across a broad range of individual traits, such as gender, age, income, and political views. Moreover, we find no additional treatment effect from adding a trust message to the promise intervention. Finally, and as a direction for future research, we provide suggestive evidence that the effect of promises seems to require an active choice, as promises made via presumed consent did not significantly reduce dishonesty.

The rest of the paper is organised as follows. Section 2 reviews the literature on (dis)honest behaviour and promises. Section 3 presents a simple theoretical framework to fix ideas about our hypotheses. Section 4 includes a detailed description of the main experimental design. In Section 5, we present the results in the three experiments. Finally, in Section 6, we discuss the implications of our observations.

1.2 Related literature

Honesty, a fundamental moral value, is positively correlated with trust, cooperation, and economic development (Gächter & Schulz, 2016). Despite its significance, dishonesty pervades social relationships, organizations, politics, and the marketplace, resulting in detrimental consequences (Cohn et al., 2019). Dishonest behaviour imposes substantial costs on firms and countries due to consumer dishonesty (Creighton et al., 2015), employee theft, and managerial dishonesty (Cialdini et al., 2004). Additionally, public sector dishonesty correlates with corruption, negatively affecting economic performance and social development (Sulitzeanu-Kenan et al., 2022). Thus, understanding the drivers of dishonest behaviour and potential remedies is important to reduce its harmful effects on society.

The rational-actor model in traditional economics predicts that individuals will lie if the expected material benefits exceed the expected cost (Becker, 1968). In contrast, a comprehensive meta-analysis of incentivized honesty experiments in behavioural economics, psychology, and sociology concluded that people lie "surprisingly little" (Abeler et al., 2019). This result is consistent with the modern social science perspective that people care about both their economic self-interest and their moral ideals and usually try to balance the two when making decisions (Awad et al., 2018; Cappelen et al., 2007; Fehr & Gächter, 2000; Gneezy et al., 2018; Kahneman et al., 1986; Sjåstad, 2019).

Previous research has shown that commitments, for instance in the form of simple promises, can be useful tools for promoting desirable actions, such as increasing public goods contributions (Koessler, 2022), sustaining cooperation (Ederer & Schneider, 2022), and fostering charitable giving (Meyer & Tripodi, 2021). Nevertheless, the literature on promises' effectiveness in promoting honesty yields mixed results. While some have found that oaths, and other formal commitments, reduce deception (Beck, 2021; Peer & Feldman, 2021), others have found them to be effective only in particular contexts and for more heavy-handed interventions (Cagala et al., 2019; Jacquemet et al., 2020). Related to promises, we should also mention the contested "signing first" effect (Shu et al., 2012). The most well-known finding from that early work was that signing a veracity statement at the top of an insurance form could reduce the amount of dishonest overclaiming further down. However, this effect failed to replicate in a large-scale replication attempt (Kristal et al., 2020) and was later retracted due to data irregularities in the original study.

Potential explanations for these inconsistencies include small study samples, publication bias, and lack of international comparisons (Gerlach et al., 2019; Koning et al., 2020). In addition, the various experimental paradigms used in the dishonesty literature capture non-comparable elements in lying behaviour and do not always isolate external motives effectively, most notably the detection risk (Gerlach & Teodorescu, 2022). Importantly, if there is a non-zero chance of detection, the request of a promise may not only affect the internal moral cost of lying – it may also affect the person’s beliefs when it comes to the probability of getting caught, and thereby confounding the direct effect of the promise.

The failed replication and mixed results suggest that promise-like interventions may be ineffective in addressing dishonesty or necessitate more precise methodologies for hypothesis testing. Our experiments address this debate by investigating explicit promises' potential impact on dishonesty using large and diverse participant samples, high statistical power, pre-registration, and open data. To our knowledge, this paper is also the first to examine the causal effect of promises in a mind-game paradigm where detection risk is zero, effectively isolating potential confounders and enabling us to identify the effect on the internal moral cost of lying. Finally, we explore whether variations in perceived trust influence the efficacy of promises – using both experimental manipulation as well as comparisons within and between two countries (Norway and the USA) that differ fundamentally in terms of general trust.

1.3 Theoretical framework

The meta-analysis by Abeler et al. (2019) identified two primary motives for truth-telling: External motives, such as social image and reputational concerns, and Internal motives, such as self-image and the purely moral cost of lying. The current research focuses on the second category. Our study explores whether promises can increase the internal moral cost of lying, even when there is an economic incentive to lie, no external reward for telling the truth, nor external punishment for acting dishonestly.

The theoretical basis for a hypothesis about a "promise effect" on the moral cost of lying draws from cognitive dissonance theory (Aronson, 1992; Festinger, 1957). This theory states that people are motivated to maintain a positive and consistent self-image. As a result, inconsistencies between self-standards (e.g. being a competent or moral person) and behaviour create an uncomfortable "cognitive dissonance" that people seek to avoid. According to this perspective, people with a self-standard of honesty will experience cognitive dissonance from lying, so they will tend to avoid it.

It is well-established that people have different moral standards concerning what they consider the right thing to do (Awad et al., 2018; Van Doesum et al., 2021). In addition, the attention to one's moral standards can also vary across time and place. We hypothesize that increasing the salience or drawing attention to the moral standard by something like a promise could increase the standards' weight in the decision process. For example, we expect that it is more dissonance-provoking to violate the standard by lying if the person has just promised to tell the truth, ultimately leading to a reduction in dishonesty. Therefore, we do not claim that a simple promise or other light-touch behavioural interventions are enough to change people's moral views and most profound convictions. Instead, we assume that promises increase the attention given to an existing moral standard of honesty.

In this experiment, we also investigate whether variations in perceived interpersonal trust change the effect of simple promises. Intuitively, it is more psychologically painful (causes more cognitive dissonance) to lie to someone who trusts us than to someone who does not. The reason for this may be reciprocity, understood in Psychological Game Theory as an aversion to feeling guilt for "letting others down" (Charness & Dufwenberg, 2010). That is, to fail to comply with what we believe that others expect from us, given that they did something for us first (Dufwenberg & Kirchsteiger, 2004). Other investigations have shown that guilt aversion is a possible explanation for honesty in laboratory experiments (Battigalli et al., 2013). Therefore, receiving a signal of trust might increase both the willingness to be honest in general, and the willingness to follow a promise of honesty in particular.

To fix these ideas in a simple model, consider the following individual utility function:

$$U = a - pq - b(c)m$$

In this equation, the first term a is the monetary gain of dishonesty. The second term is the expected monetary cost of dishonesty, with p being the probability of being caught and q the fine if caught. Finally, the third term is the intrinsic cost of dishonesty, where $b(c)$ is the weight or attention placed on the moral cost of lying $m.$

We interpret the moral cost of lying, m, as how far from a moral ideal of honesty the individual perceives a lie to be. In the conceptual framework explained before it can be understood as how large the cognitive dissonance is produced by being dishonest in a given situation. Moreover, we assume that the weight, or attention $b(c)$, we give to this moral cost changes in different contexts, c. So, for instance, this weight may be lower in a situation where cheating is perceived as widespread or general trust is low.

Our primary hypothesis is that promising to be honest reduces lying behaviour. In our experiment, participants are in a situation where the promise is inconsequential and the risk of detection of lying is zero, and thus, the decision whether to lie depends only on comparing the monetary gain with the intrinsic moral cost. In other words, we test if making a promise increases the moral cost of lying. Following the framework explained before, we propose that a soft intervention as the promise does not change the moral standard, m, but increases attention to that standard, $b(c)$. Furthermore, we hypothesize that a trust message reduces lies by changing the context, which potentially also interacts with the effect of the promise.

1.4 Experiment design

We implemented our three experiments with online samples in Norway and the US. Online experiments are a useful tool to study dishonest behaviour. Evidence shows that individuals behave similarly in online experiments and in laboratory environments (Prissé & Jorrat, 2022). In addition, dishonesty levels measured in online experiments correlate with moral behaviour in other experiments (Schild et al., 2021) and norm violations in daily life (Dai et al., 2018).

Our experiment is based on the mind-game paradigm (Jiang, 2013; Kajackaite & Gneezy, 2017). After completing a consent form, participants read the instructions for a Guessing game. Participants win the game if they guess the result of a six-sided dice roll. Before the dice roll, participants are asked to think of the number they believe will show up. We do not ask for their prediction, so the guess is always private. Then participants then observe the result of the dice roll and report the number they saw and if they won. We remind them that if they thought of the correct number they enter a raffle for a 100 USD bonus in addition to their participation fee. Since the participant's guesses are completely private, they could misreport having won without being detected. After the experiment, participants answered a socioeconomic questionnaire. We present detailed instructions for the experiment in Appendix 4. The IRB at the Norwegian School of Economics approved the study with the application number NHH-IRB 31/21. We pre-registered our hypothesis in the platform AsPredicted (Nums. 97210 and 97208).

We compare the baseline situation to some experimental variations to measure the effect of promises on honesty levels. Our three experiments differ in the country and the conditions included. The conditions were:

1. *Control condition*: participants roll the dice without reading any additional message.
2. *Trust & Promise*: participants read a message that said "We trust you" and then answered a question about whether they could confirm that they would report honest information in the study (Yes or No).
3. *Promise*: participants answered a question about whether they could confirm that they would report honest information in the study (Yes or No).
4. *Trust*: participants read a message that said "We trust you".

As in similar dice-rolling designs that study dishonesty, our identification strategy exploits that even when it is impossible to observe lies at the individual level, it is possible to infer dishonesty at the group level (Fischbacher & Föllmi-Heusi, 2013). No matter the participant's prediction, the probability of winning the Guessing game is $\frac{1}{6}$. So, if there are only honest participants, the expected proportion of winners would approximate $\frac{1}{6}$. Consequently, differences in group-level proportions of reported winners indicate varying honesty levels. It should also be mentioned that two out of 2,879 individuals answered No when asked to make a promise of providing honest information. We include these individuals in all analysis, and doing so has no consequences for the results.

In addition to the traditional identification strategy, we employ the method proposed by Hugh-Jones (2019) to estimate lying rates in binary dishonesty experiments (see Appendix 2). This Bayesian estimation technique contrasts with the conventional frequentist approach prevalent in the literature. The Bayesian method provides insights into the potential distribution of lying rates, offering increased precision in smaller samples. Though all our experiments are well powered, we use this alternative approach as a robustness test for its usefulness in comparing distributions as opposed to point-estimates. We observe that the results derived from this technique align with those obtained using the traditional approach discussed in this section.

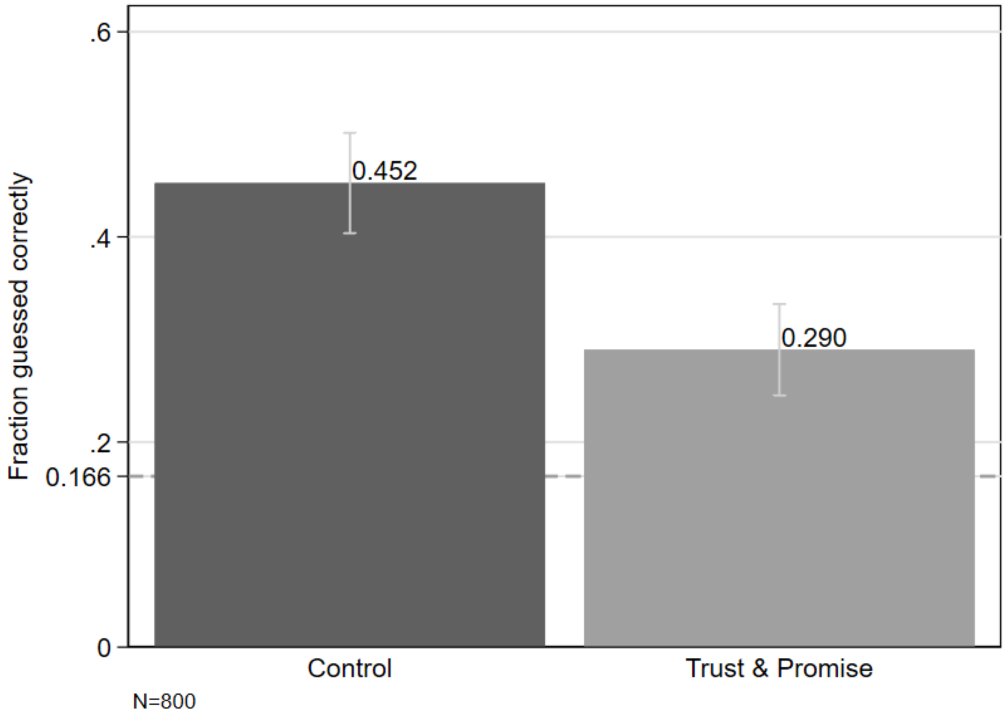
1.5 Results

1.5.1 Experiment 1

In our first study we recruited 800 participants from the general population of Norway via the survey provider Norstat (53.5% female; mean age 47 years (SD=17.95)). Participants were randomly assigned to either the *Control* or the combined *Promise and Trust* condition. Appendix 1 outlines participants' socioeconomic characteristics across the three experiments.

Figure 1 presents the main result. By comparing the fraction of people who declared that they won in the two groups, we observe a large and significant treatment effect of promises on dishonesty. In the *Control* condition, 45% reported guessing the dice roll outcome correctly, exceeding the theoretical 16.66% prediction assuming complete honesty. Conversely, in the *Promise and Trust* treatment group, only 29% stated they won, reflecting a reduction of dishonesty of 16 percentage points ($p<0.001$, two-sided proportion tests unless otherwise specified). This result suggests a 57% decrease in dishonest participants due to the *Promise and Trust* treatment. Appendix 3 confirms the finding using regression analysis, with and without control variables. Thus, our initial study demonstrates that combining a promise with a signal of trust can substantially mitigate dishonesty in the mind game.

Chapter 1. Figure 1. Results of Experiment 1



1.5.2 Experiment 2

The objective of our second study was twofold: first, to directly replicate Experiment 1 using participants from the same population, and second, to determine the driving force behind the effect observed in the combined Promise and Trust treatment. In other words, to disentangle the effects of the act of making a promise, the trust signal, and their combination. So, we recruited 1,600 participants from the general population of Norway through Norstat (50% female; mean age 50.83 years (SD=17.44)). Participants were randomly assigned to one of four groups: *Control*, combined *Promise and Trust*, *Promise* only, or *Trust* only condition.

Considering the effects before adding socioeconomic covariates, 36% of participants in the Control group reported winning. This rate is comparable to the 33% in the Trust group. However, in the Promise only and the combined Promise and Trust groups, 28.5% and 30% said they won, respectively. Thus, these observations suggest that making a promise reduced dishonest behaviour by 40.6%.

Chapter 1. Table 1. Regression analysis of main treatment effects in experiment 2 and 3

	Exp 2: Norway		Exp 3: U.S.		Both US and Norway	
	(1)	(2)	(3)	(4)	(5)	(6)
Promise & Trust	-0.05	-0.06*	-0.05**	-0.06**	-0.05**	-0.06***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Promise	-0.08**	-0.09***	-0.06**	-0.06***	-0.06***	-0.07***
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Trust	-0.03	-0.04	-0.03	-0.03	-0.03	-0.03*
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
U.S.						0.07***
						(0.02)
Constant	0.36***	0.43***	0.49***	0.81***	0.44***	0.65***
	(0.02)	(0.09)	(0.02)	(0.12)	(0.01)	(0.07)
Controls	NO	YES	NO	YES	NO	YES
N	1600	1600	3240	3240	4840	4840

Note: OLS regressions with robust standard errors in parentheses. The outcome variable is a dummy equal to one if the subject reported a correct guess (zero otherwise). Controls: age, gender, sector of employment, education, political stance, believe in God (0-10), believe in free will (0-10). * p<0.05, ** p<0.01, ***<0.0001

Table 1 (columns 1 and 2) confirms this finding using OLS regression analysis. The table presents point estimates of each treatment's effect on the likelihood of reporting winning relative to the Control group. Three observations standout: first, the main result from Study 1 replicates, with the combined Promise and Trust treatment reducing dishonesty (p=.086), although with a smaller effect size (6 p.p. vs 16 p.p.). Second, after splitting up the combined treatment, we found an equally strong (9 p.p.) and statistically significant effect of making a promise (p=.01), while providing a signal of trust has no effect (p=.193). Consequently, the dishonesty reduction in the Promise and Trust treatment is only determined by making a promise.

1.5.3 Experiment 3

The aim of our third study was threefold. First, we conducted a conceptual replication of the previous experiments with participants from a different country with a more diverse population. Second, we increased the sample size to detect even smaller effects, thus determining whether making a promise was indeed the crucial factor. Third, we incorporated a *promise-by-default* treatment to elucidate the underlying mechanism of making a promise.

We recruited 4805 participants from the general population of the United States via the survey platform Prolific (49% female; average age 37.17 years (SD=13.23)). A total of 3240 participants were randomly assigned to either a *Control* group, a combined

Promise and Trust treatment, a *Promise* only treatment, or a *Trust* only treatment. Additionally, 794 participants were randomised to a treatment where the answer "Yes" to the promise question had been pre-selected as the default. Similarly, 771 participants were randomised to a treatment where the pre-selected answer to the promise question was "No".

In this experiment, 48% of participants in the *Control* group reported winning, implying a 38.44% dishonesty rate among those who saw a losing outcome. This rate is similar to the 45.89% found in the *Trust* group, implying a null-effect of the trust message. In contrast, in the *Promise* condition 42.92% participants reported winning, equating to a 31.51% proportion of dishonest reports among those who lost. In the combined *Trust and Promise* group 43.31% claimed they won. As in Experiment 2, these observations suggest that making a promise reduces dishonesty while that the message of trust is ineffective.

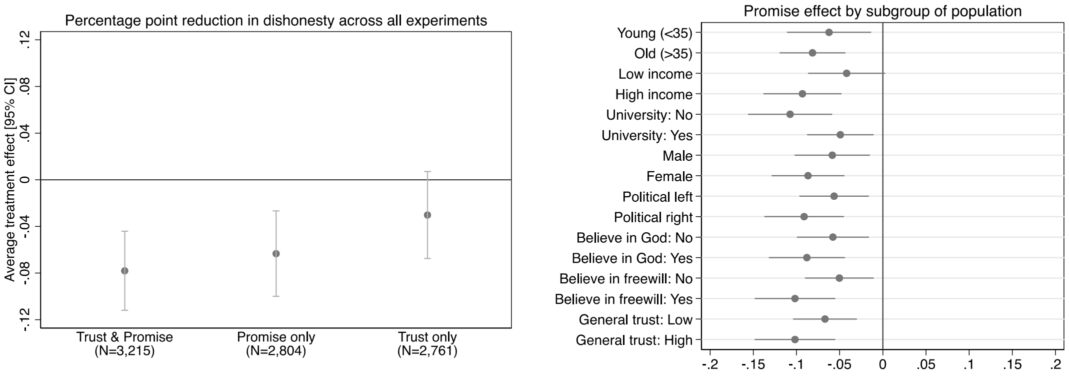
We present these results in Table 1 with and without the consideration of covariates. (column 3 and 4). There is a striking resemblance between the results in the United States and Norway. In study 3, the Promise and Trust treatment reduces dishonesty by 6 percentage points (p=.023). Similarly, when we split up the combined treatment, there is, again, an equally strong (6 p.p.) and significant effect of making a promise (p=.01), but no effect of providing a signal of trust (p=.242). Hence, also in the United States, the reduction in dishonesty caused by the Promise and Trust treatment is exclusively determined by the act of making a promise.

Interestingly, despite the robust and large effect of making an *active* promise on dishonesty, a separate analysis reveals that there is no statistically significant reduction in dishonesty in the group where we pre-selected the answer “Yes” to the promise question (p=.103) nor when the pre-selected answer was “No” (p=0.212). One interpretation of this finding is that decision-makers needs to be sufficiently engaged in the act of making of a promise, for the promise to have any consequences on behavior.

1.5.4 Heterogeneity analysis

As a final piece of evidence, Figure 2 (Panel A) and columns (5) and (6) in Table 1 present the average treatment effects on dishonesty across all our studies. In Panel B of Figure 2, we present the effect of making a promise across different subgroups of the population. As observed in each of the independent studies, the analysis shows that there is a large and significant effect of making a promise, but no effect of signaling trust neither by itself nor combined with the promise. The analysis also reveals that effect of making a promise is observed in all subgroups of the population, suggesting that the underlying mechanism behind promises has broad appeal and is generalizable to many populations.

Chapter 1. Figure 2. Average treatment effects across all studies (Panel A). Average treatment effect of making a promise in subgroups of the population (Panel B)



1.6 Discussion

Exchanging promises are a common practice in social life to signal future moral behaviour. Recently, they have also gained attention as alternative management practices to curb dishonesty in the corporate world (de Bruin, 2016). Unlike increased vigilance or more stringent rules, commitments of honesty are cheap and easy to implement (Hilo-Merkovich et al., 2021). However, does it help to make a promise in reducing dishonesty? We answer: Yes. In three large experiments with representative samples from two countries, we found that promises are effective in reducing dishonesty. In contrast, we observed that a simple message of trust does not promote honesty in itself nor moderate the effect of making a promise. Finally, we found suggestive evidence that promises need to be active and unpersuaded to work.

Our results have implications for theory in moral decision-making. First, previous evidence of the effect of simple promises had been questioned because of small sample sizes and failures in replication (Koning et al., 2020). Our study gives robust evidence in favour of the idea that unbinding promises are useful tools to nudge people towards honesty. Also, the recent literature on the potential use of pledges to reduce dishonesty has focused on those who are morally charged or use ceremonial language (Cagala et al., 2019; Jacquemet et al., 2019). We extend that literature by showing that a simple confirmatory question can be enough to commit people to give honest and correct information. Finally, our results suggest that actively committing to being honest might be more effective than doing that in a persuaded way, which is an observation congruent with Cognitive Dissonance Theory (Aronson, 1992; Festinger, 1957).

Although we have found robust evidence showing that promises enhance the moral cost of lying, we do notice that there is still scarce evidence of similar effects in the field. Indeed, the few published field experiments where researchers implemented similar interventions show null results (Bhanot, 2017; Martuza et al., 2022). One plausible explanation for this inconsistency is that field settings often are associated with a detection risk, which we intentionally and effectively could abstain from. Future research would benefit from more closely studying the interplay between promises and detection risk, in order to identify when and how simple promises can be effective in reducing the harmful effects of lying in real-life scenarios.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4), 1115–1153. https://doi.org/10.3982/ecta14673

Aronson, E. (1992). The Return of the Repressed: Dissonance Theory Makes a Comeback. *Psychological Inquiry*, 3(4), 303–311. https://doi.org/10.1207/s15327965pli0304_1

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Battigalli, P., Charness, G., & Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior and Organization*, 93, 227–232. https://doi.org/10.1016/j.jebo.2013.03.033

Beck, T. (2021). How the honesty oath works: Quick, intuitive truth telling under oath. *Journal of Behavioral and Experimental Economics*, 94. https://doi.org/10.1016/j.socec.2021.101728

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Economic Analysis of the Law: Selected Readings*, 255–265. https://doi.org/10.1002/9780470752135.ch25

Bhanot, S. P. (2017). Cheap promises: Evidence from loan repayment pledges in an online experiment. *Journal of Economic Behavior and Organization*, 140, 246–266. https://doi.org/10.1016/j.jebo.2017.04.007

Bjørnskov, C. (2021). Civic honesty and cultures of trust. *Journal of Behavioral and Experimental Economics*, 92, 101693. https://doi.org/10.1016/j.socec.2021.101693

Cagala, T., Glogowsky, U., & Rincke, J. (2019). *Content Matters: The Effects of Commitment Requests on Truth-Telling*. Retrieved from https://ssrn.com/abstract=3432445

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3), 818-827. DOI: 10.1257/aer.97.3.818

Charness, G., & Dufwenberg, M. (2010). Bare promises: An experiment. *Economics Letters*, 107(2), 281–283. https://doi.org/10.1016/j.econlet.2010.02.009

Cialdini, R. B., Petrova, P. K., & Goldstein, N. J. (2004). The Hidden Cost of Organizational Dishonesty. *MIT Sloan Management Review*, 45(3), 67–73. Retrieved from https://www.researchgate.net/publication/241364589

Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, 365(6448), 70–73. https://doi.org/10.1126/science.aau8712

Creighton, S., Johnson, G., Robertson, P., Law, J., & Szymkowiak, A. (2015). Dishonest behavior at self-service checkouts. In *HCI in Business: Second International Conference*, HCIB 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings 2 (pp. 267-278). Springer.

Dai, Z., Galeotti, F., & Villeval, M. C. (2018). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science*, 64(3), 1081-1100. https://doi.org/10.1287/mnsc.2016.2616

de Bruin, B. (2016). Pledging Integrity: Oaths as Forms of Business Ethics Management. *Journal of Business Ethics*, *136*(1), 23–42.https://doi.org/10.1007/s10551-014-2504-1

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2), 268–298. https://doi.org/10.1016/j.geb.2003.06.003

Ederer, F., & Schneider, F. (2022). Trust and Promises over Time. *American Economic Journal: Microeconomics*, 14(3), 304–320. https://doi.org/10.1257/mic.20200049

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic perspectives*, *10*(3), 103-118. DOI: 10.1257/jep.10.3.103

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford University Press.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise-an experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525–547. https://doi.org/10.1111/jeea.12014

Hilo-Merkovich, R., Peer, E., & Feldman, Y. (2021). *Affidavit Aversion: Public Preferences for Trust-Based Policy Instruments*. Available at SSRN 3940999.

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. Nature, 531(7595), 496–499. https://doi.org/10.1038/nature17160

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The Truth About Lies: A Meta-Analysis on Dishonest Behavior. *Psychological Bulletin*, 145(1), 1–44. https://doi.org/10.1037/bul0000174.supp

Gerlach, P., & Teodorescu, K. (2022). Measuring dishonest behavior: Hidden dimensions that matter. *Current Opinion in Psychology*, 101408. https://doi.org/10.1016/j.copsyc.2022.101408

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, 108(2), 419–453. https://doi.org/10.1257/aer.20161553

Hugh-Jones, D. (2019). True lies: Comment on Garbarino, Slonim and Villeval (2018). *Journal of the Economic Science Association*, 5, 255-268. https://doi.org/10.1007/s40881-019-00069-x

Jacquemet, N., Luchini, S., Rosaz, J., & Shogren, J. (2019). Truth telling under oath. *Management Science*, *65*(1), 426–438. https://doi.org/10.1287/mnsc.2017.2892

Jacquemet, N., Luchini, S., Malézieux, A., & Shogren, J. F. (2020). Who'll stop lying under oath? Empirical evidence from tax evasion games. *European Economic Review*, 124. https://doi.org/10.1016/j.euroecorev.2020.103369

Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior and Organization*, 93, 328–336. https://doi.org/10.1016/j.jebo.2013.04.003

Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, 102, 433–444. https://doi.org/10.1016/j.geb.2017.01.015

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American economic review*, 728-741. DOI: 10.1017/cbo9780511803475.019

Kristal, A. S., Whillans, A. V, Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *PNAS*. https://doi.org/10.1073/pnas.1911695117/-/DCSupplemental

Koessler, A. K. (2022). Pledges and how social influence shapes their effectiveness.

Journal of Behavioral and Experimental Economics, *98*. https://doi.org/10.1016/j.socec.2022.101848

Koning, L., Junger, M., & van Hoof, J. (2020). Digital signatures: a tool to prevent and predict dishonesty? *Mind and Society*, 19(2), 257–285. https://doi.org/10.1007/s11299-020-00237-1

Martuza, J. Bin, Skard, S. R., Løvlie, L., & Thorbjørnsen, H. (2022). Do honesty-nudges really work? A large-scale field experiment in an insurance context. Journal of Consumer Behaviour. https://doi.org/10.1002/cb.2049

Meyer, C. J., & Tripodi, E. (2021). Image concerns in pledges to give blood: Evidence from a field experiment. *Journal of Economic Psychology*, 87. https://doi.org/10.1016/j.joep.2021.102434

Peer, E., & Feldman, Y. (2021). Honesty pledges for the behaviorally-based regulation of dishonesty. *Journal of European Public Policy*, 28(5), 761–781. https://doi.org/10.1080/13501763.2021.1912149

Prissé, B., & Jorrat, D. (2022). Lab vs online experiments: no differences. *Journal of Behavioral and Experimental Economics*, 101910. https://doi.org/10.1016/j.socec.2022.101910

Rabesandratana, T. (2022). France introduces research integrity oath. *Science*, 377(6603), 251–251. https://doi.org/10.1126/science.add9092

Schild, C., Lilleholt, L., & Zettler, I. (2021). Behavior in cheating paradigms is linked to overall approval rates of crowdworkers. *Journal of Behavioral Decision Making*, 34(2), 157–166. https://doi.org/10.1002/bdm.2195

Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *PNAS*, 109(38), 15197-15200. https://doi.org/10.1073/pnas.1209746109

Sjåstad, H. (2019). Short-sighted greed? Focusing on the future promotes reputation-based generosity. *Judgment and Decision making*, 14(2), 199-213. https://doi.org/10.1017/S1930297500003430

Sulitzeanu-Kenan, R., Tepe, M., & Yair, O. (2022). Public-sector honesty and corruption: field evidence from 40 countries. J*ournal of Public Administration Research and Theory*, 32(2), 310-325. https://doi.org/10.1093/jopart/muab033

Tannenbaum, D., Cohn, A., Zünd, C. L., & Maréchal, M. A. (2022). What do cross-country surveys tell us about social capital?. *Review of Economics and Statistics*, 1-30. https://doi.org/10.1162/rest_a_01245

Van Doesum, N. J., Murphy, R. O., Gallucci, M., Aharonov-Majar, E., Athenstaedt, U., Au, W. T., ... & Van Lange, P. A. (2021). Social mindfulness and prosociality vary across the globe. *Proceedings of the National Academy of Sciences*, 118(35), e2023846118. https://doi.org/10.1073/pnas.2023846118

Vonasch, A. J., & Sjåstad, H. (2021). Future-orientation (as trait and state) promotes reputation-protective choice in moral dilemmas. *Social Psychological and Personality Science*, 12(3), 383-391. https://doi.org/10.1177/1948550619899257

Vésteinsdóttir, V., Joinson, A., Reips, U. D., Danielsdottir, H. B., Thorarinsdottir, E. A., & Thorsdottir, F. (2019). Questions on honest responding. *Behavior Research Methods*, 51(2), 811–825. https://doi.org/10.3758/s13428-018-1121-9

Appendix

Appendix 1: participant's characteristics by experiment

Experiment 1

	Condition		
	Control	**Promise & Trust**	**Total**
Frequency	400	400	800
Female	54.75%	52.25%	53.50%
Age	47.56	46.43	47.00
(In years)	(18.53)	(17.35)	(17.95)
Believes in God	3.4725	3.4125	3.4425
(0:Not at all, 10:Totally)	(3.52)	(3.54)	(3.53)
Trust in others	6.1825	6.2225	6.2025
(0:Not at all, 10:Totally)	(1.97)	(1.82)	(1.89)
Political Opinion	5.0725	4.785	4.92875
(0:Left, 10:Right)	(2.40)	(2.35)	(2.38)
High school or less	0.575	0.5825	0.57875
(0: No, 1:Yes)	(0.49)	(0.49)	(0.49)
Works in public sector	24.50%	30.00%	27.25%

Experiment 2

	Condition				
	Control	**Trust**	**Promise**	**Promise & Trust**	**Total**
Frequency	400	400	400	400	1,600
Female	52.00%	46.75%	49.50%	51.75%	50.00%
Age	51.35	50.64	50.53	50.82	50.84
(In years)	(17.01)	(17.61)	(17.80)	(17.38)	(17.44)
Believes in God	3.738	3.745	3.668	3.378	3.632
(0:Not at all, 10:Totally)	(3.57)	(3.61)	(3.69)	(3.50)	(3.59)
Trust in others	6.568	6.468	6.248	6.513	6.449
(0:Not at all, 10:Totally)	(1.78)	(1.80)	(1.94)	(1.78)	(1.83)
Political Opinion	5.123	4.783	5.033	4.650	4.897
(0:Left, 10:Right)	(2.47)	(2.40)	(2.38)	(2.43)	(2.42)
High school or less	0.640	0.563	0.603	0.598	0.601
(0: No, 1:Yes)	(0.48)	(0.50)	(0.49)	(0.49)	(0.49)
Works in public sector	23.00%	26.50%	27.25%	25.25%	25.50%

Experiment 3

	Condition						
	Control	**Trust**	**Promise**	**Promise (Honest)**	**Promise (Dishonest)**	**Promise & Trust**	**Total**
Frequency	770	791	834	794	771	845	4,805
Female	50.91%	49.94%	48.20%	46.73%	48.25%	47.81%	48.62%
Age	37.08	37.45	36.65	37.21	37.25	37.42	37.17
(In years)	(13.53)	(13.37)	(12.69)	(13.25)	(13.23)	(13.38)	(13.24)
Believes in God	4.632	4.432	4.468	4.540	4.414	4.388	4.478
(0:Not at all, 10:Totally)	(4.02)	(4.10)	(4.07)	(4.10)	(4.10)	(4.10)	(4.08)
Trust in others	5.594	5.475	5.372	5.370	5.484	5.505	5.466
(0:Not at all, 10:Totally)	(2.13)	(2.27)	(2.32)	(2.38)	(2.33)	(2.31)	(2.29)
Political Opinion	3.529	3.343	3.487	3.456	3.422	3.609	3.476
(0:Left, 10:Right)	(2.86)	(2.84)	(2.90)	(2.88)	(2.87)	(2.78)	(2.85)
High school or less	0.634	0.603	0.641	0.594	0.603	0.618	0.616
(0: No, 1:Yes)	(0.48)	(0.49)	(0.48)	(0.49)	(0.49)	(0.49)	(0.49)
Works in public sector	24.68%	25.41%	26.86%	26.07%	26.98%	27.34%	26.24%

All experiments

	Condition						
	Control	**Trust**	**Promise**	**Promise (Honest)**	**Promise (Dishonest)**	**Trust & promise**	**Total**
Frequency	1,570	1,191	1,234	794	771	1,645	7,205
Female	52.17%	48.87%	48.62%	46.73%	48.25%	49.85%	49.47%
Age	43.386	41.877	41.147	37.210	37.245	42.869	41.297
(In years)	(17.05)	(16.17)	(15.93)	(13.25)	(13.23)	(16.49)	(15.98)
Believes in God	4.109	4.202	4.208	4.540	4.414	3.905	4.175
(0:Not at all, 10:Totally)	(3.82)	(3.95)	(3.96)	(4.10)	(4.10)	(3.86)	(3.94)
Trust in others	5.992	5.809	5.656	5.370	5.484	5.925	5.766
(0:Not at all, 10:Totally)	(2.05)	(2.17)	(2.24)	(2.38)	(2.33)	(2.12)	(2.20)
Political Opinion	4.328	3.826	3.988	3.456	3.422	4.148	3.953
(0:Left, 10:Right)	(2.77)	(2.78)	(2.84)	(2.88)	(2.87)	(2.65)	(2.80)
High school or less	0.620	0.589	0.629	0.594	0.603	0.604	0.608
(0: No, 1:Yes)	(0.49)	(0.49)	(0.48)	(0.49)	(0.49)	(0.49)	(0.49)
Works in public sector	24.20%	25.78%	26.99%	26.07%	26.98%	27.48%	26.19%

Appendix 2: regression results Experiment 1

Table 1. Regression analysis of main treatment effect in experiment 1

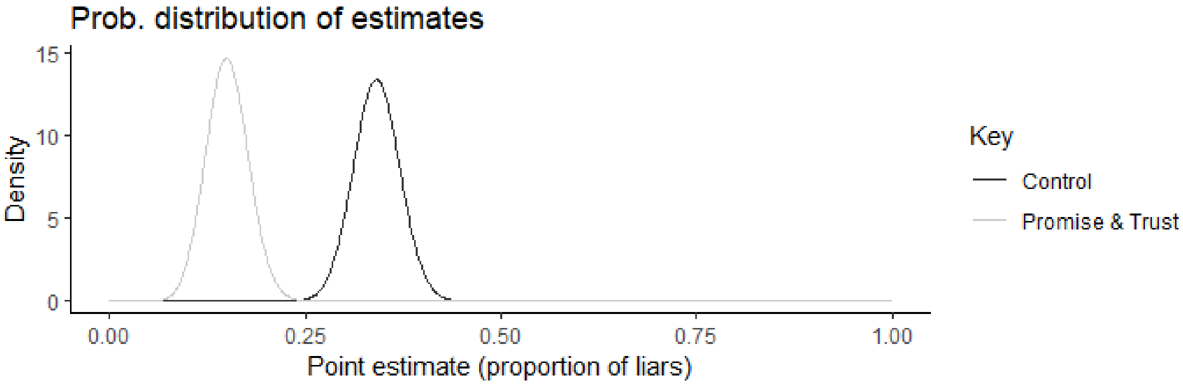
	(1)	(2)
Promise & Trust	-0.162***	-0.157***
	(0.034)	(0.034)
Constant	0.452***	0.599***
	(0.025)	(0.122)
Controls	NO	YES
N	800	800

Note: OLS regressions with robust standard errors in parentheses. The outcome variable is a dummy equal to one if the subject reported a correct guess (zero otherwise). Controls: age, gender, sector of employment, education, political stance, believe in God (0-10), believe in free will (0-10). * p<0.05, ** p<0.01, ***<0.0001

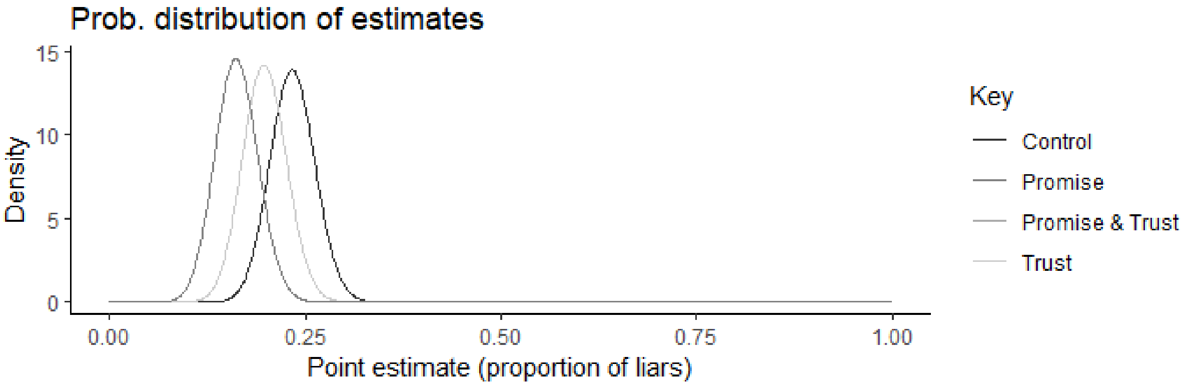
Appendix 3: Robustness using Hugh-Jones (2019) method

We provide estimates of the proportion of dishonest participants in the three experiments using the method suggested by Hugh-Jones (2019). This Bayesian method is based in estimating the distribution of the excess number of winning reports with respect to the expectation under full honesty. The main advantage of this method in contrast to a frequentist approach is that it is possible to use CIs of the distribution of the possible proportion of liars. In all estimations we impose a uniform prior distribution.

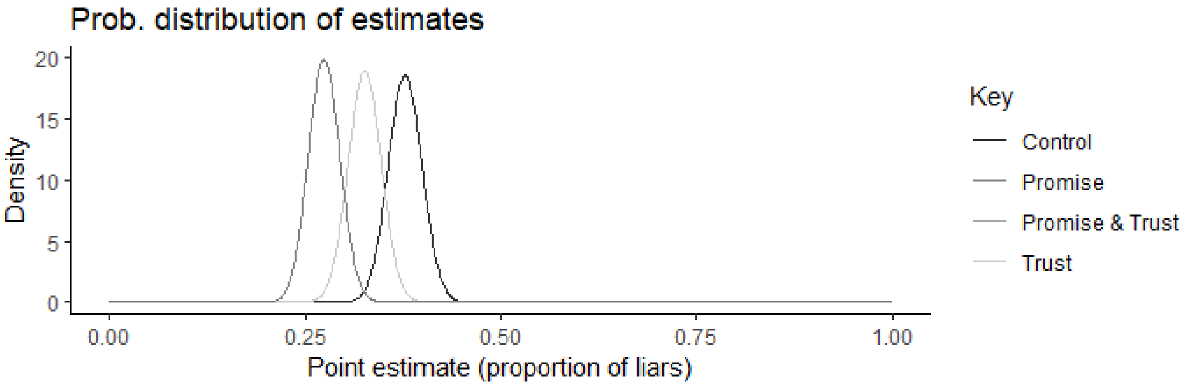
Experiment 1



Experiment 2



Experiment 3

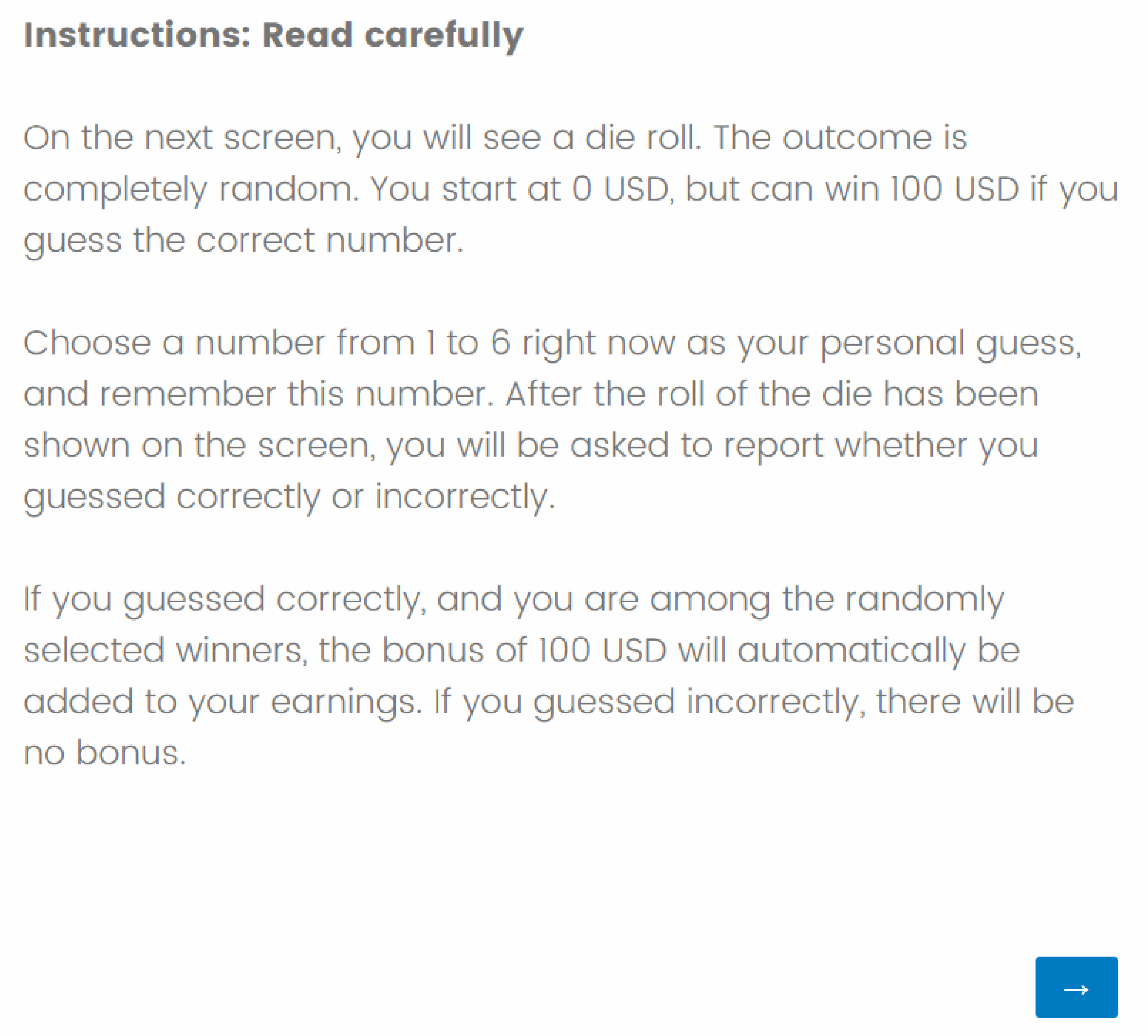


Summary table

				Prop. Liars	95% CI	
		N	**Won**	**(Estimated)**	**Low**	**High**
Experiment 1	**Control**	400	180	0.34	0.281	0.40
	Promise & Trust	400	116	0.149	0.09	0.20
Experiment 2	**Control**	400	144	0.233	0.177	0.29
	Promise & Trust	400	120	0.161	0.11	0.25
	Trust	400	132	0.197	0.145	0.25
	Promise	400	114	0.143	0.08	0.19
Experiment 3	**Control**	770	370	0.377	0.33	0.42
	Promise & Trust	845	333	0.274	0.24	0.31
	Trust	791	346	0.33	0.281	0.37
	Promise	834	330	0.275	0.24	0.31

Appendix4: Materials and instructions (US Experiment)

Control condition



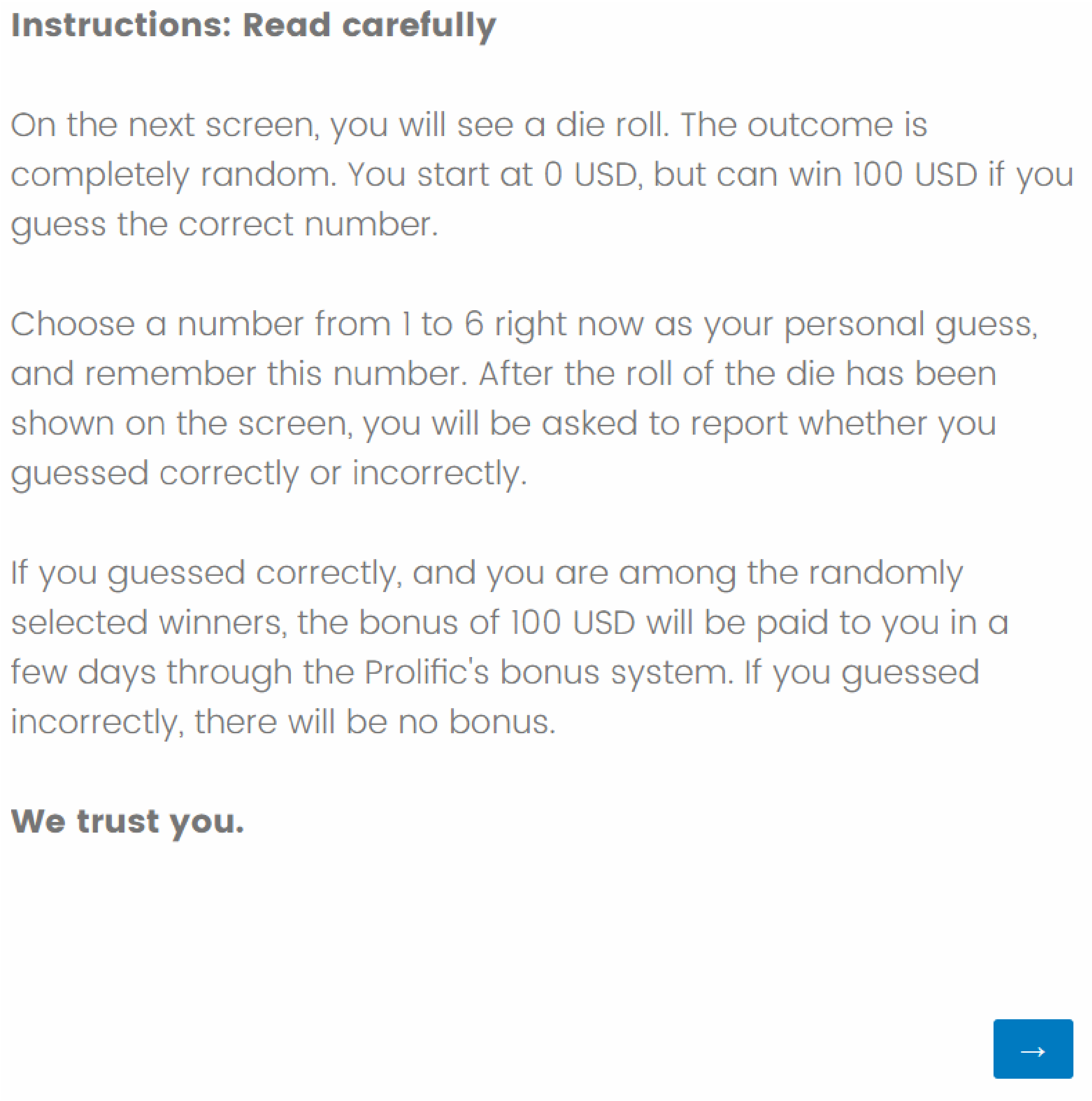
Instructions: Read carefully

On the next screen, you will see a die roll. The outcome is completely random. You start at 0 USD, but can win 100 USD if you guess the correct number.

Choose a number from 1 to 6 right now as your personal guess, and remember this number. After the roll of the die has been shown on the screen, you will be asked to report whether you guessed correctly or incorrectly.

If you guessed correctly, and you are among the randomly selected winners, the bonus of 100 USD will automatically be added to your earnings. If you guessed incorrectly, there will be no bonus.

Trust condition



Instructions: Read carefully

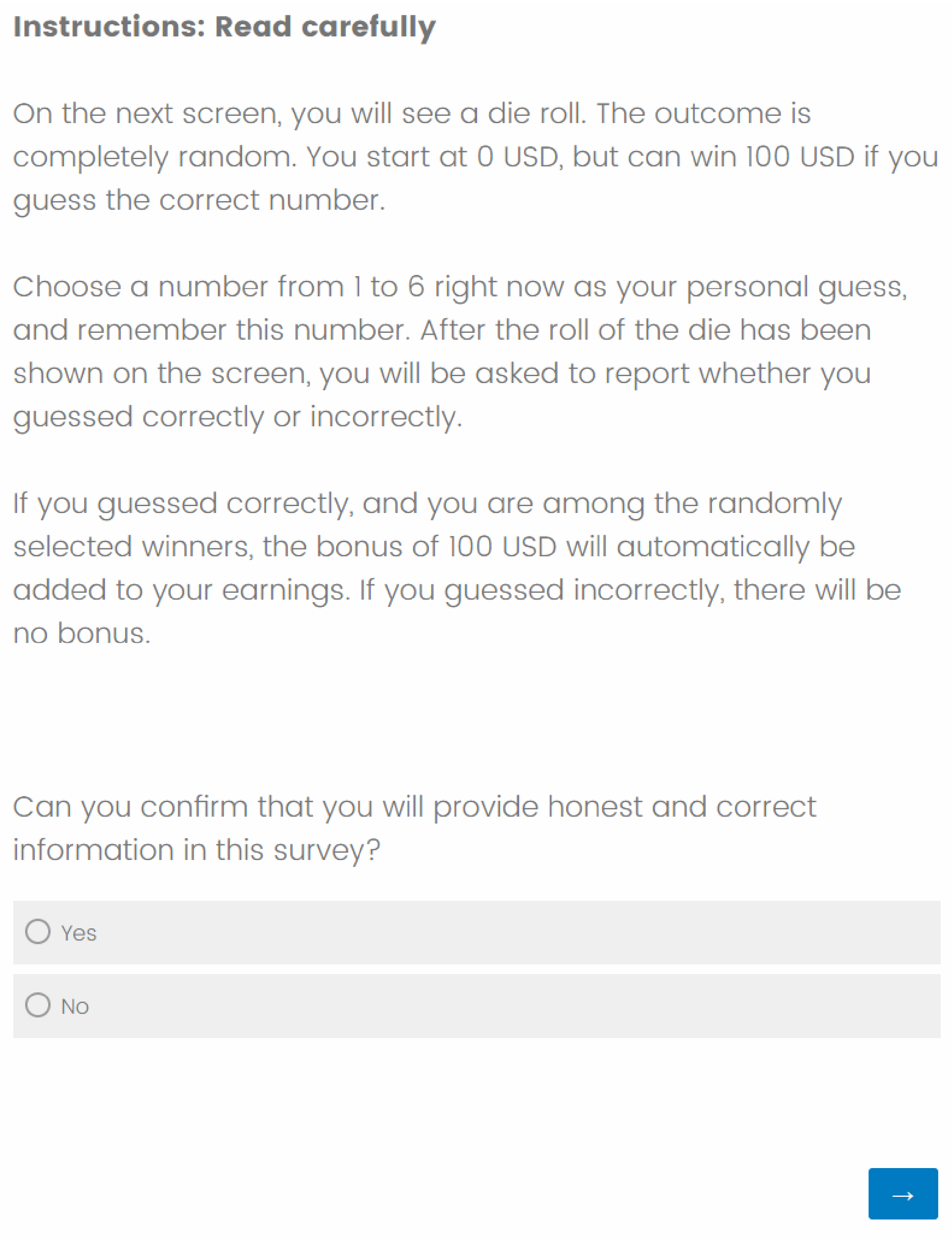
On the next screen, you will see a die roll. The outcome is completely random. You start at 0 USD, but can win 100 USD if you guess the correct number.

Choose a number from 1 to 6 right now as your personal guess, and remember this number. After the roll of the die has been shown on the screen, you will be asked to report whether you guessed correctly or incorrectly.

If you guessed correctly, and you are among the randomly selected winners, the bonus of 100 USD will be paid to you in a few days through the Prolific's bonus system. If you guessed incorrectly, there will be no bonus.

We trust you.

Promise condition



Instructions: Read carefully

On the next screen, you will see a die roll. The outcome is completely random. You start at 0 USD, but can win 100 USD if you guess the correct number.

Choose a number from 1 to 6 right now as your personal guess, and remember this number. After the roll of the die has been shown on the screen, you will be asked to report whether you guessed correctly or incorrectly.

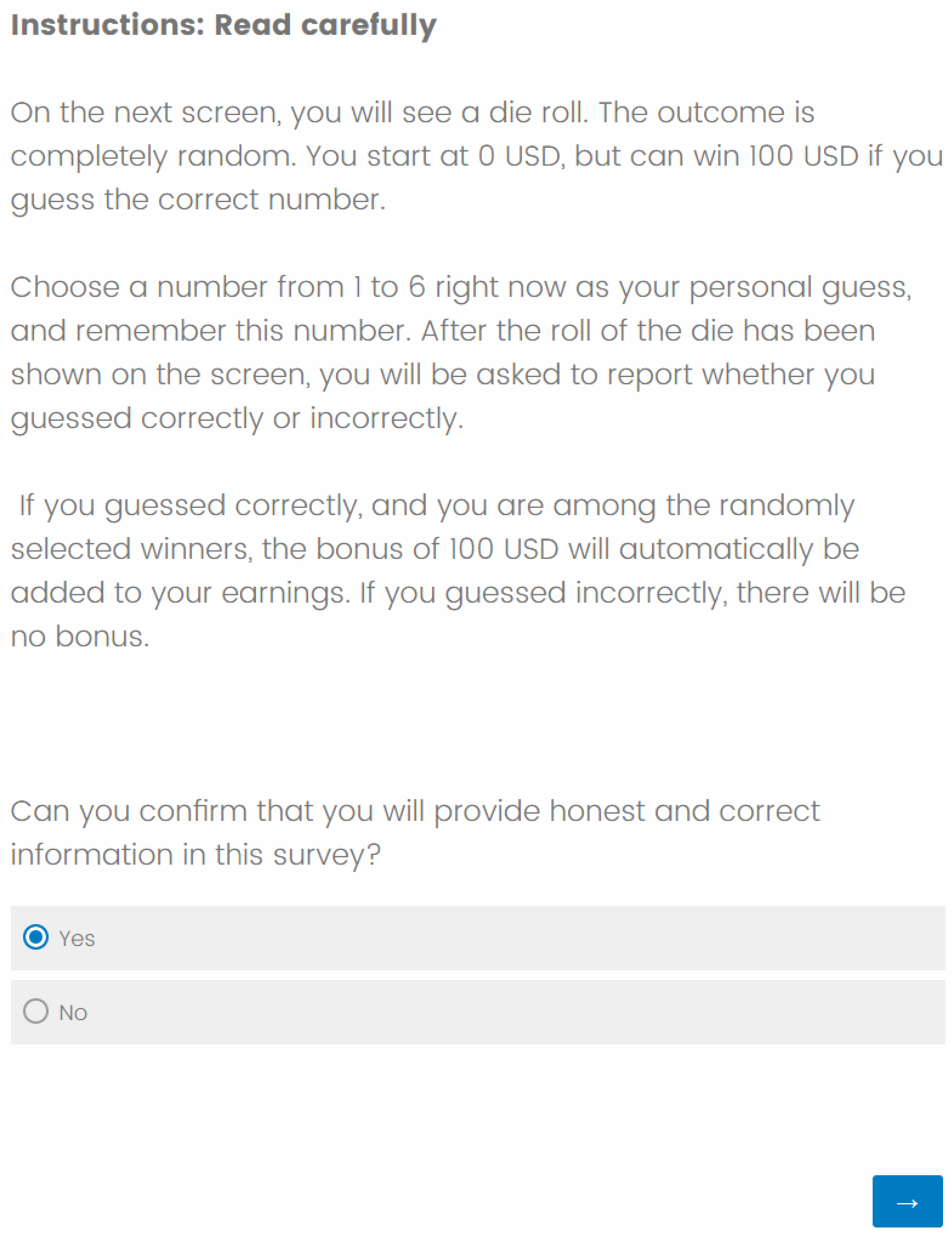
If you guessed correctly, and you are among the randomly selected winners, the bonus of 100 USD will automatically be added to your earnings. If you guessed incorrectly, there will be no bonus.

Can you confirm that you will provide honest and correct information in this survey?

○ Yes

○ No

Promise condition (Honest default)



Instructions: Read carefully

On the next screen, you will see a die roll. The outcome is completely random. You start at 0 USD, but can win 100 USD if you guess the correct number.

Choose a number from 1 to 6 right now as your personal guess, and remember this number. After the roll of the die has been shown on the screen, you will be asked to report whether you guessed correctly or incorrectly.

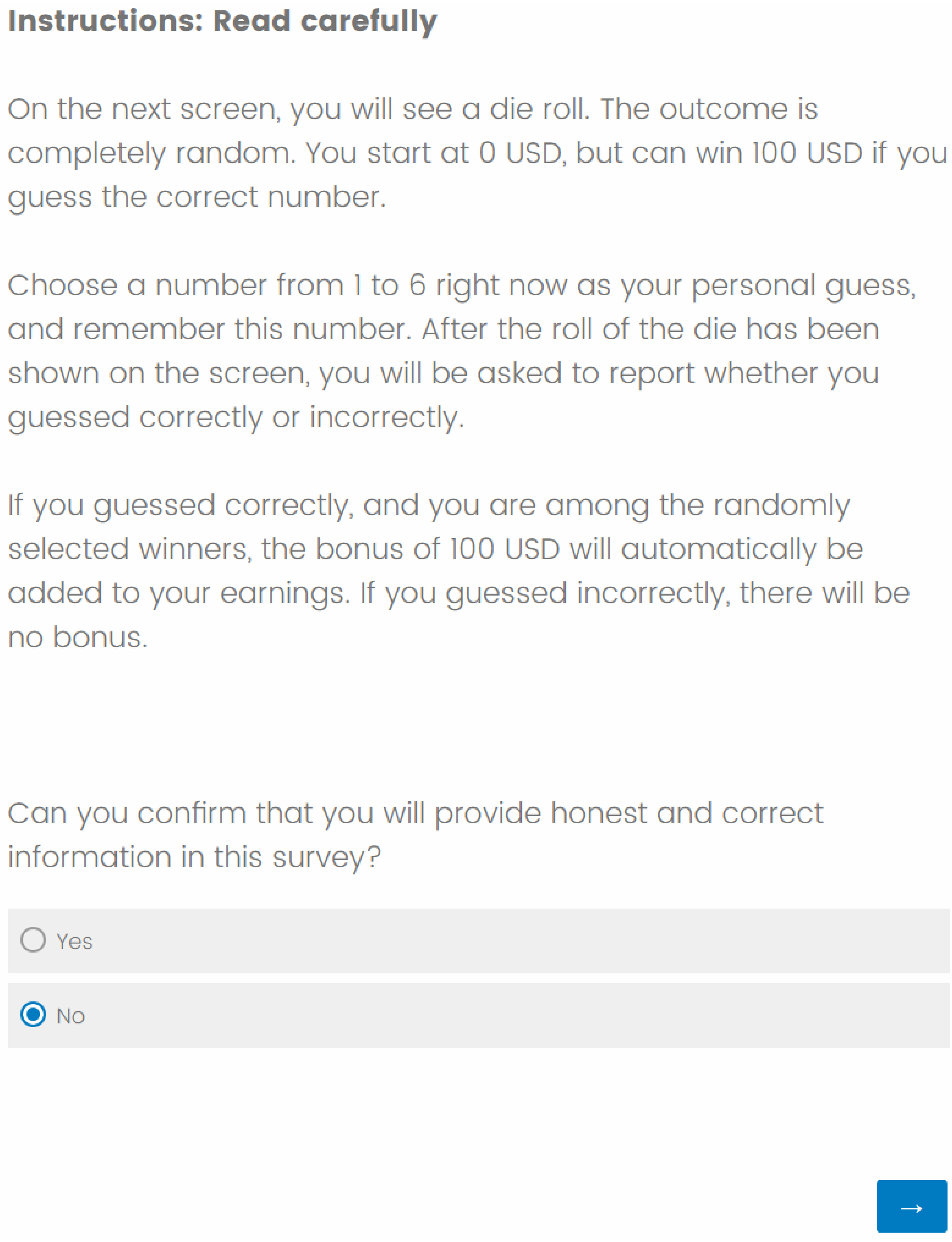
If you guessed correctly, and you are among the randomly selected winners, the bonus of 100 USD will automatically be added to your earnings. If you guessed incorrectly, there will be no bonus.

Can you confirm that you will provide honest and correct information in this survey?

● Yes

○ No

Promise condition (Dishonest default)



Instructions: Read carefully

On the next screen, you will see a die roll. The outcome is completely random. You start at 0 USD, but can win 100 USD if you guess the correct number.

Choose a number from 1 to 6 right now as your personal guess, and remember this number. After the roll of the die has been shown on the screen, you will be asked to report whether you guessed correctly or incorrectly.

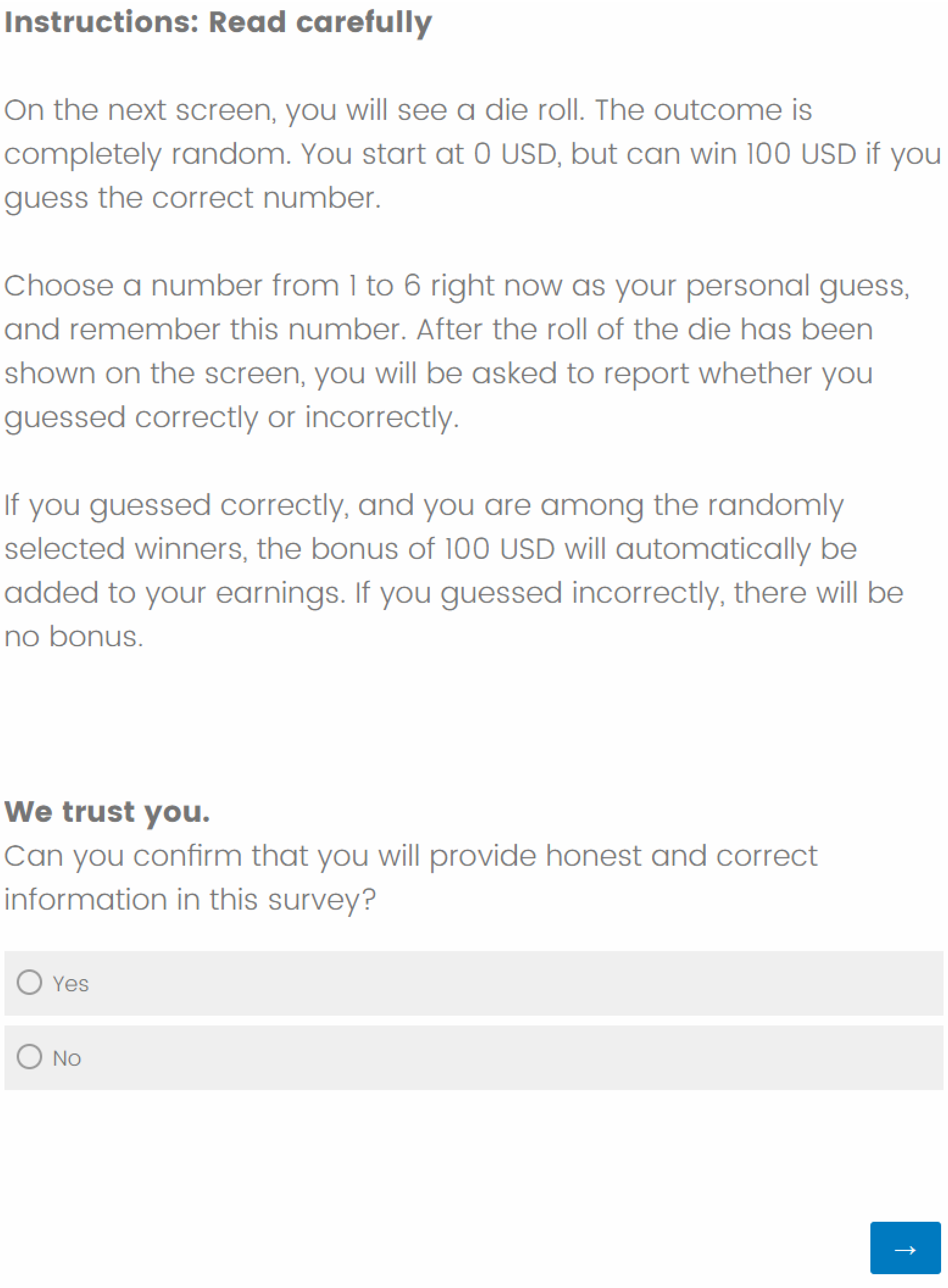
If you guessed correctly, and you are among the randomly selected winners, the bonus of 100 USD will automatically be added to your earnings. If you guessed incorrectly, there will be no bonus.

Can you confirm that you will provide honest and correct information in this survey?

○ Yes

● No

Promise & Trust condition



Instructions: Read carefully

On the next screen, you will see a die roll. The outcome is completely random. You start at 0 USD, but can win 100 USD if you guess the correct number.

Choose a number from 1 to 6 right now as your personal guess, and remember this number. After the roll of the die has been shown on the screen, you will be asked to report whether you guessed correctly or incorrectly.

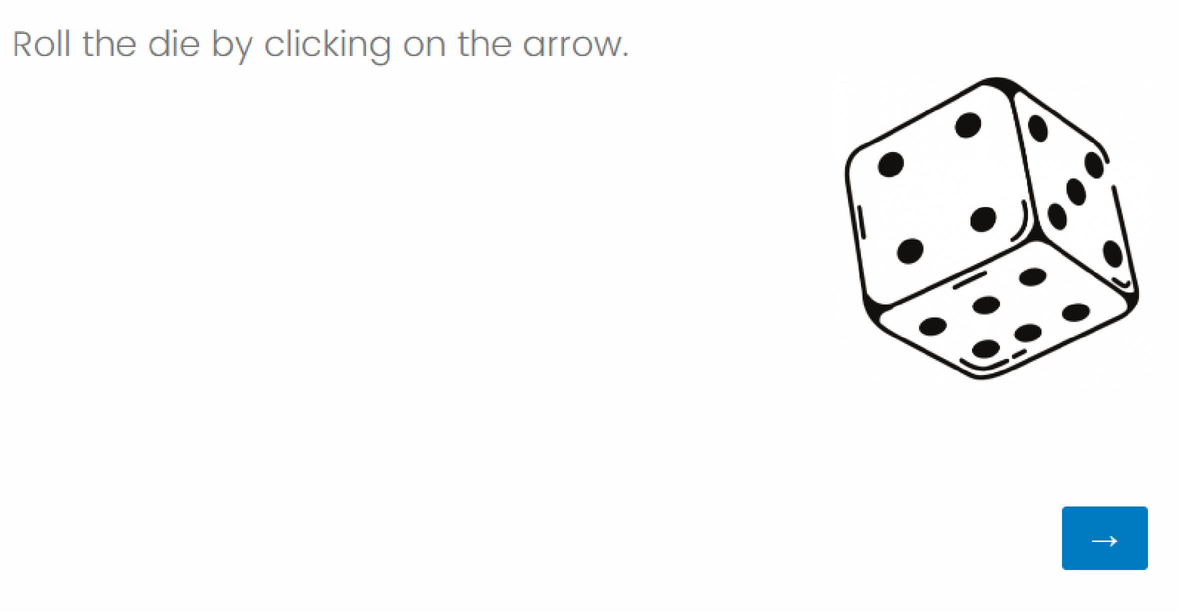
If you guessed correctly, and you are among the randomly selected winners, the bonus of 100 USD will automatically be added to your earnings. If you guessed incorrectly, there will be no bonus.

We trust you.
Can you confirm that you will provide honest and correct information in this survey?

○ Yes

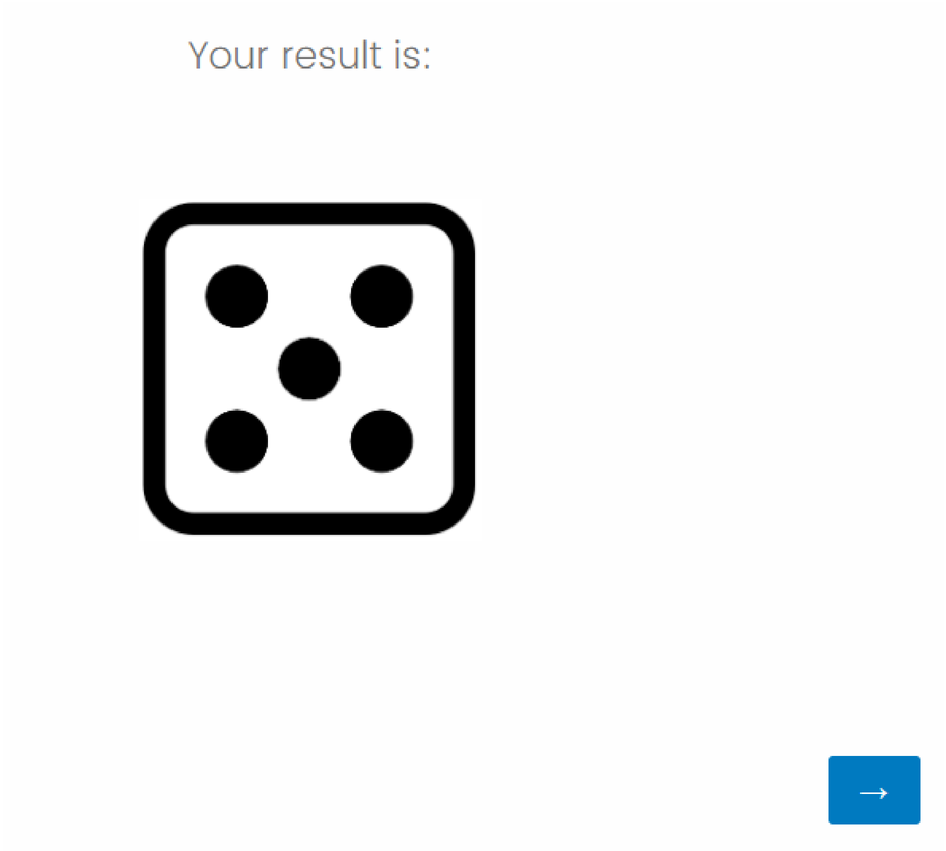
○ No

Roll-dice screen



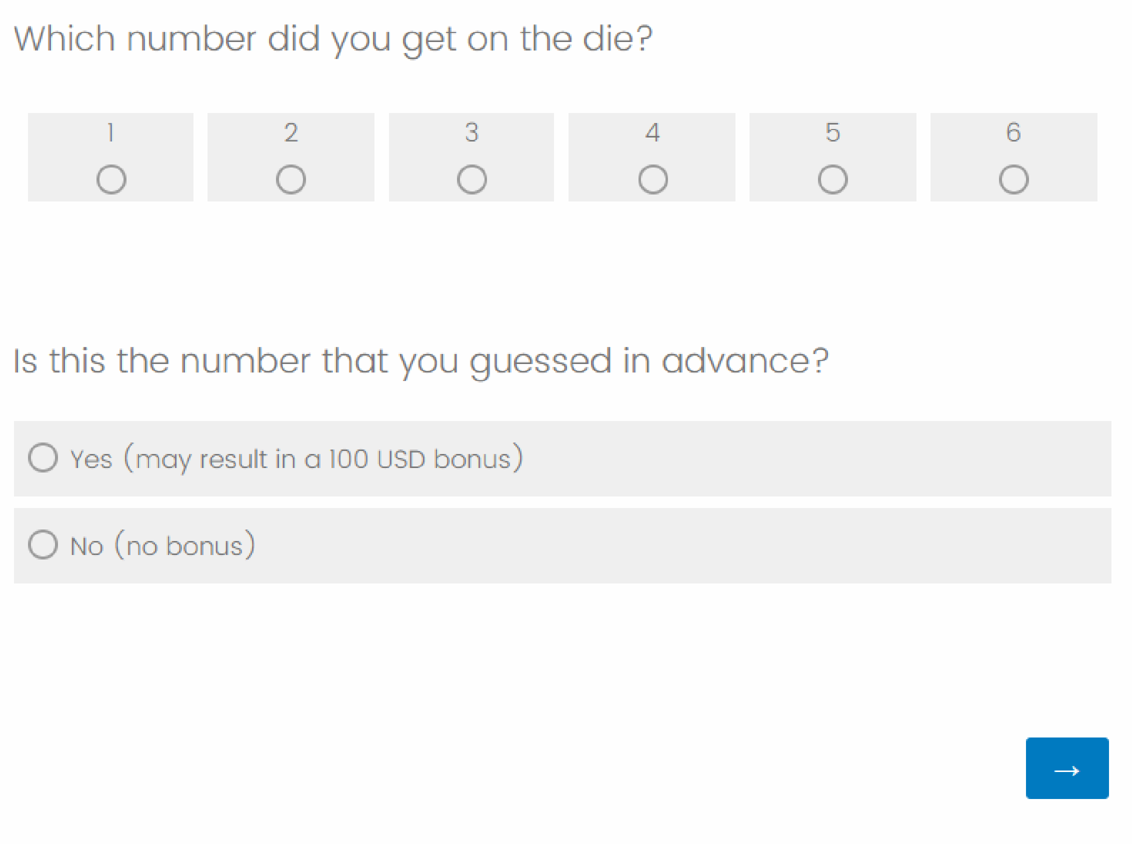
Roll the die by clicking on the arrow.

Result screen (Example)



Your result is:

Report result



Which number did you get on the die?

1	2	3	4	5	6
○	○	○	○	○	○

Is this the number that you guessed in advance?

○ Yes (may result in a 100 USD bonus)

○ No (no bonus)

Chapter 2: Dishonesty at the doctor's office. What influences patients' disclosure of accurate information?

Pablo Soto-Mota
Norwegian School of Economics

Andrés Castañeda Prado MD.
Universidad Nacional Autónoma de México

Alfonso Gulias-Herrero MD. DPhil.
National Institute of Medical Sciences and Nutrition "Salvador Zubirán"

Adrián Soto-Mota MD. DPhil.
National Institute of Medical Sciences and Nutrition "Salvador Zubirán"

Abstract

Dishonesty is often motivated by image concerns. However, lying to appear more socially desirable can come at a cost. There is plenty of evidence that some patients misreport sensitive information to their doctors to avoid feeling judged or because they are ashamed of their habits. Misreporting by patients is a cause for concern because it can significantly affect diagnoses, treatment and, ultimately, patients' health. In this study, we examine whether it is possible to increase the accuracy of the information patients disclose to their doctors. We conducted a field experiment in Mexico City, covering 6 clinics and 1,164 patients. We experimentally varied the script doctors used to ask about two health-related habits: alcohol and vegetable consumption. This setup enabled us to test four simple question formulations designed to motivate truth-telling. In addition, we took advantage of the random assignment of patients to female or male doctors in three clinics to explore the effect of gender alignment on patients' reports. We find that the framing we tested does not affect patients' answers. In a smaller sample, however, we also observe some evidence of an effect of gender alignment on the information patients share. Male patients report drinking less alcohol to female doctors than to their male counterparts. Our results imply that simple strategies used by doctors to deter patients from providing socially desirable answers might not be enough to improve the quality of their reports. However, the alignment of some characteristics may enhance honesty.

Keywords: Doctor-patient relationship, sensitive questions, lying, gender alignment

JEL: D91, I12, C93

2.1 Introduction

There is plenty of evidence that some patients lie to their doctors by hiding, distorting and fabricating relevant information. In observational studies, researchers have documented that patients lie about their smoking habits (Almadana Pacheco et al., 2020), drug use (Caulley et al., 2020), psychosocial problems (Mule et al., 2022) and general wellbeing (Forder et al., 2020). Levy et al. (2018) estimated that between 61% and 81% of patients have avoided disclosing information to their doctors on many health-related topics, such as having an unhealthy diet, not complying with prescriptions or disagreeing with the clinician's recommendation. A similar study documented that patients also avoid disclosing information when faced with imminent threats such as depression and abuse (Levy et al., 2019). Misreporting by patients is a cause for concern among clinicians since it limits their capacity to help (Fagerlin, 2019).

Levy et al. (2018; 2019) observed that the leading causes of misreporting by patients are related to image concerns. Patients do not want to feel judged; they do not want to know how wrong their behaviour is or they are ashamed of their health-related habits. Agaku et al. (2014) also found that patients lie when they are worried about their privacy. Similar studies, based on self-reports, have found that patients also lie because they disagree with the opinion of their physicians (Sidora-Acoleo et al., 2008), because they fear discrimination (Mosack et al., 2013) or they have generally low trust in the health system (Churchill et al., 2000). In a survey we conducted on Mexican patients, described in Appendix 1, people expressed similar reasons for lying to their doctor, and that their misreporting is motivated by image-promotion. In other words, they present themselves as having healthier habits than they actually have.

Patients' misreporting behaviour also impedes the collection of relevant statistical information for the design of health-related policies. For example, Castelo-Branco et al. (2010) observed that people answer questions about sexual habits differently in anonymous surveys and clinical consultations, making it difficult to obtain precise measurements. In addition, Turner et al. (2009) reported that participants in a randomised control trial lied about the effect of a treatment designed to improve their health out of politeness to researchers and to avoid criticism. Since patients' dishonesty increases costs for individuals and society, it is a cause of concern for policymakers.

In this study, we explore factors that influence why people lie to improve their image during clinical consultations. We focus on reports of two health-related habits that patients tend to lie about: alcohol consumption and following a healthy diet (measured as eating vegetables). The vast majority of patients who lie about these habits underreport their alcohol consumption and overreport their vegetable intake. Consequently, we interpret an increase (decrease) in reported alcohol consumption (vegetable consumption) across otherwise comparable groups as increases in honesty. We ask two research questions. Firstly, can we increase patient reports of health-related habits by changing how doctors frame their questions? Secondly, what is the role of gender matching between patients and doctors in misreporting behaviour?

To answer these questions, we conducted a field experiment in collaboration with six COVID-19 triage clinics in Mexico City during the second half of 2021. Each doctor experimentally varied how the questions about alcohol and vegetable consumption were framed. We compare a control group using standard, direct framing with alternatives suggested by the literature: (i) reminding patients about the importance of the information; (ii) giving patients a physical format; (iii) asking these questions in a non-judgmental way; (iv) giving patients enough time to remember and answer. Secondly, we took advantage of the natural randomisation of patient assignment to doctors in three clinics. This provides an exogenous variation in the gender alignment of patients and doctors. We use this randomisation to explore the role of gender matching on patient reports.

The unique environment of our field experiment has some advantages for identifying causal effects. Firstly, all the patients attended the clinic for the same reason: they were worried about having COVID-19. This reduced the effect of differences in health issues. In addition, all the clinics were free and open to the public. Moreover, interviews in these clinics were standardised among doctors, reducing the potential differential effects of their consultation style. Finally, it was very unlikely that doctors and patients had any previous clinical interaction or a realistic expectation of meeting again, reducing the effect of past interactions.

In our analysis, we found that the four tested variations of the questions do not affect patients' answers. Since we used the question framing commonly recommended to doctors as potential solutions to dishonesty, this null effect is meaningful for clinical practice. It indicates that the strategies doctors use may give them a false sense of security about their patients' reports, which could be detrimental to their health results. Notably, in a post-study survey, we found that doctors in a different sample relied on similar strategies to those we tested to reduce patients' dishonesty, which presumably had no effect.

In addition, our analyses suggest an effect of gender alignment on male patients. In particular, consistent with the image-promotion mechanism, male patients report healthier habits to female doctors. When interviewed by a female doctor, the number of men who admit to drinking alcohol is halved. This translates to an overall decrease in the average report of alcohol of 0.25 SD. Nonetheless, this analysis would have gained statistical power from a bigger and more diverse pool of doctors. Therefore, although this result should not be interpreted as conclusive, it does suggest a direction for future research.

Our results have economic and health policy implications. Nonetheless, they should be interpreted within the limitations of our field experiment. While we observed actual interactions between doctors and patients, they were in the context of COVID-19 and thus outside regular clinical practice. Future research should confirm our observations in other settings that are closer to regular clinical practice and explore the mechanisms behind our findings. Finally, our null results on the effect of the framing of questions are an invitation to other studies to focus on finding simple, effective and scalable alternatives to help patients provide truthful reports.

The rest of this paper is structured as follows. In the next section, we present the contribution our study makes to the related literature. We then describe the experimental research design, the context of the study, our pre-registered hypotheses and our empirical strategy. In the fifth section, we provide an overview of the characteristics of our sample. The sixth section presents the results from the field experiment and discusses their significance. Finally, we summarise our findings and discuss their implications.

2.2 Related literature and contribution

People value their image. Individuals make an effort to ensure they are liked, respected and perceived as socially adequate (Ariely et al., 2009). However, sometimes these efforts involve lying. For example, people lie about their physical traits in online dating apps (Toma et al., 2008), about their competence during job interviews (Weiss & Feldman, 2006) and about pro-social behaviours such as voting (Dellavigna et al., 2017). Participants in laboratory experiments lie to appear more skilful (Falk & Szech, 2020), confident (Ewers & Zimmermann, 2015) and, ironically, they also lie to appear more honest (Barron, 2019; Bašić & Quercia, 2022). Previous literature has focused on situations where people's motivation for lying is financial (Barron et al., 2021) or where maintaining a good image has clear extrinsic benefits, for example, a date or a job. There is less evidence of situations where lying to improve one's image without a clear extrinsic motive actually comes at a cost. An example of such a situation is when patients misreport relevant health information during clinical consultations with their doctors. In these cases, patients try to maintain a positive image and avoid feeling judged for their habits at the cost of jeopardising the effectiveness of their treatment and future health.

In addition to being relevant to the literature on lying for image, this paper relates to the impact of matching socioeconomic characteristics of patients and doctors on health outcomes. Previous observational studies have found that similarities between doctors and patients in characteristics such as gender (Greenwood et al., 2020; Tsugawa et al., 2017), race (Hill et al., 2020) and socioeconomic status (Kristiansen & Sheng, 2022) improve clinical results such as survival rates and patient satisfaction (Schmittdiel et al., 2000). Likewise, in a laboratory setting, Howe et al. (2022) observed that white male patients react more quickly to treatment when their doctor is also white and male. They hypothesised that the effect was driven by biases related to stereotypes of the warmth and competence of some doctors. Similarly, Alsan et al. (2019) noted in a field experiment that patients are more willing to accept invasive check-ups when their race aligns with that of their doctor.

Gender alignment between patients and doctors might affect health outcomes due to differences in doctors' consultation styles (Roter et al., 2002) and interpretation of patients' reports (Adams et al., 2008). While male doctors focus on technical aspects, female doctors pay more attention to preventive and screening questions (Bertakis, 2009; Henderson & Weisman, 2001). In addition, doctor-patient communication might differ when genders align. Mohajer and Endut (2020) observed in recorded clinical consultations in Malaysia that male doctors made more cooperative interruptions with their male patients than with their female patients. The effect of gender alignment between doctors and patients on other aspects of their interaction is mixed. Some studies suggest that gender match only has a positive effect when the patient picks the doctor and not when the doctor is randomly assigned (Lambert, 2016; Schmittdiel et al., 2000).

This study advances the existing literature in three ways. Firstly, we document that image concerns may lead to dishonest behaviours when patients visit doctors, which could have negative health consequences. Secondly, to the best of our knowledge, our study is the first to use a field experiment to test common simple alternatives to framing sensitive questions in clinical settings. Finally, this paper is the first to estimate the effect of gender matching on the accuracy of patients' reports to their doctors, which can help us to understand the observed impacts of the patient-doctor gender match on health outcomes.

2.3 Experimental research design

In our study, we compare patients' reporting of two health-related habits: alcohol and vegetable consumption. We randomise patients in two ways. Firstly, we vary how doctors frame their questions about health-related habits. Secondly, we take advantage of the randomisation of patients to male and female doctors. We test whether the gender of the doctor or the way they frame their questions increases the patient's honesty. We complement our observations in the field experiment with two online surveys. The first survey, distributed before the field experiment to a different sample of 1,065 Mexican patients, sheds light on the prevalence, topics and direction of misreporting to their doctors. We distributed the second survey after the trial ended to a sample of 154 Mexican doctors to explore their perspectives on patients' dishonesty and their strategies to prevent this behaviour. We refer to these results throughout the text and present the details of each survey in Appendices 1 and 2, respectively.

2.3.1 Context of the experiment

We conducted the experiment during clinical consultations at six different triage COVID-19 temporary clinics and hospitals between July and December 2021. The Government of Mexico City opened these clinics in April 2020 to help in the early detection of serious COVID-19 cases (CDMX, 2020a). In these clinics, patients received an initial evaluation of their symptoms and were given medical advice (CDMX, 2020b). During the months that our experiment ran, each of these clinics received between 15 and 25 patients per day (Ramírez, 2021).

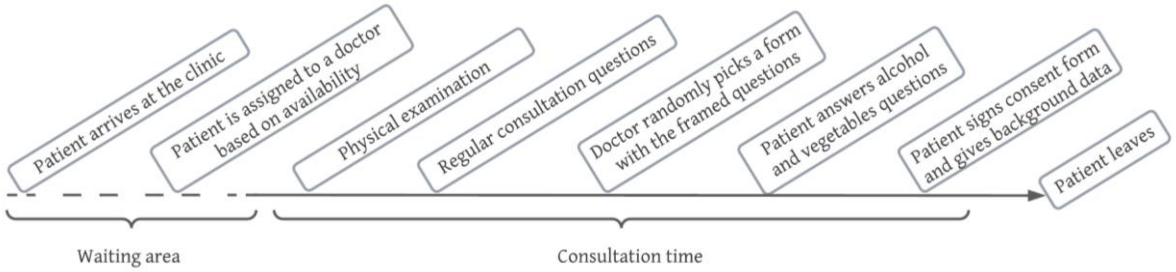
The unique setting of our experiment offered four advantages that increased the internal validity of our analyses. Firstly, all the patients went to the clinic for the same reason: concerns related to COVID-19. This commonality eliminated non-observed differences in patient reporting across different medical conditions. Secondly, all clinics were free and open to the public. This reduced differences in behaviours due to the cost of attending a hospital. Thirdly, the interaction between doctors and patients was standardised; doctors had to evaluate the gravity of each case and not directly treat the patient. This reduced the effect of different consultation styles. Fourthly, it was unlikely that a patient and a doctor had any previous clinical interaction. Likewise, there was little chance the patient would encounter the same doctor in the future. Therefore, the effect of previous encounters with a doctor or the expectation of a future relationship was minimal.

We recruited 28 doctors across the six clinics, 20 men and 8 women, who worked at the clinics and volunteered to conduct the experiment. They were all between 24 and 31 years of age and had recently graduated from medical school. We gave them precise instructions on how to implement the treatments. We also informed them that the study's objective was to analyse different aspects of the doctor-patient relationship. However, we did not explain our research questions in detail to them. We kept in contact with them during the weeks of the experiment to distribute materials and collect data.

2.3.2 Procedure

Figure 1 presents the typical procedure patients had to go through when visiting the clinics participating in the study. After arriving at the clinic, the patient received a queue number and waited in a designated area. Then, a health worker assigned the patient to a doctor based on the patient's number and the doctor's availability. It was impossible to know the exact number of patients arriving each day or when a doctor would be available. So, neither the patients nor the doctors could select each other conditional on attending one of the clinics. Therefore, we can safely assume that the assignment of patients to doctors within each clinic was as good as a random assignment. In addition, the eight female doctors who volunteered to conduct the experiment worked in three of the six clinics. Consequently, patients in those clinics were randomly assigned to either a male or a female doctor.

Chapter 2. Figure 1. Experiment procedure


Patient arrives at the clinic
Patient is assigned to a doctor based on availability
Physical examination
Regular consultation questions
Doctor randomly picks a form with the framed questions
Patient answers alcohol and vegetables questions
Patient signs consent form and gives background data
Patient leaves
Waiting area
Consultation time

Patients answered the health-related questions that we analyse in our experiment during the clinical consultation. Each doctor implemented all the experimental conditions explained in Section 2.4.2. We gave each doctor a set of folders with the instructions and formats for each treatment in random order. We instructed doctors to take a new one for each patient. In Appendix 10, we present the instructions for each treatment, including the formats the doctors followed when asking health-related questions.

The clinical consultation began with a physical examination and questions to determine the severity of COVID-19 symptoms. Then, the doctors took the instructions for one of the treatments from the folder. They asked about alcohol and vegetable consumption in accordance with the instructions. The experiment modified no other part of the clinical consultation.

At the end of the consultation, the doctor asked the patient if they consented to having their data used for research purposes. If the patient agreed, the doctor asked them to fill out a form comprising basic socioeconomic data such as age and education level. Following standard practices and the ethical guidelines outlined in our Institutional Review Board (IRB) applications, we did not collect any data on participants who did not agree to share their information. At the end of the study, we asked doctors about the number of patients who did not consent to sharing their information, which they reported only occurred in a handful of cases. Therefore, we have no reason to suspect that self-selection in the experiment endangers the identification of causal effects.

2.3.3 Reasons for focusing on alcohol and vegetable consumption

The reasons for focusing on alcohol and vegetable consumption are threefold. Firstly, limiting alcohol consumption and increasing the consumption of vegetables are important actions to prevent various adverse health conditions. There is a scientific consensus that alcohol is harmful to health (Nyberg et al., 2020) as well as evidence that consuming vegetables helps prevent various health conditions (Wallace et al., 2020). This is relevant in the Mexican context. In the National Health Survey in Mexico, 33% of adults stated that they consume alcoholic beverages at least once a month. Moreover, 60% of those who consume alcohol said they did so to excess at least once during that period (INSP, 2021). Likewise, Mexico has a high prevalence of chronic illnesses linked to poor nutritional habits (Popkin, 2015; Pérez-Ferrer et al., 2020). For example, 45% of respondents in the national survey answered that they eat vegetables less than three days a week. Moreover, in the survey of Mexican clinicians shown in Appendix 2, doctors answered that information about these habits is relevant for guiding the treatment of people with symptoms of COVID-19.

Secondly, alcohol and vegetable consumption are behaviours that are suitable for identifying causal effects in our experiment, as standard questions can be asked about both behaviours. However, doctors cannot easily and objectively test these behaviours, and thus must rely on patients' reports. In the case of alcohol, we used a measure based on units of alcohol per week that we present in Appendix 10; this measure is based on the one first developed by Sobell and Sobell (1995). For vegetable consumption, clinicians asked participants how many days per week they consumed at least one portion of green vegetables.[1]

Finally, the existing evidence shows that patients misreport their alcohol and vegetable consumption consistently to promote image. In the patient survey described in Appendix 1, most respondents who admitted to having lied about these topics stated they underreported their alcohol consumption (87.5%) and overreported their vegetable consumption (96.7%). Consequently, holding everything else constant, for any true average consumption of alcohol or green vegetables, if two comparable groups have different average levels of reported consumption, we interpret the difference as driven by differences in honesty.

2.3.4 Experimental conditions

2.3.4.1 Variations in the framing of the questions

To investigate the effect of how doctors ask patients for information, we vary how they frame their questions about alcohol and vegetable consumption. In the literature on survey methodology, this is defined as a *loading strategy* since we are not changing anything else in the environment other than the wording to encourage respondents to answer truthfully (Näher & Krumpal, 2012). We compare these alternatives to a *control* condition, in which doctors ask a standard, direct question.

Our treatments are inspired by the previous literature on why patients lie to their doctors.; for example, not wanting to be judged, being ashamed of their health-related habits (Levy et al., 2018) or privacy concerns (Agaku et al., 2013). Similarly, we also based our treatments of common recommendations to physicians on strategies for preventing misreporting by patients. Our treatments attempt to reduce these concerns in four different ways: reminding patients of the importance of their information; reassuring

[1] During the experiment, we asked patients about their consumption of "green vegetables". We emphasise "green" to avoid frequent confusion at the clinics about what a vegetable is. In the text, we use "green vegetable" and "vegetable" interchangeably.

them of the privacy of their answers; generating a non-judgemental environment; giving them time to reflect. Notably, in the post-trial survey of clinicians, shown in Appendix 2, we noticed that clinicians say that they follow similar strategies to prevent misreporting by patients.[2]

Control Group: Doctors asked about health behaviours directly. For alcohol, the doctors asked, "*How many alcoholic beverages do you drink during an average week? Both during weekdays and weekends.*" For vegetables, they asked "*How many times a week do you eat at least one portion of green vegetables?*"

T1 (Importance): Before asking about health behaviours, doctors emphasised the importance of answering honestly to get a proper diagnosis by saying, "*I need you to tell me the truth about the following questions so that I can help you*."

T2 (Format): Patients answered a written questionnaire about their health behaviours.

T3 (Non-judgemental question): While conducting the consultation, the physician said a sentence to the patient to increase empathy and trust, and indicate that the clinic is a non-judgmental environment. The sentence was the same for all doctors: *"I am asking you this because my friends sometimes eat too few vegetables and drink beer*."

T4 (Reflect): The doctor instructed the patient to think carefully about their health behaviours and gave them enough time to remember accurately. The doctor said, "*Take 10 seconds to remember precisely how many drinks you had in the last two weeks*." Then, "*And now, try to remember, how many days did you have green vegetables in the last week?*"

2.3.4.2 Assignment of patients to male and female doctors

We took advantage of the randomisation of patients to male and female doctors in three of the six participating clinics. Since doctors and patients could not pick each other, being assigned to a doctor of a particular gender was as good as random conditional on attending the clinic. Therefore, there were two experimental conditions for male and female patients: gender aligned and gender non-aligned consultations.

[2] Since medical visits are private, we have no means of measuring compliance on the part of the doctors. However, we have no reason to believe that this was a problem. Doctors voluntarily agreed to implement these treatments during the preparation phase, and even when asked directly during the months of implementation, they did not report avoiding or having a preference for any of them.

To evaluate the effect of gender alignment, we pre-registered a comparison of patients' reports of alcohol and vegetable consumption in each condition. We exclude the data from the hospitals without female clinicians in this analysis. In Appendix 4, we present a balance table of the assignment of patients to male and female doctors.

Male doctors interviewed significantly more patients than female doctors. This is because of the lower number of female doctors in the study, the rotation of the doctors in their positions and the unpredictable development of the pandemic. Although this significantly limits the statistical power of our analysis, we observe balance across conditions for other variables.

2.4 Empirical strategy: hypotheses and estimating equations

Our pre-registered hypotheses are the following:[3]

H1. Question framing will have an effect on the reports. The average report of alcohol consumption (vegetable consumption) will be higher (lower) under the treatment conditions than in the control group.

H2. Gender matching will have an effect on the reports. The average report of alcohol consumption and vegetable consumption will be different for gender-matched interviews than for interviews without gender matching.

To test our first hypotheses, we estimate the following equation:

$$Y_i = \alpha_i + \beta_d Treatment_d + \phi_p + \delta f(\mathrm{X}_i) + \varepsilon_{ihd} \quad (1)$$

In Equation (1), Y_i is either the alcohol consumption or vegetable consumption report. $Treatment_d$ is a vector of binary variables, one per treatment condition. Therefore, $d \in \{Importance, Format, Non-judgemental,\ Reflect, Control\}$. X_i is a vector of covariates that includes age, gender and education level of the patient, their level of oxygen as a proxy of health at that moment, and a binary variable that takes the value of 1 if they regularly go to the doctor. Since doctors work at different clinics, have different abilities and characteristics, we also add doctor-fixed effects, ϕ_p, to increase precision.

To test our hypothesis about the effect of gender concordance on patients' reports, we estimate the following equation:

$$Y_i = \alpha_i + \beta_1 M.Pat_i + \beta_2\, F.Dr_h + \beta_3 (M.Pat. \times F.Dr)_{ihd} + \phi_h + \delta f(\mathrm{X}_i) + \varepsilon_{ihd} \quad (2)$$

[3] Before conducting our experiment, we preregistered our hypotheses in the OSF Registry. We also share other resources there such as the data, replication do-files and materials. These documents can be accessed at: https://archive.org/details/osf-registrations-jk2yq-v1

As in the previous case, in Equation (2), Y_i is either alcohol consumption or vegetable consumption report. $M.Pat_i$ is a binary variable that takes the value of 1 if the patient is male, and $F.Dr_h$ takes the value of 1 if the doctor is female. X_i is the vector of covariates previously described. Consistent with the previous analysis, we add ϕ_h as a vector of hospital fixed effects. Finally, we cluster standard errors at the doctor level using bootstrapping methods.

In Equation (1), the parameters β_d are the causal effect of question framing on patients' reports. Similarly, in Equation (2), the parameters of interest are β_2 and β_3. First, β_2 shows whether female patients' reports are different when they meet a female doctor than when they meet a male doctor. Likewise, β_3 is the causal effect of meeting a female doctor instead of a male doctor for male patients relative to female patients. Hence, β_2 and β_3 provide the causal effect of gender concordance for male and female patients.

In both analyses, our identification strategy relies on the idea that given the randomisation across groups, based on the absence of dishonesty, all conditions should report the same level of average alcohol and vegetable consumption. In addition, based on the literature and our survey to patients, we can safely assume that some patients are dishonest and that the vast majority of them do so to protect image. This implies that all groups, everything else constant, differ from the true levels of consumption in the same direction. Therefore, we can interpret differences in the average reports of consumption across groups as differences in dishonesty. However, we cannot make claims about how far these reports are from the truth.

2.5 Sample characteristics

The participants in the field experiment were adult patients who attended one of the six collaborating hospitals in Mexico City in the second half of 2021. Appendix 3.1 provides a list of the names of the participating hospitals. In accordance with our research protocol, we excluded patients younger than 18 years old; patients who could not read or write and were not accompanied by someone who could assist them in understanding the informed consent letter; patients who, for whatever reason, could not answer the questions by themselves; and patients who refused to participate in the study. Moreover, the implementation of the experiment was dependent on the availability of doctors at the hospitals and the development of the pandemic.

Table 1 presents the socioeconomic characteristics. In congruence with our power calculations and pre-registered analysis, the experiment had 1,164 participants[4]. The average age of participants was 40.5 years old (SE = 15.92), 55% identified as female, and 65% had an education lower or equal to high school level. In Appendices 4 and 5, we show that, as expected, the experimental groups are balanced on observables, so we can also assume that they are balanced in their true level of consumption of alcohol and vegetables.

Chapter 2. Table 1. Sample characteristics.

		Total
N		1,164
Age		40.595
(In years)		(15.927)
Female patient		0.559
		(0.497)
Oxygen saturation		94.290
(In percentage)		(2.297)
Visits the doctor frequently		0.205
(1 = Yes, 0 = No; self-reports)		(0.404)
Female doctor		4.38%
≤ High school		65.81%
Hospital	**CEDA**	2.13%
	INSZ	21.81%
	RL	2.64%
	TLC	15.08%
	TLP	25.55%
	XCH	32.79%

Note: SD in parentheses. We present the names of the hospitals and characteristics of participants per hospital in the Appendix.

Clinics differed from each other in both the type and number of patients they received during the experiment. In addition, a different proportion of doctors from each clinic participated in the experiment. The table in Appendix 3.2 shows the sample characteristics per clinic. The main reason for the observed differences is that these clinics are located in areas of Mexico City with populations of diverse socioeconomic status.

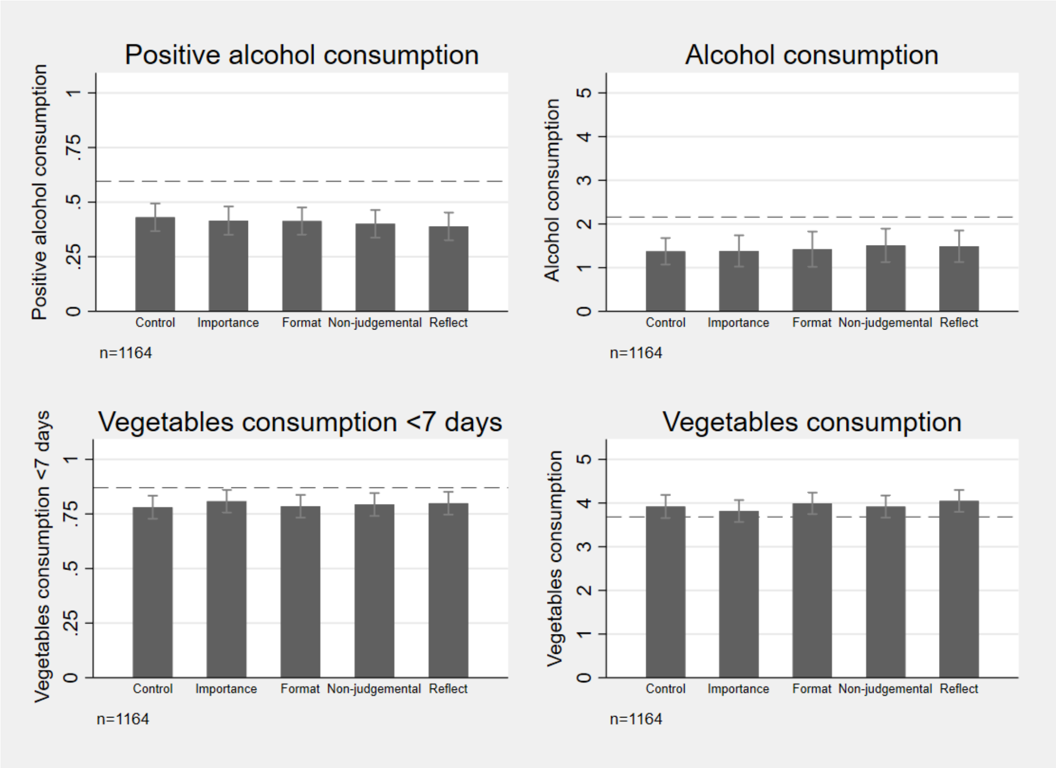
[4] Using the G-Power software (Faul et al., 2007), we calculated that with a statistical power of 0.8 and a significance level of 0.05, we needed 200 participants in each treatment group to detect a minimum standardised effect of 0.28 (MDE). An effect of this size is equivalent to an increase of one drink in the average report of alcohol consumption. Since there is no precedent in the literature on this type of experiment, we used information from the 2016-2017 National Survey of Drug, Alcohol and Tobacco Consumption to estimate the MDE. We present details of these calculations in Appendix 4.

2.6 Experimental Results

2.6.1 Results of the framing of the questions

We do not observe any effect of the different framing of questions on the patients' reports. The left-side panels of Figure 2 show the proportion of patients who reported drinking alcohol and not eating vegetables daily. The right-side panels show the average alcohol and vegetable consumption of all patients. The dotted line in each panel is the average answer on alcohol and vegetable consumption in the anonymous pre-test survey of patients presented in Appendix 1. Although the samples in the field experiment and the survey differ, we use the survey as a baseline to contextualise patients' reports in a situation where image concerns should be less important.

Chapter 2. Figure 2. Results by treatment



Note: Error bars indicate 95% CIs.

In Table 2, we present the results from the estimation of Equation (1). Columns (1) and (4) present the estimations of the effects of question framing on alcohol and vegetable consumption as described in Equation (1). In columns (2) and (5), we present similar results but using as a dependent variable admitting to drinking alcohol and eating vegetables less than seven days a week. Finally, in columns (3) and (6), we use as a dependent variable the consumption of those who admit to drinking alcohol or not eating vegetables every day. We observe null effects for all treatments.

Chapter 2. Table 2. Results by treatment

	(1)	(2)	(3)	(4)	(5)	(6)
	Alcohol consumption			**Vegetable consumption**		
	Alc.	**Alc. > 0**	**Alc. if > 0**	**Vg.**	**Vg. < 7**	**Vg. if < 7**
Importance	0.036	-0.013	0.103	-0.115	0.034	-0.019
	(0.234)	(0.043)	(0.467)	(0.177)	(0.037)	(0.140)
Format	0.021	-0.025	0.242	0.066	0.009	0.100
	(0.253)	(0.043)	(0.514)	(0.175)	(0.037)	(0.138)
Non-judgmental	0.188	-0.018	0.464	0.000	0.016	0.031
	(0.234)	(0.042)	(0.456)	(0.180)	(0.037)	(0.140)
Reflect	0.058	-0.046	0.330	0.126	0.023	0.220
	(0.219)	(0.042)	(0.419)	(0.175)	(0.037)	(0.143)
Constant	4.434	0.913*	5.490	1.747	1.070	3.807*
	(2.945)	(0.521)	(8.702)	(3.074)	(0.799)	(2.083)
Baseline level	1.373	0.43	3.191	3.92	0.78	3.054
(SD)	(2.375)	(0.496)	(2.706)	(2.074)	(0.415)	(1.444)
R^2	0.147	0.184	0.106	0.134	0.102	0.049
N	1164	1164	477	1164	1164	923

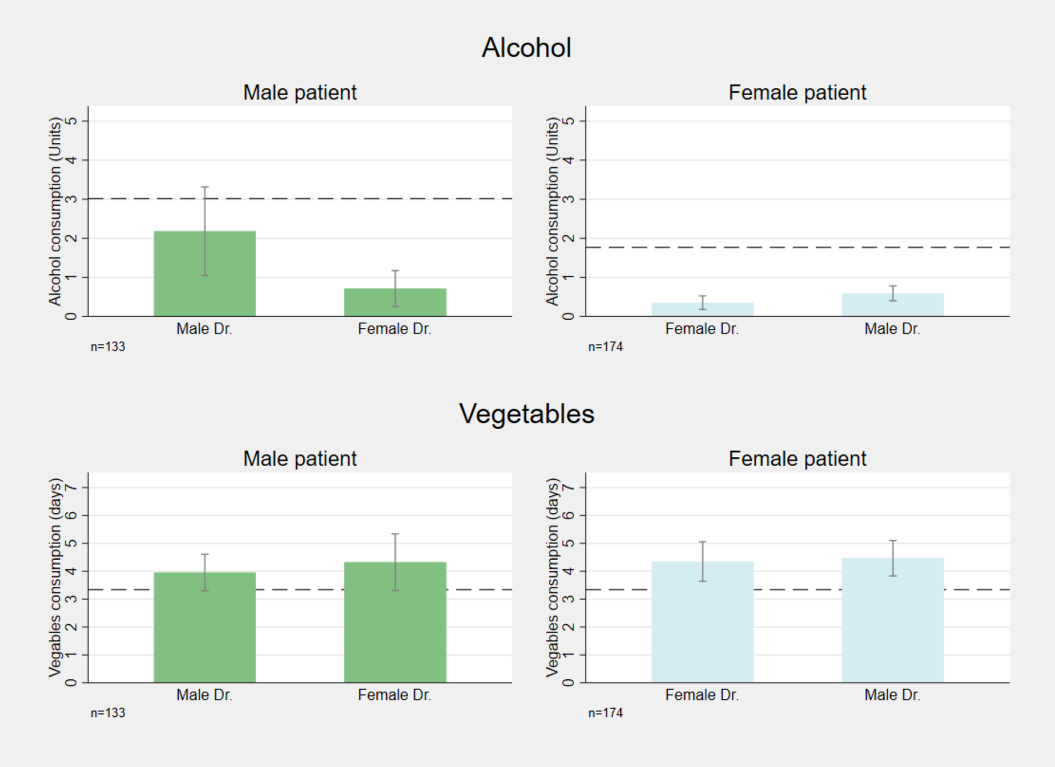
Note: All columns include controls: age, oxygen level during physical examination, education and if the patient visits hospital frequently, and hospitals' and doctors' fixed effects. Robust SEs in parentheses. Baseline level is the mean of the control group. *p<0.1, ** p<0.05, ***p<0.001.

2.6.2 Results of gender alignment

In Figure 3, we present the results of the gender matching between doctors and patients. As before, the dotted line is the average level of reported consumption in the pre-test survey. Focusing on male patients, we note a difference in their reported alcohol consumption depending on the gender match or mismatch with their doctor. Men assigned to a male doctor state that they drink three units per week. In sharp contrast, men in gender-mismatched consultations state that they drink one unit of alcohol per week. If the gender of the doctor was insignificant, we would expect these numbers to be identical and close to what patients reported in the online survey. Therefore, our result suggests that male patients underreport their alcohol intake more to female doctors than to their male counterparts.

Likewise, male patients report eating green vegetables on average one more day per week when they meet a female doctor compared to when they meet a male doctor. Nonetheless, this difference is not statistically significant. We do not observe any equivalent effects for female patients.

Chapter 2. Figure 3. Results for gender alignment



Note: 95% CIs.

In Table 3, we present the results from the estimation of Equation (2). As in Table (2), columns (1) and (4) present the estimations of the effects of gender alignment on alcohol and vegetable consumption. In addition, in columns (2) and (5), we use as a dependent variable a binary variable that takes the value of 1 when the patient admits to drinking alcohol and not consuming vegetables every day. Thus, this variable shows the proportion of individuals who report behaving in a fully desirable way. Finally, in columns (3) and (6), we use as a dependent variable the consumption of those who accept drinking alcohol or not eating vegetables daily.

We observe that, in general, men report drinking more alcohol and eating fewer vegetables than women. In addition, we see that the effect of female doctors on male patients' reports is primarily explained by the lower proportion of men who admit to drinking alcohol. In other words, an effect on the intensive margin of alcohol drinkers. In line with the image concern hypothesis, when a female doctor conducts the interview, 28.2% fewer men state that they drink at least one unit of alcohol per week.

Chapter 2. Table 3. Results for gender alignment

	(1)	(2)	(3)	(4)	(5)	(6)
	Alcohol consumption			**Vegetable consumption**		
	Alc.	**Alc. > 0**	**Alc. if > 0**	**Vg.**	**Vg. < 7**	**Vg. if < 7**
Male patient	1.583***	0.245***	2.075***	-0.581**	0.052	-0.480**
	(0.414)	(0.060)	(0.768)	(0.261)	(0.053)	(0.212)
Female doctor	-0.399	-0.031	-0.537	-0.017	-0.088	-0.430
	(0.247)	(0.101)	(0.825)	(0.538)	(0.111)	(0.389)
Male Pat. & Fem. Dr.	-1.210**	-0.282**	-0.764	0.625	-0.047	0.520
	(0.521)	(0.136)	(1.307)	(0.710)	(0.140)	(0.497)
Constant	4.676	0.064	24.794	13.820***	-0.760	10.000***
	(7.272)	(1.107)	(26.704)	(5.031)	(1.220)	(3.530)
R^2	0.100	0.136	0.135	0.080	0.052	0.094
N	307	307	102	307	307	224

Note: All columns include controls: age, oxygen level during physical examination, education and if the patient visits hospital frequently and hospitals' fixed effects. Bootstrap SEs in parentheses. *p<0.1, ** p<0.05, ***p<0.001.

2.6.3 The doctors' perspective on misreporting by patients

How doctors and patients interact is key to the effectiveness of treatment (Pearson & Raeke, 2000). Communication is one of the most important elements of interactions in clinical settings (Ha & Longnecker, 2010). In this study, we did not find any effect of the framing of questions about health-related habits on the honesty of patients. However, our results suggest that gender matching impacts patients' truthfulness when speaking to their doctors. In particular, men lie more about sensitive topics when interviewed by female clinicians.

To complement our findings and make sense of their potential significance for health, we distributed a survey to a sample of Mexican doctors (N = 152). Appendix 2 presents the details of this part of the study. In the survey, we explore to what extent doctors are aware of the misreporting behaviour and their strategies to correct it. We found that doctors think that most patients lie about health-related habits, predicting that 64% of patients lie to them. In contrast, only 30% of participants in our survey of patients state that they lie. Since patients may also lie in online questionnaires, we interpret this figure as a lower bound, indicating that at least a third of patients have lied to their doctors. These misreports endanger the accuracy of diagnoses, the effectiveness of their treatment and, ultimately, patients' health.

Doctors also indicated that they use methods like the ones we tested to attempt to reduce the problem. In the survey, they stated that they try to remind their patients about the importance of their answers and make efforts to make them feel comfortable and assure them about the privacy of their information. These strategies correspond to common recommendations found in the medical and non-academic literature. The fact that these methods are found to be ineffective could be useful for informing professionals and motivating studies to explore ways of improving the doctor-patient relationship.

Finally, we asked doctors about which type of patients they believe lie more than others. Our results reveal that both male and female doctors believe that male patients misreport more to female doctors. In addition, we designed a vignette study, described in Appendix 2, in which doctors answered how credible they found male patient reports of healthy habits in comparison to those that accepted having a less healthy lifestyle. We found that female doctors interpreted the reports of “healthy” patients as less credible compared to male doctors. This result was robust even after controlling for years of experience and the number of patients seen per week. Although more research is needed to investigate whether female doctors adjust more than their male colleagues for patients' reports in regular practice, our results suggest that more awareness is needed on the effect of gender matching to ensure they can provide better care for their patients.

2.7 Conclusions

The economics literature on dishonesty has focused on situations where lying is financially beneficial to individuals. Recent studies have focused on scenarios where individuals lie to acquire and protect a socially desirable image. However, in most of these studies, participants have extrinsic motives for creating a good image. For example, a person might lie in a job interview to make a good impression, and they might get the job. Little is known about situations where creating a falsely positive image harms the individual. An example of such a situation is patients lying to their doctors: they might present a healthier version of themselves to enhance their social image, but doing so puts their health at risk. Although lying to a doctor potentially carries a high cost, evidence shows that it is still prevalent among patients.

In this study, we examine whether patient reports on two health-related habits change when questions are framed differently. Likewise, we test whether patients' answers change when their gender matches their doctor's. We conducted a field experiment in Mexico City in association with six COVID-19 triage clinics and hospitals during 1,174 unique medical consultations. In our study, doctors varied how they asked questions about alcohol and vegetable consumption. These variations were inspired by the literature on patients’ dishonesty, survey design studies and what Mexican doctors say they do to prevent misreporting. In addition, patients were assigned to either a male or a female doctor. We observed no treatment effects with respect to how doctors asked these questions. However, we found suggestive evidence that male patients lie more to female doctors than to their male counterparts. That is, when reporting to female doctors, men claim to drink less and eat more vegetables than when reporting to male doctors. We do not observe similar effects for female patients.

In a follow-up survey of Mexican doctors, we observed three main findings. Firstly, doctors know that some patients lie during clinical consultations. Secondly, they believe that male patients lie more to female doctors. Finally, respondents answered that they frame questions in the same way we asked doctors to do in our study to minimise the likelihood of patient misreporting. For example, they give them information about why their answers matter or try to make them feel comfortable. However, doctors should be aware that this might not be enough. Other studies should look for more effective ways to enhance the accuracy of patients' reports.

We interpret our results on the effect of doctor gender as suggestive evidence that patients lie to their physicians because of image concerns. For reasons that future research should investigate in more detail, male patients might feel a stronger need to be liked by or to impress female doctors. In line with this reasoning, previous research has, for example, found that men make larger charity donations to female recipients (List & Price, 2009). An alternative mechanism expressed by the doctors in our study suggests that male patients might feel less comfortable revealing their true habits to female physicians because they see them as motherly figures. Future research should replicate our results in other contexts, such as job interviews or surveys, to explore possible explanatory mechanisms.

References

Adams, A., Buckingham, C. D., Lindenmeyer, A., McKinlay, J. B., Link, C., Marceau, L., & Arber, S. (2008). The influence of patient and doctor gender on diagnosing coronary heart disease. *Sociology of Health and Illness*, *30*(1), 1–18. https://doi.org/10.1111/j.1467-9566.2007.01025.x

Agaku, I. T., Adisa, A. O., Ayo-Yusuf, O. A., & Connolly, G. N. (2014). Concern about security and privacy, and perceived control over collection and use of health information are related to withholding of health information from healthcare providers. *Journal of the American Medical Informatics Association*, *21*(2), 374–378. https://doi.org/10.1136/amiajnl-2013-002079

Almadana Pacheco, V., Benito Bernáldez, C., Luque Crespo, E., Perera Louvier, R., Rodríguez Fernández, J. C., & Valido Morales, A. S. (2020). Do COPD patients lie about their smoking habit? *Atencion Primaria*, *52*(8), 523–528. https://doi.org/10.1016/j.aprim.2020.05.014

Alsan, M., Garrick, O., & Graziani, G. (2019). Does diversity matter for health? Experimental evidence from Oakland. *American Economic Review*, *109*(12), 4071–4111. https://doi.org/10.1257/aer.20181446

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–555. https://doi.org/10.1257/aer.99.1.544

Barron, K. (2019). *Markets and Choice Lying to appear honest* (Discussion Paper No. SP II 2019-307). Retrieved from www.wzb.eu

Barron, K., Kajackaite, A., Saccardo, S.,(2021). *Lying for Image*. Retrieved from http://dx.doi.org/10.2139/ssrn.3977054

Bašić, Z., & Quercia, S. (2022). The influence of self and social image concerns on lying. *Games and Economic Behavior*, *133*, 162–169. https://doi.org/10.1016/j.geb.2022.02.006

Bertakis, K. D. (2009). The influence of gender on the doctor-patient interaction. *Patient Education and Counseling*, *76*(3), 356–360. https://doi.org/10.1016/j.pec.2009.07.022

Castelo-Branco, C., Palacios, S., Ferrer-Barriendos, J., & Alberich, X. (2010). Do patients lie? An open interview vs. A blind questionnaire on sexuality. *Journal of Sexual Medicine*, *7*(2 PART 2), 873–880. https://doi.org/10.1111/j.1743-6109.2009.01575.x

Caulley, L., Takhar, A., Bast, F., Surda, P., & Hopkins, C. (2020). Nondisclosure of cocaine use in the rhinology practice: A retrospective study of 27 patients. *Clinical Otolaryngology*, *45*(4), 608–610. https://doi.org/10.1111/coa.13535

CDMX. (2020a). Conferencia de prensa 02 de abril de 2020e. Retrieved 7 August 2022, from https://servicios.covid19.cdmx.gob.mx/comunicacion/nota/version-estenografica-conferencia-de-prensa-02-de-abril-de-2020

CDMX. (2020b). Permanece Ciudad de México en Semáforo Epidemiológico Naranja con Alerta del 3 al 9 de agosto. Retrieved 7 August 2022, from https://servicios.covid19.cdmx.gob.mx/comunicacion/nota/permanece-ciudad-de-mexico-en-semaforo-epidemiologico-naranja-con-alerta-del-3-al-9-de-agosto

Churchill, R., Allen, J., Denman, S., Williams, D., Fielding, K., & von Fragstein, M. A. R. T. I. N. (2000). Do the attitudes and beliefs of young teenagers towards general practice influence actual consultation behaviour? *BJGP*, 50(461), 953-957.

Dellavigna, S., List, J. A., Malmendier, U., & Rao, G. (2017). Voting to tell others. *Review of Economic Studies*, *84*(1), 143–181. https://doi.org/10.1093/restud/rdw056

Eckerd, S., DuHadway, S., Bendoly, E., Carter, C. R., & Kaufmann, L. (2021). On making experimental design choices: Discussions on the use and challenges of demand effects, incentives, deception, samples, and vignettes. *Journal of Operations Management*, *67*(2), 261–275. https://doi.org/10.1002/joom.1128

Eifler, S., & Petzold, K. (2019). *Validity Aspects of Vignette Experiments: Expected "What-If" Differences Between Reports of Behavioral Intentions and Actual Behavior*. Retrieved from www.wiley.com/go/Lavrakas/survey-research

Ewers, M., & Zimmermann, F. (2015). Image and misreporting. *Journal of the European Economic Association*, *13*(2), 363–380. Retrieved from https://www.jstor.org/stable/24538904

Fagerlin, A. (2019). When patients lie | AAMC. Retrieved 28 April 2020, from https://www.aamc.org/news-insights/insights/when-patients-lie

Falk, A., & Szech, N. (2020). Competing Image Concerns: Pleasures of Skill and Moral Values, (066), 1–21. Retrieved from https://polit.econ.kit.edu/downloads/papers/WP_Pleasures_of_skill_Falk_Szech.pdf

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Forder, P. M., Rich, J., Harris, S., Chojenta, C., Reilly, N., Austin, M. P., & Loxton, D. (2020). Honesty and comfort levels in mothers when screened for perinatal depression and anxiety. *Women and Birth*, *33*(2), e142–e150. https://doi.org/10.1016/j.wombi.2019.04.001

Greenwood, B., Hardeman, R., Huang, L., & Sojourner, A. (2020). Physician–patient racial concordance and disparities in birthing mortality for newborns. *PNAS*, *117*(35), 21194–21200.

Ha, J. F., & Longnecker, N. (2010). Doctor-patient communication: a review. Ochsner Journal, 10(1), 38-43. doi: 10.3329/jbcps.v32i2.26036.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(8), 2395–2400. https://doi.org/10.1073/pnas.1416587112

Henderson, J. T., & Weisman, C. S. (2001). Physician Gender Effects on Preventive Screening and Counseling: An Analysis of Male and Female Patients' Health Care Experiences. *Medical Care*, *39*(12), 1281–1292.

Hill, A., Jones, D., & Woodworth, L. (2020). Physician-Patient Race-Match Reduces Patient Mortality. *SSRN Electronic Journal*. Retrieved from https://www.aamc.org/data/workforce/reports/439214/workforcediversity.html

Howe, L. C., Hardebeck, E. J., Eberhardt, J. L., Markus, H. R., & Crum, A. J. (2022). White patients' physical responses to healthcare treatments are influenced by provider race and gender. *Proceedings of the National Academy of Sciences*, *119*(27), 2017. https://doi.org/10.1073/pnas.2007717119

INSP. (2021). *Encuesta Nacional de Salud y Nutrición 2021 sobre Covid-19. Resultados Nacionales*. Retrieved from https://ensanut.insp.mx/encuestas/ensanutcontinua2021/informes.php

Kristiansen, I.L., and Sheng, S.Y., Doctor Who? The Effect of Physician-Patient Match on The SES-Health Gradient (19 July 2022). CEBI Working Paper Series, 2022, Available at http://dx.doi.org/10.2139/ssrn.4152145

Lambert, M. J. (2016). Does Client-Therapist Gender Matching Influence Therapy Course or Outcome in Psychotherapy? *Evidence Based Medicine and Practice*, *02*(02). https://doi.org/10.4172/2471-9919.1000108

Levy, A. G., Scherer, A. M., Zikmund-Fisher, B. J., Larkin, K., Barnes, G., & Fagerlin, A. (2018). Prevalence of and Factors Associated With Patient Nondisclosure of Medically Relevant Information to Clinicians. *JAMA Network Open*, *1*(8). https://doi.org/ilCilli?w252.1WcB06BX1.CmlisCnmok

Levy, A. G., Scherer, A. M., Zikmund-Fisher, B. J., Larkin, K., Barnes, G., & Fagerlin, A. (2019). Assessment of Patient Nondisclosures to Clinicians of Experiencing Imminent Threats. *JAMA Network Open*, *2*(8), 1–6.

List, J. A., & Price, M. K. (2009). The role of social connections in charitable fundraising: Evidence from a natural field experiment. Journal of Economic Behavior & Organization, 69(2), 160-169. https://doi.org/10.1016/j.jebo.2007.08.011

Mendel, R., Hamann, J., Traut-Mattausch, E., Bühner, M., Kissling, W., & Frey, D. (2010). "What would you do if you were me, doctor?": Randomised trial of psychiatrists' personal v. professional perspectives on treatment recommendations. *British Journal of Psychiatry*, *197*(6), 441–447. https://doi.org/10.1192/bjp.bp.110.078006

Mohajer, L., & Endut, N. (2020). The role of gender and status in communication between doctors and patients in Malaysian contexts. *Kajian Malaysia*, *38*, 89–108. https://doi.org/10.21315/KM2020.38.S1.6

Mosack, K. E., Brouwer, A. M., & Petroll, A. E. (2013). Sexual identity, identity disclosure, and health care experiences: is there evidence for differential homophobia in primary care practice?. *Women's Health Issues*, 23(6), e341-e346. https://doi.org/10.1016/j.whi.2013.07.004

Mule, V., Reilly, N. M., Schmied, V., Kingston, D., & Austin, M. P. V. (2022). Why do some pregnant women not fully disclose at comprehensive psychosocial assessment with their midwife? Women and Birth, 35(1), 80–86. https://doi.org/10.1016/j.wombi.2021.03.001

Näher, A. F., & Krumpal, I. (2012). Asking sensitive questions: The impact of forgiving wording and question context on social desirability bias. Quality and Quantity, 46(5), 1601–1616. https://doi.org/10.1007/s11135-011-9469-2

Nyberg, S. T., Singh-Manoux, A., Pentti, J., Madsen, I. E. H., Sabia, S., Alfredsson, L., … Kivimäki, M. (2020). Association of Healthy Lifestyle with Years Lived without

Major Chronic Diseases. JAMA Internal Medicine, 1–9. https://doi.org/10.1001/jamainternmed.2020.0618

Pearson, S. D., & Raeke, L. H. (2000). Patients' trust in physicians: many theories, few measures, and little data. Journal of General Internal Medicine, 15(7), 509-513. doi: 10.1046/j.1525-1497.2000.11002.x.

Pérez-Ferrer, C., Auchincloss, A. H., Barrientos-Gutierrez, T., Colchero, M. A., de Oliveira Cardoso, L., Carvalho de Menezes, M., & Bilal, U. (2020). Longitudinal changes in the retail food environment in Mexico and their association with diabetes. Health and Place, 66(February). https://doi.org/10.1016/j.healthplace.2020.102461

Popkin, B. M. (2015). Nutrition Transition and the Global Diabetes Epidemic. Current Diabetes Reports, 15(9), 1–8. https://doi.org/10.1007/s11892-015-0631-4

Ramírez, E. (2021, January). La experiencia en un módulo de triage. Gaceta, Facultad de Medicina. Retrieved from https://gaceta.facmed.unam.mx/index.php/2021/01/12/la-experiencia-en-un-modulo-de-triage/

Roter, D. L., Hall, J. A., & Aoki, Y. (2002). Physician Gender Effects in Medical Communication A Meta-analytic Review. JAMA, 288(2), 756–764. Retrieved from http://jama.jamanetwork.com/

Schmittdiel, J., Grumbach, K., Selby, J. V, & Quesenberry, C. P. (2000). Effect of Physician and Patient Gender Concordance on Patient Satisfaction and Preventive Care Practices. Journal of General Internal Medicine, 15, 761–769. https://doi.org/https://doi.org/10.1046/j.1525-1497.2000.91156.x

Sidora-Arcoleo, K., Yoos, H. L., Kitzman, H., McMullen, A., & Anson, E. (2008). Don't ask, don't tell: parental nondisclosure of complementary and alternative medicine and over-the-counter medication use in children's asthma management. *Journal of Pediatric Health Care*, 22(4), 221-229. https://doi.org/10.1016/j.pedhc.2007.07.001

Sobell, L. C., & Sobell, M. B. (1995). Alcohol Consumption Measures. Retrieved 2 April 2020, from http://pubs.niaaa.nih.gov/publications/AssessingAlcohol/sobell.pdf

Tilburt, J. C., Miller, F. G., Jenkins, S., Kaptchuk, T. J., Clarridge, B., Bolcic-Jankovic, D., … Curlin, F. A. (2010). Factors that Influence Practitioners' Interpretations of Evidence from Alternative Medicine Trials: A Factorial Vignette Experiment Embedded in a National Survey. Med Care, 48(4), 341–348.

Toma, C. L., Hancock, J. T., & Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. Personality and Social Psychology Bulletin, 34(8), 1023–1036. https://doi.org/10.1177/0146167208318067

Tsugawa, Y., Jena, A. B., Figueroa, J. F., Orav, E. J., Blumenthal, D. M., & Jha, A. K. (2017). Comparison of hospital mortality and readmission rates for medicare patients treated by male vs female physicians. JAMA Internal Medicine, 177(2), 206–213. https://doi.org/10.1001/jamainternmed.2016.7875

Turner, A. N., De Kock, A. E., Meehan-Ritter, A., Blanchard, K., Sebola, M. H., Hoosen, A. A., … Ellertson, C. (2009). Many vaginal microbicide trial participants acknowledged they had misreported sensitive sexual behavior in face-to-face interviews. Journal of Clinical Epidemiology, 62(7), 759–765. https://doi.org/10.1016/j.jclinepi.2008.07.011

Wallace, T. C., Bailey, R. L., Blumberg, J. B., Burton-Freeman, B., Chen, C. y. O., Crowe-White, K. M., … Wang, D. D. (2020). Fruits, vegetables, and health: A comprehensive narrative, umbrella review of the science and recommendations for enhanced public policy to improve intake. Critical Reviews in Food Science and Nutrition, 60(13), 2174–2211. https://doi.org/10.1080/10408398.2019.1632258

Weiss, B., & Feldman, R. S. (2006). Looking Good and Lying to Do It: Deception as an Impression Management Strategy in Job Interviews. Journal of Applied Social Psychology (Vol. 36).

Ethics statement

The IRB at the *Norwegian School of Economics* approved the protocol of our experiment, application Number NHH-IRB 26/21. Moreover, the IRB of *The National Institute of Health Sciences and Nutrition "Salvador Zubirán"* approved the protocol with reference 3612. The data collection complied with good clinical practice standards. We anonymised the data immediately after the end of the experiment period. We share all materials, including translations into English, in an open repository (here).

CRediT authors' statement

Pablo Soto-Mota: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project administration. **Andrés Castañeda**: Conceptualisation, Methodology, Investigation, Writing – Review & Editing, Resources. **Adrián Soto-Mota:** Investigation, Writing – Review & Editing, Resources. **Alfonso Gulias:** Investigation, Writing – Review & Editing.

Acknowledgements

We thank the doctors and administrative personnel at the six COVID-19 triage clinics who volunteered to conduct this study. In addition, we would like to thank Eirik Berger, Daniel Parra, Mathias Ekström, Andreas Haller, Sam Hirshman, Ernesto Gavassa-Perez, Arne Nasgowitz, and Akshay Moorthy for their advice and comments on earlier drafts of this paper. Likewise, we would like to thank those who attended and provided useful feedback during the presentation of the preliminary results of this project during the *2nd NHH-Max Planck Institute Workshop*, the *FAIR Midway Conference*, the brown-bag Seminar organised by the *People in Government Lab* at *Oxford University*, the *8th Conference "Sobre México"* at Universidad Iberoamericana Mexico City, the UCSD Spring School in Behavioural Economics, and the Newcastle Experimental Economics Workshop 2022.

Appendix

Appendix 1: Survey of patients

Objective

Before conducting the field experiment, we distributed an online survey to obtain a first approximation of whether, how and when patients avoid disclosing truthful information to their doctors. In addition, we asked for levels of consumption of alcohol and vegetables to compare them with reports in the clinical setting. Finally, we explored the critical assumption for the analysis of the field experiment: that patients misreport their information in a socially desirable way. Thus, lying about the consumption of alcohol and vegetables tends to go in opposite directions. Our survey was based on a study published by Levy and co-authors (2018), where they used a non-probabilistic online sample of patients in the US to explore the factors associated with patients' nondisclosure of information.

Method

Our questionnaire comprised a total of 26 questions and took 10 minutes to complete. The questionnaire can be accessed in the Online Appendix at the OSF repository for the project. After obtaining informed consent, the survey had three sections. First, participants answered an online questionnaire survey about the effect of different framing of questions on patients' answers about health-related habits. In the second part, we asked participants whether they had lied to their doctors and their predictions of similar behaviour in others. Finally, the third part of the survey asked for basic background information. We recruited a sample of adults in Mexico through social media; they answered the survey from 27 May to 17 June 2021.

Participants

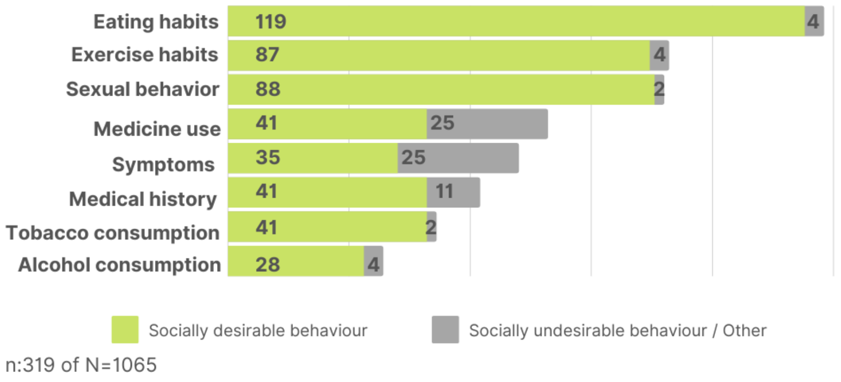
The survey link had 1,245 interactions. Of these interactions, all except three agreed to sign the informed consent letter. After we excluded responses from those who did not complete the questionnaire (n = 177) or did not accept the consent form (n = 3), we had 1,065 observations for the analysis.

Results

In our sample, 29.95% of respondents (n = 319) admitted that they had lied or withheld information from their doctors. The most cited reasons for lying are related to not feeling empathy from their doctor, being embarrassed and avoiding information. The five most common reasons are not wanting to be judged (75.9% of patients who admitted to having lied; n = 319), being ashamed of admitting their behaviour (42.0%), avoiding criticism of their behaviour (33.9%), not wanting to take too much time (17.2%), and not wanting to be seen as a difficult patient (16.0%).

The most lied about topics in our survey were eating habits (39% of respondents), sexual behaviour (32%) and exercise (29%). These topics are all not readily observable to the doctor and have a social desirability dimension. A critical assumption of the analysis is that patients who lie do so in a socially desirable way. Figure 4 explores the validity of this assumption. We observe that people report lying in socially acceptable ways across all categories. Exceptions are medicine use and symptoms, where it is not clear what would be more socially accepted.

Figure 4. Frequency and social desirability of lies



Finally, some participants (n = 208) were asked directly about their alcohol and green vegetable consumption in the survey experiment, with 37.98% answering that they do not drink alcohol and 12.5% stating that they eat at least one portion of vegetables daily. On average, participants said they drink 2.158 (SD = 3.289) units of alcohol a week (Male Avg. = 3.015, SD = 4.1; Female Avg. = 1.777, SD = 2.787). Furthermore, on average, they answered that they eat a portion of green vegetables 3.682 (SD = 1.9) days a week (Male Avg. = 3.343, SD = 1.765; Female Avg. = 3.883, SD = 1.946). We use these numbers as references for the field experiment.

Respondents in this survey, though diverse, are not representative of the Mexican population. When compared to the patients in the field experiment, we observed that they have a similar age (M = 42.25 years, SD = 14.04), though a higher proportion of survey participants were female (66.48% of respondents) and were more educated since only

17.68% had an education lower or equal to high school. Nonetheless, we have no reason to believe that these differences would imply differences in true alcohol and vegetable consumption habits between the survey participants and field experiment participants.

Appendix 2: Survey of doctors

Objective

To help interpret our results and explore the implications of the results for health, we designed and implemented an online survey on a non-probabilistic sample of medical doctors. In the survey, we explored their beliefs and experiences of misreporting by patients. We also included a vignette study to analyse whether doctors find some reports more credible than others in a context similar to our field experiment. In particular, we analysed whether male and female doctors differ in how credible they perceive the same information to be.

Vignette studies are commonly used in social sciences research (Hainmueller et al., 2015). Vignettes are a simplified description of real-life scenarios that, by varying specific elements, help explore how individuals make evaluations in multidimensional scenarios (Eckerd et al., 2021). These instruments are not meant to be externally valid but measure approximate behaviour by understanding intentions (Eifler & Petzold, 2019). In medical sciences, vignette studies have been used to analyse how physicians interpret evidence from medical trials (Tilburt et al., 2010) and the perspective of doctors during different types of doctor-patient communication (Mendel et al., 2010).

Method

The survey lasted, on average, 10 minutes and comprised 26 questions. The Online Appendix provides a translation of the questionnaire. The questionnaire had four sections. The first was the informed consent letter and confirmation that participants were medical doctors who attended to patients in Mexico. The second was the vignette experiment on the credibility of patients' reports. There was then a section about their perception and beliefs around the problems of misreporting by patients. Finally, the fourth section was a basic background questionnaire that included questions about respondents' medical practice and education. We recruited a sample of Mexican medical doctors. We distributed our questionnaire through medical channels on social media and personal messaging from 11 to 15 July 2022.

In our vignette study, we described a similar situation as our field experiment. We randomly assigned participants to one of two experimental conditions. In the first (*Unhealthy report*), the patient reported alcohol and vegetable consumption closer to the average of what male patients told male doctors in our field experiment: 4 alcoholic drinks per week and 3 days with at least one portion of green vegetables. In the second condition (*Healthy report*), the patient declared they had healthier levels of consumption than those we observed: 1 alcoholic drink per week and eating green vegetables 6 days a week. The text of the vignette was the following:

> *José is a 45-year-old patient with high school education. He is at the clinic because he has had a fever, dry cough, shortness of breath and general malaise. He suspects that he has COVID-19 although he says that he has complied with all the recommended protection measures and has received two doses of vaccines.*
>
> *Upon inspection, you observe that his oxygen saturation level is 90%. He has normal pulse and pressure. However, he has had diabetes for five years and is overweight.*
>
> *When asked about his habits, José answers that he exercises moderately every two or three days, that he drinks* ***[Unhealthy:4/ Healthy:1]*** *alcoholic beverages a week, that he eats at least one serving of green vegetables* ***[Unhealthy:3, Healthy:6]*** *days a week, and that he only smokes one cigarette a week when he goes out with his friends.*

We asked respondents about the likelihood of the patient's report being accurate. We expected to see differences in the credibility of the reports on the consumption of alcohol and vegetables. Specifically, we expected doctors to think that the *unhealthy reports* were more credible than the *healthy* ones. In addition, we also asked how credible the reports were on factors that did not vary, namely, about the protective measures against COVID-19 and his exercise and smoking habits. We expected to see no differences in the credibility of these reports. Finally, we asked doctors how important these questions were to diagnosing and guiding the treatment of this patient.

Participants

Of all the survey link interactions (N = 279), we eliminated those with incomplete answers (n = 103) and those who did not confirm they were medical doctors (n = 22). This left us with 154 valid answers. In our sample, 44.15% of respondents identified as female. The average age was 34.81 years (SD = 10.41), and the average number of years of experience attending to patients was 7.31 (SD = 5.41). Half of our sample consisted of specialists, and 72.73% worked full- or part-time in public hospitals. Finally, 68.63% attended to at least 21 patients a week.

Results

Medical doctors predict that most patients lie about health-related behaviours during clinical consultations (mean = 0.649, SD = 1.575). Male and female doctors do not differ in this prediction. Likewise, we do not observe systematic differences in the perceptions of misreporting by different groups. However, male and female doctors perceive that male patients lie more to female doctors (Appendix 3.6).

Doctors believe that the topics that patients lie most about are eating habits (73.37% of respondents), exercise (57.792%), alcohol consumption (48.052%) and sexual behaviour (38.312%). These results align with those from the survey of patients. Similarly, doctors also believe that the main reason for misreporting by patients is image concerns, which aligns with patients' answers. These concerns are not wanting to be judged (77.92%), not wanting to admit to shameful behaviour (64.29%) and avoiding criticism of their behaviour (51.30%). However, 45.46% of physicians said that one of the three main reasons why patients lie is to avoid having to make difficult changes in their habits, while only 9% of patients said that was one of the reasons.

Table 3. Results vignette study in the survey to doctors.

	(1) **Alcohol** **b/se**	**(2)** **Alcohol** **b/se**	**(3)** **Vegetables** **b/se**	**(4)** **Vegetables** **b/se**
Credibility of the healthy report	-0.715	-0.732	-0.554	-0.631
[Less alc. & more vegs = 1]	(0.678)	(0.708)	(0.639)	(0.631)
Female Dr.	1.380**	1.281*	0.015	0.212
[female Dr. = 1]	(0.696)	(0.723)	(0.675)	(0.656)
Healthy report & Female Dr.	-1.726*	-1.688*	-0.564	-0.443
	(0.973)	(0.994)	(0.857)	(0.854)
Specialist		-0.101		0.161
[Is specialist = 1]		(0.689)		(0.588)
Works in Public H.		0.745		0.996**
[Works in Public Hosp. = 1]		(0.589)		(0.466)
>30 patients p. week		-0.653		0.508
[Attends > 30 patients p. week = 1]		(0.500)		(0.465)
Experience		0.007		0.043
[In years]		(0.063)		(0.057)
Constant	4.738***	4.593***	4.190***	2.725***
	(0.484)	(0.780)	(0.514)	(0.686)
Controls	NO	YES	NO	YES
R^2	0.064	0.057	0.008	0.037
N	154	154	154	154

Note: OLS regression with robust standard errors in parenthesis. *p<0.01, **p<0.005, ***p<0.001

As expected, in the vignette study, participants in both experimental conditions perceived the reports without variations as equally credible (Appendix 2.1). However, we found significant differences across conditions in the reported levels of credibility of alcohol and vegetable consumption before adding controls. In particular, the condition where the patient reported healthier levels of consumption of alcohol and vegetables was perceived as less credible. On closer inspection, and after adding controls and the interaction of the gender of the doctor, we observe this effect only for the consumption of alcohol. Finally, we found that this effect is driven by female doctors. That is, female doctors find the patient's healthy report more credible than male doctors.

A2.1 Survey to doctors: perceptions of lying among patients

		Male Drs.	Female Drs.	All
	N	86	68	154
Who lies more?	Male patients	34.88%	35.29%	35.07%
	Female patients	5.81%	7.35%	6.49%
	They lie the same	59.30%	57.35%	58.44%
Who lies more?	Younger patients	30.23%	17.65%	24.68%
	Older patients	24.42%	30.88%	27.27%
	They lie the same	45.35%	51.47%	48.05%
Who lies more?	Patients in Private Clinics	10.47%	8.82%	9.74%
	Patients in Public Clinics	41.86%	48.53%	44.81%
	They lie the same	47.67%	42.65%	45.46%
Male patients lie more to…	Female doctors	47.67%	47.06%	47.40%
	Male doctors	3.49%	0.00%	1.95%
	To both equally	48.84%	52.94%	50.65%
Female patients lie more to…	Female doctors	20.93%	11.77%	16.88%
	Male doctors	19.77%	17.65%	18.83%
	To both equally	59.30%	70.59%	64.29%
Younger patients lie more to…	Female doctors	17.44%	16.18%	16.88%
	Male doctors	6.98%	5.88%	6.49%
	To both equally	75.58%	77.94%	76.62%
Older patients lie more to…	Female doctors	17.44%	27.94%	22.08%
	Male doctors	6.98%	10.29%	8.44%
	To both equally	75.58%	61.77%	69.48%

A2.2 Survey to doctors: tests in vignette study

	C1	C2	Combined	Diff.
	Mean/SD	Mean/SD	Mean/SD	Diff/SE
COVID-19 precautions	4.539	4.013	4.276	0.527
	(2.650)	(2.586)	(2.623)	(0.422)
Exercise habits	3.513	3.590	3.551	-0.077
	(2.676)	(2.713)	(2.686)	(0.434)
Alcohol***	5.355	3.872	4.614	1.483
	(3.097)	(3.004)	(3.130)	(0.492)
Vegetables*	4.197	3.397	3.797	0.800
	(2.989)	(2.360)	(2.710)	(0.433)
Tobacco use	3.461	2.795	3.128	0.666
	(3.202)	(2.655)	(2.947)	(0.473)
Credibility index*	4.213	3.562	3.887	0.652
	(2.374)	(2.137)	(2.273)	(0.364)

Note: *p<0.1, ** p<0.05, ***p<0.001

Appendix 3: Names of the hospitals

Hospital	**CEDA**	Central de Abastos (Iztapalapa)
	INSZ	Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán
	RL	Hospital Rubén Leñero
	TLC	Jurisdicción Santitaria de Tláhuac
	TLP	Clínica Covid Tlalpan
	XCH	Clínica Covid Xochimilco

Appendix 4: Balance table with respect to treatment

		Treatment				
		Control	**T1 Importance**	**T2 Format**	**T3 Friends**	**T4 Memory**
N		237	224	242	232	229
Age		40.211	40.723	39.628	40.720	41.764
(In years)		(16.546)	(16.237)	(14.779)	(15.963)	(16.157)
Female patient		0.549	0.580	0.537	0.599	0.533
		(0.499)	(0.495)	(0.500)	(0.491)	(0.500)
Oxygen saturation		94.333	94.214	94.384	94.241	94.271
(In percentage)		(1.856)	(1.843)	(1.814)	(3.391)	(2.210)
Visits the doctor frequently		0.181	0.174	0.198	0.246	0.227
(1=Yes, 0=No; self-reports)		(0.386)	(0.380)	(0.400)	(0.431)	(0.420)
Female doctor		3.80%	4.02%	5.79%	2.59%	5.68%
≤ High school		66.67%	62.95%	66.12%	68.53%	64.63%
Hospital	**CEDA**	1.79%	3.31%	1.29%	2.18%	2.15%
	INSZ	22.77%	21.90%	21.12%	21.83%	21.99%
	RL	1.79%	2.07%	1.72%	2.62%	2.23%
	TLC	15.18%	14.88%	15.95%	14.41%	15.12%
	TLP	25.00%	25.21%	28.02%	25.33%	25.60%
	XCH	33.48%	32.65%	31.90%	33.62%	32.90%

Note: SD in parentheses. Percentages are from the total in each hospital.

Appendix 5: Balance table with respect to gender match

		Without gender matching			With gender matching		
		Male Dr.	Female Dr.	Total	Male Dr.	Female Dr.	Total
N		154	31	185	102	20	122
Age (In years)		47.156	42.581	46.389	51.284	39.450	49.344
		(16.719)	(12.999)	(16.215)	(15.861)	(13.809)	(16.102)
Oxygen saturation (In percentage)		94.305	94.452	94.330	93.951	93.350	93.852
		(2.100)	(2.173)	(2.107)	(2.103)	(5.102)	(2.798)
Visits the doctor frequently (1=Yes, 0=No; self-reported)		0.435	0.419	0.432	0.461	0.400	0.451
		(0.497)	(0.502)	(0.497)	(0.501)	(0.503)	(0.500)
≤ High school		65.58%	54.84%	63.78%	64.71%	50.00%	62.30%
Treatment	**T1-Importance**	24.03%	22.58%	23.78%	18.63%	10.00%	17.21%
	T2-Format	20.78%	16.13%	20.00%	17.65%	20.00%	18.03%
	T3-Friends	20.13%	29.03%	21.62%	20.59%	25.00%	21.31%
	T4-Memory	18.18%	12.90%	17.30%	21.57%	10.00%	19.67%
	T1-Importance	16.88%	19.36%	17.30%	21.57%	35.00%	23.77%

Note: SD in parentheses. Percentages are from the total in each hospital.

Appendix 6: Sample characteristics by hospital

		Hospital						
		CEDA	INSZ	RL	TLC	TLP	XCH	Total
N		25	256	26	176	298	383	1,164
Age (In years)		37.520	48.426	48.731	38.670	41.326	35.326	40.595
		(12.346)	(16.154)	(17.255)	(15.118)	(16.353)	(13.424)	(15.927)
Female patient		0.400	0.582	0.577	0.545	0.581	0.543	0.559
		(0.500)	(0.494)	(0.504)	(0.499)	(0.494)	(0.499)	(0.497)
Oxygen saturation (In percentage)		94.760	94.426	90.731	94.693	94.128	94.352	94.290
		(2.047)	(1.854)	(4.341)	(1.854)	(2.998)	(1.651)	(2.297)
Visits the doctor frequently (1=Yes, 0=No; self-reported)		0.000	0.484	0.423	0.131	0.258	0.010	0.205
		(0.000)	(0.501)	(0.504)	(0.338)	(0.438)	(0.102)	(0.404)
Female doctor		60.00%	12.11%	19.23%	0.00%	0.00%	0.00%	4.38%
≤ High school		68.00%	63.28%	57.69%	69.32%	59.40%	71.28%	65.81%
Treatment	**T1-Importance**	20.70%	26.92%	20.46%	19.46%	20.37%	20.36%	20.27%
	T2-Format	19.92%	15.39%	19.32%	18.79%	19.58%	19.24%	19.25%
	T3-Friends	20.70%	19.23%	20.46%	20.47%	20.63%	20.79%	20.70%
	T4-Memory	19.14%	15.39%	21.02%	21.81%	19.32%	19.93%	20.10%
	T1-Importance	19.53%	23.08%	18.75%	19.46%	20.10%	19.67%	19.68%

Note: SD in parentheses. Percentages are from the total in each hospital.

Appendix 7: Comfort treatments

	(1) More Comfort b/se	(2) More Comfort b/se
T1-Importance	0.009	0.010
	(0.047)	(0.046)
T2-Format	-0.003	-0.001
	(0.046)	(0.045)
T3-Friends	-0.023	-0.020
	(0.046)	(0.045)
T4-Memory	0.023	0.021
	(0.046)	(0.045)
Constant	0.532***	-0.456
	(0.032)	(0.536)
Controls + FEs	NO	YES
R^2	-0.003	0.039
N	1164	1164
Note: *p<0.1, ** p<0.05, ***p<0.001		

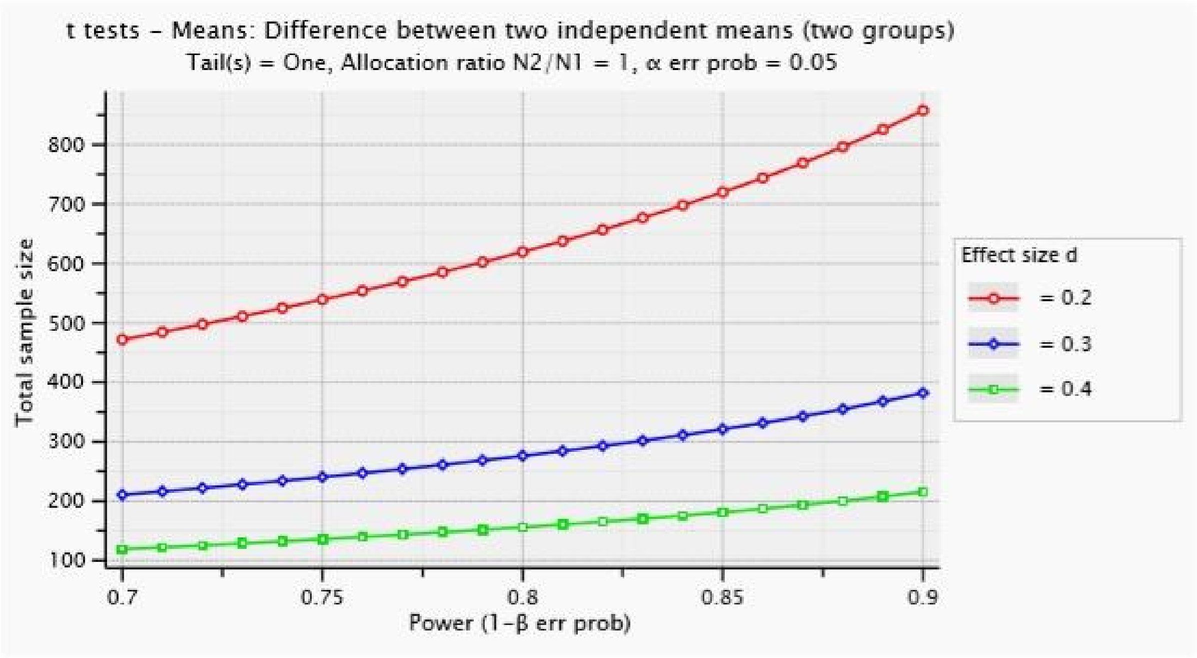
Appendix 8: Comfort gender

	(1) More Comfort b/se	(2) More Comfort b/se
Male patient	-0.028	-0.008
	(0.043)	(0.032)
Female doctor	-0.084	-0.151
	(0.112)	(0.107)
Male patient & Female doctor	0.168*	0.152**
	(0.083)	(0.071)
Constant	0.496***	-1.410
	(0.068)	(1.563)
Controls + Hospital FEs	NO	YES
R^2	-0.004	0.025
N	457	457
Note: *p<0.1, ** p<0.05, ***p<0.001		

Appendix 9: Power Calculations

According to the 2016-2017 National Survey of Drug, Alcohol and Tobacco Consumption, respondents in Mexico City who reported that they had consumed alcohol in the last month said that they consumed an average of 4.3 drinks each time they did so (SD = 3.43; Mean for females = 3.5, SD = 1.81; Mean for men = 5.3, SD = 4). One alcohol unit equals one can of beer, one glass of wine or one ounce of liquor. An effect size of 0.28 is related to an average increase of 23% in the alcohol consumption report. That is, with 200 participants in each group (1,000 in total), we will be able to detect an effect size of 0.28 with a statistical power of 0.8 and a significance level of 0.05. In the clinical literature, this is considered an effect of medium size (Ferguson, 2009).

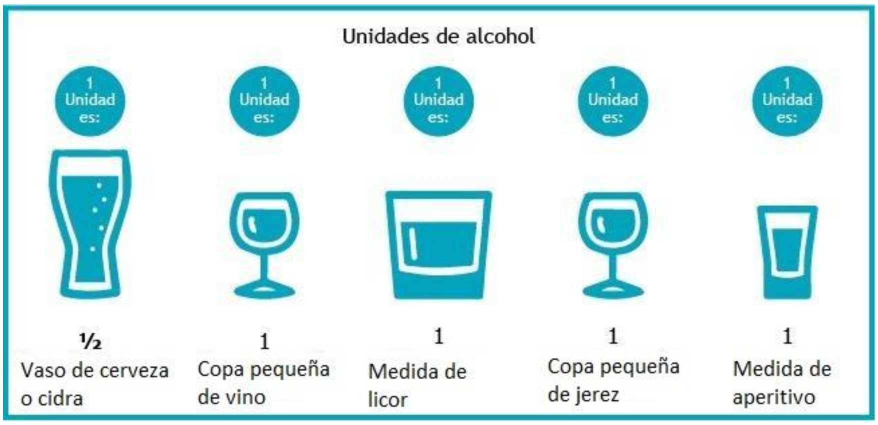
Figure 5. Power calculations



Appendix 10: Materials

We present the original material in Spanish in the OSF page of this project.

Guide for units of alcohol



Translation: *Une unit is equivalent to: ½ glass of beer lager or cider, 1 small glass of wine, 1 measure of spirit, 1 small glass of Jerez, 1 small glass of an aperitif drink.*

Format for all doctors

Name of the doctor:
Date:

1. Sex of the patient
2. Age in years
3. Does the patient receive medical care regularly (for example because of a chronic illness?)
4. Level of oxygen saturation

(To be completed according to the type of question).

1. How many alcoholic beverages do you drink during an average week? Both during weekdays and weekends.
2. How many times a week do you eat at least one portion of green vegetables?

Addition for the treatment: Importance

Before the question, please say "I need you to tell me the truth about the next questions to be able to help you better." (Show the guide for units of alcohol.)

1. How many alcoholic beverages do you drink during an average week? Both during weekdays and weekends.
2. How many times a week do you eat at least one portion of green vegetables?

Addition for the treatment: Format

Give this format to the patient!

How many alcoholic beverages do you drink during an average week? Both during weekdays and weekends? (Image of units of alcohol)
I drink ____ units of alcohol per week.

How many times a week do you eat at least one portion of green vegetables?
I eat green vegetables _____ days a week.

Addition for the treatment: Non-judgemental

Before the question, please say "I am asking you this because my friends sometimes eat too few vegetables and drink beer." (Show the guide for units of alcohol.)

1. How many alcoholic beverages do you drink during an average week? Both during weekdays and weekends.
2. How many times a week do you eat at least one portion of green vegetables?

Addition for the treatment: Memory

Before the question, please say "Take 10 seconds to remember precisely how many drinks you had in the last week." (Show the guide for units of alcohol.)

1. How many alcoholic beverages do you drink during an average week? Both during weekdays and weekends.

Wait 10 seconds and confirm the answer. Then say:

2. How many times a week do you eat at least one portion of green vegetables?

Wait 10 seconds and confirm the answer.

Addition for the treatment: Control

1. How many alcoholic beverages do you drink during an average week? Both during weekdays and weekends.
2. How many times a week do you eat at least one portion of green vegetables?

Wait 10 seconds and confirm the answer.

Informed consent

Informed consent

We would like to use these answers in connection with academic research on the interaction between patients and doctors. All your information will be confidential and will not be used for other purposes. If you agree to let us use this information for research, please answer the following and sign at the end.

1. Level of education (highest level)
2. Postal code
3. State to what extent you agree with the following statement: In this interview I felt:
   - *More comfortable than in any other medical appointment*
   - *As comfortable as in any other medical appointment*
   - Less comfortable than in any other medical appointment

Name and signature:

If you would like to know more about this research project and the way we process your information, please write to Pablo Ignacio Soto Mota: xxxxxxxxx@gmail.com

Chapter 3: Causing harm with others. An experiment on diffusion of responsibility and social norms.

Pablo Soto-Mota
Norwegian School of Economics, FAIR, Norway

Adrian Vargas-López
Institute of Environmental Planning, Leibniz University Hannover, Germany

Why does individuals in groups tend to cause more harm than isolated individuals? This paper explores the mechanisms of diffusion of responsibility and social norms. Using an online experiment with 1,801 participants, we ask two questions: Are individuals in larger groups of culprits more willing to collaborate in generating harm? Do perceived social norms affect the willingness to collaborate in generating harm? In the experiment, participants were pivotal in a group decision whether to eliminate a charity donation. They could benefit privately from favouring the elimination, though their individual decision was crucial for the group. We compared groups with different numbers of culprits and different strengths of the perceived social norm. As predicted by diffusion of responsibility, we found that a larger proportion of participants were willing to cause harm in groups with more culprits. As predicted by social norms, we found that participants were more likely to eliminate the donation when they knew that a larger proportion of individuals in their group had favoured eliminating it. Overall, this paper contributes to the literature on ethical decision-making and the factors that influence group behaviour.

Keywords: Misconduct, Group thinking, Diffusion of responsibility, Social norms, Group size

JEL: C92, D01, D23

3.1 Introduction

Sharing responsibility for unethical actions is a factor leading individuals to voluntarily join groups that cause harm (El Zein et al., 2019). For instance, members of a shooting squad may experience less guilt when shooting collectively than if they acted alone, though the outcome does not change for their victims. Similarly, teenagers in groups are less likely to help bullying victims in comparison to when they are alone (Thornberg & Jungert, 2014), and individuals may overlook moral transgressions of other group members to maintain the groups' reputation more than what they do with single individuals (Ashokkumar et al., 2019).

Although the previous examples refer to different situations and environments, they all suggest that belonging to a group provides potential justifications for collaborating in generating harm. The existing literature emphasises several underlying reasons for such an effect. For example, individuals in groups can justify their selfish actions by claiming that the harmful outcome was someone else's responsibility (Bartling & Fischbacher, 2012; McGloin & Piquero, 2009), that their actions were non-decisive (Duffy & Tavits, 2008), that someone else would have done the same (Bartling & Özdemir, 2023), that others in the group approve the behaviour (Danilov et al., 2018), or that the guilt or praise of an action is diffused among all who collaborated in harming others (Bandura et al., 1975). However, these justifications are often intertwined making it is difficult to disentangle their effects. We address this gap in the literature by conducting a controlled online experiment that cleanly identifies the causal effect of two channels: social norms and the diffusion of responsibility.

We answer two research questions. First, holding all else constant, are individuals in larger groups of potential accomplices more willing to collaborate in generating harm? Second, do perceived social norms affect the willingness to favour a group decision that negatively affect others? Following the literature, we refer to the effect of the group size of accomplices as "diffusion of responsibility". This captures the intuition that moral responsibility decreases, everything else constant, by the number of group member causing harm. Likewise, we understand social norms in this context as the perceived fraction of people doing wrong out of the population or reference group.

In our experiment, participants were pivotal in a group decision on whether to eliminate a charity donation or receive a private monetary benefit. We compare groups with a different number of potential culprit accomplices and two levels of the perceived social norms. In our experiment, all participants knew they were pivotal in the group's choice, and we informed them that their behaviour would not affect or be known by others in their group. Thus, we isolate the effects of group size and social norms from other mechanisms.

We observe that being in a group with more potential accomplices increases the likelihood that a participant accepted eliminating a charity donation for private benefit. Increasing the number of culprits from 2 to 4 caused a 5-percentage point increase in the elimination of the donation. This implies that 19% fewer individuals rejected to collaborate when they were in a smaller group, and thus had fewer culprits to diffuse their moral responsibility on, in comparison to those who were in a larger group. This result indicates that diffusion of responsibility is a mechanism of why people in groups tend to behave worse than isolated individuals. Moreover, we found evidence supporting the significance of perceived social norms on group members' decisions – participants were more likely to eliminate the charity donation when they knew that a larger proportion of individuals in their group favoured that option. Pivotal participants voted 4.5% more in favour of eliminating the donation when they knew all previous voters did that (perceived social norm 1) compared to when they knew that only half of them did the same (perceived social norm 0.5). Furthermore, in exploratory analyses, we observe that these results correlate positively with participants' expectations of others' behaviour. In other words, in groups with more accomplices, and when individuals perceived that the social norm was in favour of eliminating the donation, participants predicted that others would be more likely to eliminate the donation. Finally, our results suggest a self-serving formation of responsibility attribution. Namely, those who saved the donation felt more responsible for the group's decision than those who eliminated it.

The rest of this paper goes as follows. In the next section, we review the literature related to the effect of groups on ethical decision-making. Section 3 presents the experiment design, a conceptual framework, and our hypotheses. In addition, we explain the method for implementation and some characteristics of our sample. Section 4 shows the results. In section 5, we conclude and discuss our findings.

3.2 Related literature

Previous studies show that groups act more selfishly and strategically than isolated individuals (Bornstein et al., 2004; Charness & Sutter, 2012; Kugler et al., 2012). Evidence from laboratory experiments has shown that groups lie more (Kocher et al., 2018; Muehlheusser et al., 2015) and distribute less money in dictator games than individuals alone (Luhan et al., 2009). The difference in behaviour between individuals and groups is often attributed to the "bystander effect". This is the observation that in that the presence of multiple witnesses – bystanders – the likelihood that an individual will help is reduced (Darley & Latane, 1968; Latané & Nida, 1981; Panchanathan et al., 2013). In other words, when others are present, an individual may be less willing to act morally if doing so is costly.

Fischer et al. (2011) reviewed the literature up to that date and observed the "bystander effect" to be a consistent finding. However, it is moderated by variables like danger, the presence of the perpetrator, the identity of other witnesses and the cost of helping. Nonetheless, it is debated what are the mechanisms behind it. Why would being part of a group or the presence of additional witnesses make individuals act more immorally? A common rationale to explain the bystander effect is that individuals may avoid helping someone in need –acting morally– when they can argue that other witnesses could step in to provide assistance. In other words, individuals free-ride the morality of others when the good action is a public good and helping is costly (Campos-Mercade, 2021). This is also referred to as "replacement logic".

Another but related explanation of the bystander effect, mainly emphasized in the economics literature, is that the cost of behaving immorally may decrease when it is possible to justify that the behaviour was not decisive or pivotal for a group decision (Bénabou et al., 2018). In simple terms, acting immorally – i.e. not helping in an emergency– may be easier if one thinks the harmful outcome will not happen in any case. Conversely, individuals are more motivated to act morally when they think their actions are critical for a good outcome to happen. For instance, Duffy and Tavits (2008) found that in elections, individuals are more willing to bear the costs of voting when they think their vote is pivotal. Moreover, selfish individuals may self-select into groups where their selfish actions are less crucial and, therefore, less morally costly (Brütt et al., 2020).

An alternative mechanism behind the bystander effect, and thus the difference between group and individual behaviour, is the intuition of "diffusion of responsibility". Individuals may feel less responsible for actions within a group because they perceive that the moral responsibility for the harm caused by the individual is divided among its members (Bandura, 2016; Guerin, 2011). Since the causal attribution between an individual decision and the harm produced is blurred in a group, its members can justify acting wrongly (Engl, 2022). For example, individuals may feel that their decision not to help someone in need – acting immorally – contributes but does not cause the harm, so they find it easier to avoid the costs of helping. Feng et al. (2016) identified the neurological bases for the diffusion of responsibility on individual engagement in pro-social behaviour. Likewise, using a vignette experiment, McGloin and Thomas (2016) observed that people expect to receive lower formal sanctions when they break the rules in larger groups. Recently, Behnk et al. (2022) showed suggestive evidence from a laboratory experiment that immoral actions gain normative acceptance when made with others.

Finally, social norms within groups may also explain differences between group and individual moral behaviour. There is plenty of evidence that perceived social norms influence people's decision-making (Bicchieri et al., 2022; Soraperra et al., 2017). This implies that the likelihood that individuals will act immorally – i.e., not providing help – will increase if they observe a larger fraction of others doing the same. Hence, as long as no one in a group pays the cost of acting morally, the rest will perceive the social norm to be in favor of acting immorally. In this spirit, Campos-Mercade (2022) observed in a laboratory setting that individuals behave more selfishly when they know they are part of a group with mostly selfish people, potentially explaining why groups, under certain conditions, behave worse than isolated individuals.

Despite the broad literature and different proposed explanations as to why groups behave more unethically than individuals alone, no previous study has identified the effect of the number of accomplices and social norms on the willingness to collaborate in causing harm – independently of other mechanisms. The reason is that these explanations interact with the idea behind pivotality aversion and replacement logic, making it hard to disentangle specific effects. Intuitively, increasing the group size decreases the probability of being pivotal and increases the chance that someone else will act ethically. In this study, we fill that gap in the literature by ensuring that all participants are pivotal and irreplaceable. Hence, neither the replacement logic nor pivotality aversion can explain their behaviour. This allows a clean identification of the importance of diffusion of responsibility and social norms. Finally, we show suggestive evidence that individuals within groups interpret the moral attribution of their actions self-servingly depending on its consequences.

3.3 Experiment design

In this section, we first describe our experiment. Then, we explain the method for implementing it. Finally, we present a theoretical framework to motivate our hypothesis and analysis. The supplementary material presents the exact instructions of the study (Appendix 12).

3.3.1 Description of the experiment

We conducted an experiment where participants had to choose between personal gain and moral responsibility. To answer our research questions, our experiment design isolates the effect of the size of the guilty group and the effect of social norms from other factors like uncertainty, efficiency concerns, and strategic behaviour.

After entering the experiment, we informed participants that they were part of a group with *S* members. We told them that the group was making a joint decision by voting on whether to do a harmful action: eliminating a charity donation. Each group member could vote *Yes* or *No* to the elimination of the donation. We explained that the group's decision rule was that the donation would be eliminated if *T* or more members voted *Yes*. Importantly, participants got a bonus for voting *Yes* independently of others' votes and the groups' final decision. Participants in this situation face a moral trade-off. On the one hand, selfish participants would always vote *Yes* since it maximizes their earnings. On the other, voting *No would eliminate the individual bonus*, but could lead to the charity getting a donation[5].

In the experiment, we vary the number of members in the group, *S,* and the threshold to eliminate the donation, *T*. We are interested in the mechanism of diffusion of responsibility, that is, that the moral cost of harming the charity decreases in larger groups of culprits. In the following, when we refer to the group size or the number of culprits if the donation is eliminated, we talk about the threshold *T*. Similarly, when we refer to perceived social norms, or, shortly, social norms, we mean something close to what the literature defines as descriptive norms (Bicchieri et al., 2022). In other words, the participants' belief on the proportion of other group members who would vote *Yes*.

[5] For example, a committee deciding on contracting a firm for a project. Some committee members would be tempted to receive bribes for their votes even when that harms their beneficiaries with higher costs. The scandals of corruption during the biding process to host the FIFA World Cup finals makes evident the realism of this situation. Several committee members took bribes for their votes to favour a specific country even when that harmed other competitors and the organization (Becker, 2013; Szymanski, 2016).

To isolate other mechanisms at play, we fix participants' expectations about the behaviour of others. All participants were sure to be decisive – pivotal – in the group decision. We do this by informing them that S-1 voted before them and that exactly T-1 voted Yes to eliminate the donation. So, we simultaneously manipulate the number of potential culprits T and the perceived social norms $\frac{T-1}{S-1}$. This manipulation allows us to cleanly analyse the causal effects of the group size and social norms on the willingness to collaborate in eliminating the donation. Intuitively, by eliminating uncertainty around others' actions, participants cannot justify voting Yes arguing that others would vote to eliminate the donation anyway.

Our experiment includes some elements that rule out other mechanisms that explain how the group size affects the willingness to become an accomplice in harmful actions. First, participants' earnings depend only on their own choices. Consequently, participants are not affected by efficiency concerns or their decision's effect on the earnings of others. Second, participants remain anonymous and do not communicate before making the decision. Third, the charity remained unnamed to avoid interactions with participants' preferences among organizations.

Finally, we informed all participants that those who voted before them would never know their vote or the group's final decision. Consequently, they would not be affected by guilt aversion or the willingness to receive side payments from other participants (Battigalli & Dufwenberg, 2007; Bellemare et al., 2019).

3.3.2 Experimental conditions and method for implementation

There were four experimental conditions in our study. Each condition described a pivotal situation in a group with *S* members, and *T* as the threshold or culprit group size. In Table 1, we show how these arrangements correspond to two levels of perceived social norms: 0.5 and 1. The former indicates that half of those who voted previously supported eliminating the donation, while the latter indicates unanimous support for elimination. With respect to the culprit group size, each level of social norm had one small group (*S:3 & T:2* for social norm "Low"0.5; *S:2 & T:2* for social norm "High": 1) and one large group (*S:11 & T:6* for social norm "Low": 0.5; *S:4 & T:4* for social norm "High":1).

We conducted the experiment on Prolific, an online platform for social science studies (Palan & Schitter, 2018). Participants received a compensation of 1.3 USD for their time, along with a guaranteed bonus of 0.5 USD for completing the experiment that was independent of their choices. They could get an additional bonus of 0.5 USD for voting *Yes*. The donation that each group could eliminate was of that same amount. Notice that since participants' payments depend on their individual decisions alone and not on the voting result, there are no efficiency concerns. Likewise, since all groups could donate 0.5 USD to the charity and all participants are sure to be pivotal, all decision-makers face the same choice: whether to take 0.5 USD for themselves at the expense of the charity.

Chapter 3. Table 1. Experimental conditions

	Culprit group size (T)	
Perceived social norm	***Small***	***Large***
Low = 50% voted *Yes*	*S:3 & T:2*	*S:11 & T:6*
High = 100% voted *Yes*	*S:2 & T:2*	*S:4 & T:4*

Note: Shows the four "pivotal conditions" tested in the experiment. Number of members in the grouo: *S*. Threshold or culprit group-size: *T*. We define the perceived social norm as $\frac{T-1}{S-1}$.

In this context, we compare the group with a small number of potential culprits with the one with a large number of potential culprits at each level of the social norm. Our dependent variable is the proportion of pivotal decision-makers who voted *Yes* in each condition. Likewise, to test the effect of the social norm, we make two comparisons. First, we contrast the answers of participants in groups with a unanimous social norm in favour of voting *Yes* (High social norm) and those in groups with an ambiguous social norm in favour of voting *Yes* (Low social norm). Second, we compare the two groups with the same number of culprits but different levels of the social norm (*S:3 & T:2* and *S:2 & T:2*).

As previously mentioned, we informed all participants that they were pivotal in their group. This information did not involve deception. To implement it, we collected real and incentivised answers for the 85 possible situations in which the groups could be. For example, a group situation in the condition *S:4 & T:4* is one where two voted before, and one voted *No* while the other voted *Yes*. Another is one where no one has yet voted.

We randomly allocated participants into one of five group situations: one pivotal and four non-pivotal. A pivotal situation is one where participants know they are the last person to vote and that their decision is critical for eliminating the donation or not. We present explanatory images of the pivotal situations in Appendix 1. Participants answered the pivotal situation first and later the non-pivotal ones. The reason is that only the pivotal situation is relevant to our research question. We informed all participants that we would implement one of their decisions. With the information on the non-pivotal choices, we guaranteed that we could build group situations, based on incentivised answers, for the pivotal situations analysed in the experiment. Our experimental design was reviewed and approved by the IRB at NHH. In Appendix 2, we present a flowchart of the experiment's procedure.

3.3.3 Theoretical framework

This section presents a theoretical framework about individual decision making in the collaboration of group-generated harm. Our objective is to make clear the ideas that are behind our hypotheses and the key idea of only focusing on the pivotal voters.

In our experiment, participants are in a group with S members and decide by voting whether to eliminate a donation. Each individual, i, may either vote *Yes*, $a_i = 1$, or vote *No*, $a_i = 0$. If T or more participants vote *Yes*, $\sum_{j=1}^{S} a_j \geq T$, the donation is eliminated. Participants receive B as a bonus if they vote *Yes*. Their bonus is independent of the vote of the rest and of the group decision. They vote simultaneously, though they have beliefs about the behaviour of others. We can write *i*'s subjective probability, or belief, that the donation will be eliminated as $P_i(\sum_{j=1}^{S} a_j \geq T)$. Finally, individuals value the donation and would take a moral cost, C_i, in case it is eliminated. Therefore, we can write the individual's utility function after voting *Yes* as:

$$U(a_i = 1) = B - P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 1\right) C_i$$

Likewise, after voting *No* their utility would be:

$$U(a_i = 0) = -P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 0\right) C_i$$

Consequently, the individual's condition for voting *Yes* reduces to:

$$B \geq \left[P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 1\right) - P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 0\right)\right] C_i$$

Notice that the term inside the brackets is a measure of the relevance of *i*'s vote on the groups' decision. It is maximal, taking the value of 1, only when the individual is pivotal.[6] In fact, it is equal to the probability of being pivotal, $P_i\left(\sum_{j \neq i}^{S} a_j = T - 1\right)$. Consequently, we can rewrite the individuals' condition to vote *Yes* as:[7]

[6] For a pivotal individual, $P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 1\right)$=1 and $P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 0\right)$=0 simultaneously.

[7] To observe why this is the case, notice that $P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 1\right) = P_i\left(\sum_{j \neq i}^{S} a_j \geq T - 1\right)$ and that $P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 0\right) = P_i\left(\sum_{j \neq i}^{S} a_j \geq T\right)$. Therefore, $P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 1\right) -$

$$B \geq P_i\left(\sum_{j\neq i}^{S} a_j = T - 1\right)C_i$$

We are interested in the idea of diffusion of responsibility. In other words, do people find it easier to collaborate in eliminating the donation when they are in groups where more people – *T* – do wrong. Hence, we seek to explore how the number of culprits, *T*, affect the moral cost C_i. Likewise, we are interested in studying how perceived social norms, understood as the expected proportion of participants who will vote *Yes,* affect the moral cost of doing wrong.

This theoretical framework illustrates that the culprit group size and social norms also can affect the subjective probability of being pivotal. For example, by increasing the threshold *T*, holding everything else constant, the probability of being pivotal reduces.[8] Likewise, if the perceived social norm changes in favour of voting *Yes*, the probability that exactly *T-1* vote *Yes* will increase as well. Consequently, to observe how the size of the culprit group and social norms affect C_i, we need to disentangle these effects from the probability of being pivotal. In our experiment, we do this by fixing the participant's beliefs of the probability of being pivotal to 1. If the participant is sure to be pivotal, $P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 1\right) = 1$, their condition for voting *Yes* is simply $B \geq C_i$. Therefore, we can cleanly analyse the effects of *T* and the perceived social norms, $\frac{T-1}{S-1}$, on C_i.

3.3.4 Hypotheses and empirical strategy

Based on our theoretical framework, we hypothesise the following for pivotal individuals:

Hypothesis 1 – Diffusion of responsibility: The moral cost of collaborating in eliminating the donation decreases in groups with a larger number of culprits. Consequently, we predict that, everything else constant, a larger group of culprits, *T,* will result in more participants voting *Yes*. Consequently, in our comparisons the acceptance rates will follow: *S:2 & T:2 < S:4 & T:4* and *S:3 & T:2 < S:11 & T:6.*

Hypothesis 2 – Perceived social norms matter: The moral cost of voting to eliminate the donation decreases in groups where participants perceive that a larger proportion

$P_i\left(\sum_{j=1}^{S} a_j \geq T \middle| a_i = 0\right)$ is equivalent to $P_i\left(\sum_{j\neq i}^{S} a_j \geq T - 1\right) - P_i\left(\sum_{j\neq i}^{S} a_j \geq T\right)$. Since *T* is an integer from 0 to *S*, that difference is the subjective probability that exactly $T - 1$ individuals other than *i* vote *Yes*: $P_i\left(\sum_{j\neq i}^{S} a_j = T - 1\right)$.

[8] Falk et al. (2020) show this formally. In addition, they found in a laboratory experiment that the willingness to generate harm (killing mice) for money increases when the probability of being pivotal is lower. Our experimental design draws on theirs to analyse how the size of the group of culprits affects the moral cost of collaborating in causing harm.

of individuals favor the generation of harm. Therefore, in our experiment, for equal culprit group sizes, the group where the pivotal individual knows that a higher proportion of other voted *Yes* will have a higher proportion of acceptance. In particular, we predict that the acceptance rates will follow this order: *S:3 & T:2* < *S:2 & T:2*. Furthermore, we predict that individuals exposed to unanimous social norms (*S:2 & T:2* and *S:4 & T:4*) will have a higher rate of *Yes* votes than individuals exposed to ambiguous social norms (*S:3 & T:2* and *S:11 & T:6*).

In an exploratory analysis, we examine the effects of social norms and the size of the group of culprits on the prediction of the percentage of participants who will vote *Yes*. That is, we test whether T and $\frac{T-1}{S-1}$ influence the prediction of the social acceptability of collaborating in the generation of harm. To do this, we use a question that we asked participants at the end of the experiment: "*If we offer this exact deal to 100 people, how many do you think would vote Yes?*" We expect that the effect will be in the same direction as our main hypotheses.

Additionally, we test participants' feelings of guilt and praise after eliminating or saving the donation across the experimental conditions. We use two questions. After the experiment, one question asks how responsible they feel about the group's decision. Another asks if they feel more or less responsible than those who voted like them in their group.

It is important to notice that these exploratory comparisons are not necessarily causal. The reason is that these questions were asked at the end of the study, after participants had answered the pivotal and not pivotal situations. Participants might answer in a self-serving way, trying to reduce the importance of their actions. It is also possible that their behaviour had changed after answering the other four similar but non-pivotal cases. The reason is that our experimental design is based on giving participants information about others' choices that could be less credible after answering other scenarios.

3.3.5 Participants

We conducted our experiment in May 2022. The experiment involved 1801 participants from the US and, on average, lasted for 8 minutes. After the experiment, participants also filled out a basic demographic questionnaire. We present their demographic characteristics in Table 2 for each experimental condition. We notice that 49 percent are female, 66 percent have completed a Bachelor's degree, and that they are, on average, 38 years old. Though the orthogonality test signals some imbalance across the groups, we consider that the differences are too small to be meaningful – i.e., one year of age. Nevertheless, in the appendix we control for these covariates in our analysis obtaining the same results.

Chapter 3. Table 2. Participants' characteristics per experimental group

	Experimental condition				**All**	**Orth. test (p-val.)**
	S:2/T:2	**S:3/T:2**	**S:4/T:4**	**S:11/T:6**		
N	453	452	452	444	1,801	
Female	0.483	0.491	0.485	0.502	0.490	0.572
% female	(0.500)	(0.500)	(0.500)	(0.501)	(0.5)	
Education	0.678	0.662	0.628	0.671	0.66	0.018
%>Bachelors	(0.468)	(0.474)	(0.484)	(0.47)	(0.47)	
Age	37.638	38.352	38.728	37.788	38.128	0.002
In years	(13.17)	(13.12)	(12.85)	(13.39)	(13.1)	
Political opinion	3.530	3.524	3.617	3.718	3.597	0.048
0: left, 10: right	(2.646)	(2.564)	(2.707)	(2.718)	(2.66)	

Note: SD in parentheses. No-female includes "Other" (N=24) and "Male" (N=894). The number of participants per experimental condition varied naturally by the randomisation during the implementation of the study.

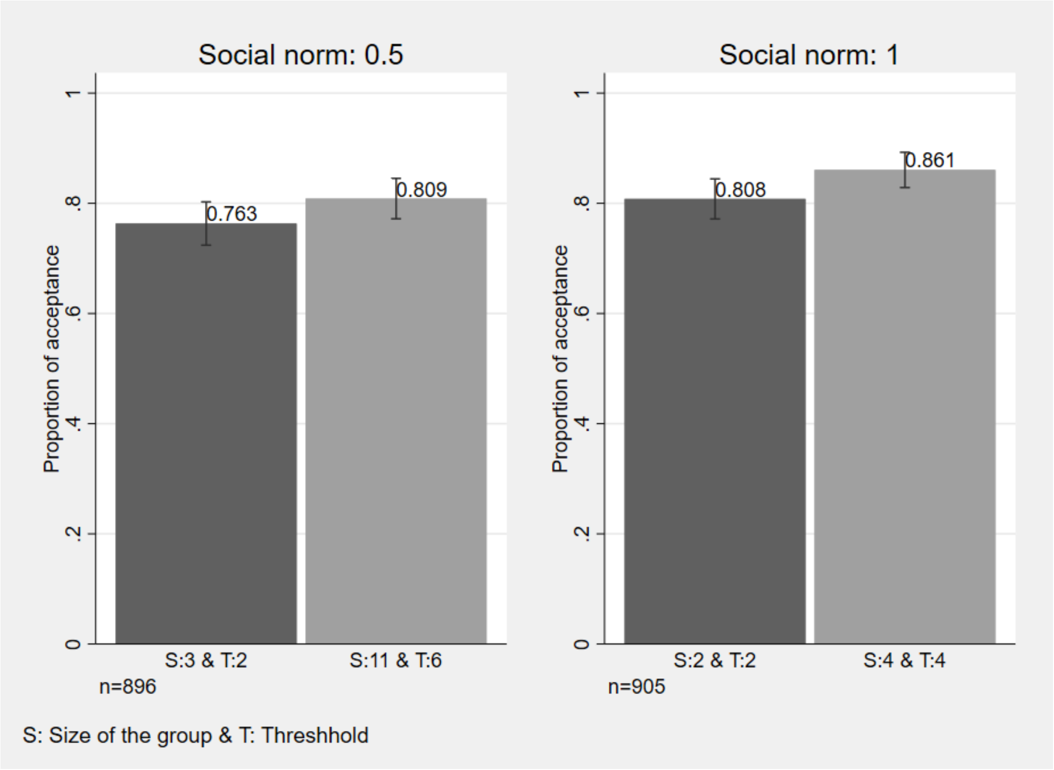
3.4 Results

3.4.1 Main results

Result 1. Diffusion of responsibility: The more culprits the higher is the likelihood that a pivotal individual collaborates in generating harm.

In Figure 2, we present the acceptance rates – i.e voting *Yes* – across conditions. The left panel displays the groups with a low level of social norm of 0.5, which implies that the pivotal voter received information that half of the previous voters had voted *Yes.* The right panel presents the groups with a unanimous social norm of 1, which implies that the pivotal voter received information that all of the previous voters had voted *Yes*. We present proportion tests of our main hypotheses in Appendix 4.

Chapter 3. Figure 1. Main results: proportions of "Yes" votes by experimental condition



Note: The error bars show the +/-95% confidence intervals.

Consistent with our predictions, a larger group of culprits caused a higher willingness to eliminate the donation. This holds irrespective of the social norm. In the case with a low level of social norm, we found a statistically significant difference in the proportion of *Yes* votes between condition *S:11 & T:6* and condition *S:3 & T:2* (*Difference* = 4.5%, *z-value* =1.652, *one-sided p-value* = 0.049). Specifically, in the *S:11 & T:6* group there was 19.4% fewer participants that saved the donation in comparison with participants in the *S:3 & T:2* group. Likewise, in the case with a high social norm, our analysis revealed a significant difference between condition *S:4 & T:4* and condition *S:2 & T:2* (*Difference* = 5.3%, *z-value* = 2.131, *one-sided p-value* = 0.017). Again, more individuals favoured the elimination of the donation in the larger group of culprits. In Appendix 6, we consider socioeconomic covariates to further test this result. Our results are the same after controlling for these covariates.

Result 2. Perceived social norms: The higher fraction of culprits there are in the population the higher is the likelihood that a pivotal individual collaborates in generating harm.

To test the second hypothesis, we first compare the two groups with social norm 0.5 with the two groups with social norm 1. By comparing participants randomly allocated into groups with different social norms, we observe that pivotal voters who see a higher proportion of previous voters voting *Yes* are more likely to vote *Yes* themselves (*Difference* = 4.8%, *z-value* =2.626, *one-sided p-value* = 0.004). Although this result supports our hypothesis, it should be noted that these groups differ slightly in the size of the group of culprits. In social norm 0.5, the threshold is either of 2 or 6 individuals (4 on average), while in social norm 1, the threshold is either 2 or 4 (3 on average).

Our second approach avoids this problem. We compare the conditions *S:3 & T:2* and *S:2 & T:2*. In these conditions, the threshold is the same, but the pivotal individual receives a different information about the social norm. Our results provide suggestive evidence that a higher social norm leads to a higher proportion of votes in favour of eliminating the donation. However, we narrowly fail to reject the null hypothesis of no difference at 95% (*Difference* = 4.5%, *z-value* =1.637, *one-sided p-value* = 0.051). However, this effect is neither qualitatively nor quantitatively different than with the first approach, suggesting that the near rejection of the null hypothesis reflects the lower statistical power in this test. As with the previous results, we observe similar results after including socioeconomic controls in Appendix 7.

Finally, we notice that both variables analysed in this experiment, culprit group size and social norms, had similar effects. Being part of a larger group or being exposed to the unanimous social norm group increased the likelihood of voting in favour of eliminating the donation by around 5 percentage points when compared to the smaller group of culprits and the ambiguous or low social norm, respectively. Moreover, our data suggests that these effects are independent from each other since when combined they are additive. Specifically, the difference between the rate of acceptance in the small group with a low social norm (S:3 & T:2) and the large group in the unanimous norm (S:4 & T:4) is of 9.8 percentage points. In other words, moving from a situation with only one other participant to diffuse moral responsibility and were the social norm was ambiguous, to a situation with three other participants to diffuse the moral responsibility and a clear social norm in favour of doing harm, reduced the proportion of participants voting to save the donation by 41.3%.

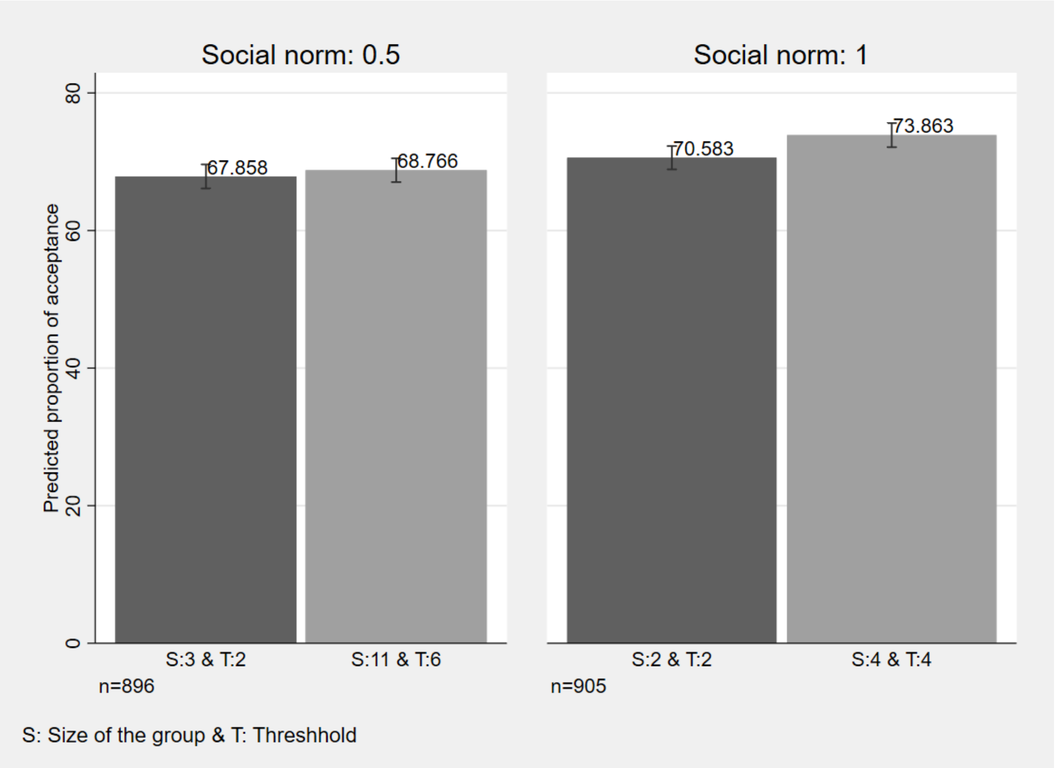
3.4.2 Exploratory results

In this section, we present the results of an exploratory analysis using questions incorporated after the experiment. With these exploratory questions we aim to make better sense of our results in two ways. First, to examine the extent to which information about social norms influenced perceptions of the acceptability of voting in favour of the elimination of the donation. Second, to explore how the individual's decision in the pivotal situation relates to feelings of guilt and praise.

Exploratory result 1. People in groups with a higher perceived social norm also predicted more people voting in favour of eliminating the donation.

Figure 3 displays the predicted percentage of participants who would vote *Yes* across conditions. We present the t-tests related to this analysis in Appendix 5. We also include a regression analysis with socioeconomic controls in Appendix 8 to obtain similar results.

Chapter 3. Figure 2. Exploratory results: percentage of predicted "Yes" votes by experimental condition



Note: The error bars show the 95% confidence intervals. Answers to the question "*If we offer this exact deal to 100 people, how many do you think would vote Yes*?"

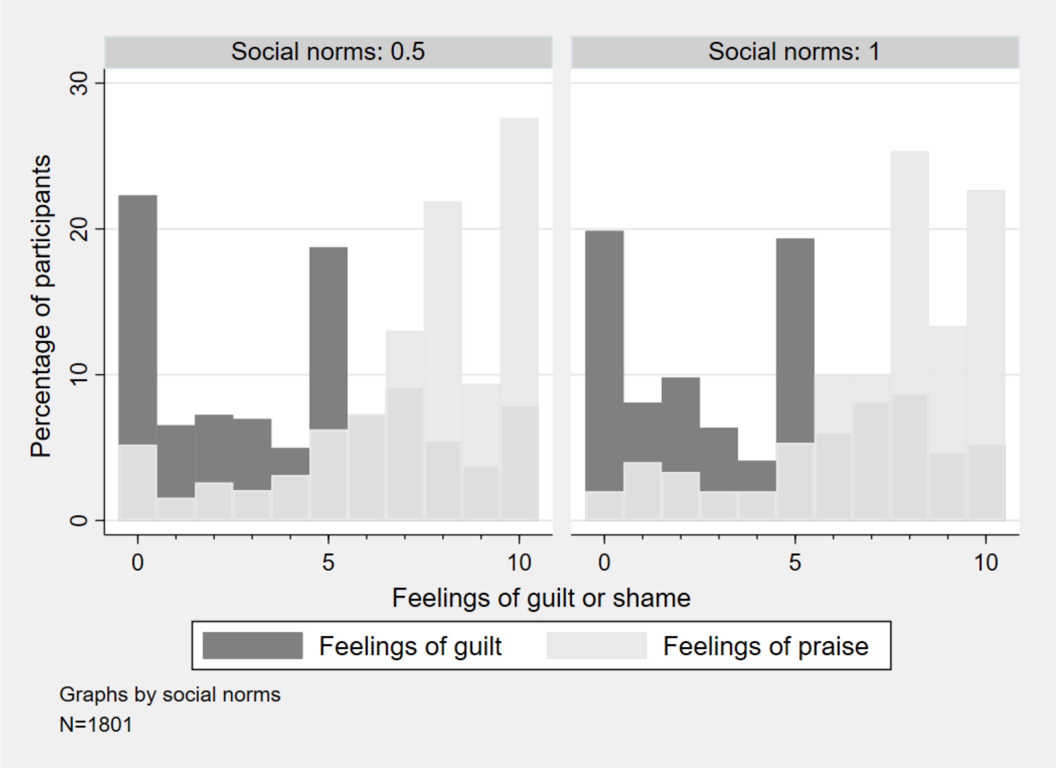
First, in all groups, participants underestimated the proportion of participants who would vote in favour of eliminating the donation, meaning that they believed others to be *less* willing to vote “Yes” than they really are. Second, we notice that, as expected, participants predicted that more of the other participants would vote *Yes* in groups with a stronger social norm in that direction. This aligns with the typical definition of social norms, which are expectations of what others would consider acceptable or normal to do (Bicchieri et al., 2022). In particular, the prediction was higher for those in a group with perceived social norm of 1 than for those in a groups with perceived social norm of 0.5 (*Difference*= 3.193, *t*= 4.43, *one-sided p-value* < 0.001), and this results also holds when we compare the two groups with the exact same threshold (*S:3 & T:2* vs *S:2 & T:2: Difference*= 2.725, *t*= 2.196, *one sided p-value*= 0.014). This suggests that our Result 2 in the previous subsection, in line with our theoretical framework, is explained by the change in the perceived acceptability of the harmful action.

Exploratory result 2. Individuals attribute moral responsibility self-servingly

Figure 4 shows histograms of responses to the question: "*From 1 to 10, how morally responsible do you feel for eliminating (saving) the donation? (0: not at all, 10: totally),* split by social norm treatment." Participants who voted *Yes* (or *No*) in the pivotal question answered the version with eliminating (or making) the donation. We interpret feelings about making the donation as praise and feelings about eliminating the donation as guilt.

We observe that those who voted in favour of eliminating the donation felt less morally responsible for the group decision than those who voted against it (*Difference*= 3.1289, *t*= 16.804, *p-value* < 0.001). In other words, people feel more personal responsibility for saving the donation than for eliminating it. In Appendix 10 we show further evidence of these self-serving perceptions of moral responsibility. Participants who voted against the elimination of the donation, and thus, saved it, felt that they were more responsible for the result than others who voted like them. However, those who voted *Yes*, and thereby eliminated the donation, reported feeling as responsible as the others. However, feelings of guilt (praise) were unrelated to the level of social norms, as a Kolmogorov-Smirnov test showed that the two levels of social norms had the same distribution of the feelings of praise (*Difference*= 0.049, *p-value*=0.986) and guilt ((*Difference=0.*026, *p-value*=0.961). In Appendix 11 we present similar results for the relative feelings of guilt (praise) and for the effect of the group size.

Chapter 3. Figure 3. Feelings of guilt and praise



3.5 Conclusions and discussion

In this study, we analysed the impact of group size and social norms on the willingness to collaborate in causing harm. Participants who were pivotal in their group had to decide whether to eliminate a donation to charity for personal gain. The study found that individuals in groups with more potential culprits were more likely to eliminate the donation – consistent with diffusion of responsibility shaping the decision. Participants were also more likely to eliminate the donation in groups with a larger fraction of culprits – consistent with perceived social norms shaping the decision. We also observed suggestive evidence of a self-serving formation of moral responsibility attribution: participants who saved the donation felt more responsible for the group's decision than those who eliminated it.

Our results contribute to the literature explaining the differences between individual and group moral behaviour. In particular, we add to the discussion on the mechanisms behind the so-called "bystander effect". Building on a simple theoretical framework and a novel experimental design we present clean evidence in favour of diffusion of responsibility and social norms as mechanisms for the bystander effect and, thus, the observation that groups behave more egoistically than isolated individuals. We also want to point out that both the theoretical framework and the experimental design could be adapted to ask further questions about moral behaviour in groups.

Some questions about moral decision-making in group situations remain open. First, as we explain in section 3, while the experiment only focused on pivotal situations, participants also answered questions in all other possible group configurations. While helpful in implementing the experiment, these answers may be less useful for analysing individual behaviour. Further research could explore other experimental designs that allow using such answers to understand how pivotality, the size of the group and social norms affect individual decisions in such situations. Second, our experiment was framed "negatively" in deciding whether to eliminate a charity donation. However, it is possible that results are not symmetrical when the decision is about generating benefits instead of harm. Also, our study tested a situation where the damage was hurting a charity, but other real-life situations often involve hurting specific individuals, such as bullying or looting. Previous evidence suggests that individuals find it easier to damage a charity in contrast to specific individuals due to a greater psychological distance (Maggian, 2019). However, this asymmetry has not been studied in a group setting. It is unclear whether we would observe larger or smaller treatment effects if the damage was against concrete people instead of a charity.

Despite these open questions, this article is relevant for the discussion on how to promote ethical behaviours in situations where costly individual actions are necessary for obtaining a common good. For example, corrupt bureaucrats may find it easier to negatively affect others if they can justify their actions as part of a network or organisation rather than as pure individual choices. Similarly, individuals may avoid simple pro-environmental behaviours if their moral responsibility toward the environment is diffused on corporations or other members of society. Also, citizens may find it harder to find the motivation to participate politically if they feel that doing so is not critical for results. Our results suggest that highlighting individual responsibility, underscoring its causal link with group choices, and making salient pro-social norms may effectively promote moral behaviour.

Acknowledgements

We thank *Centre for Ethics and Economics* for funding this project. Also, we are grateful with Bertil Tungodden, Akshay Moorty, Daniel Parra, and Ernesto Gavassa-Perez for their comments on our design and early results. We also thank assistants at the first NHH & Max Plank Institute Workshop, The Choice Lab Coffee Meeting (NHH), and the *Seminario de Filosofía de la Economía* (UNAM) for their useful feedback on our preliminary results. Finally, we are equally grateful with Mathias Ekström, and Arne Nasgowitz for comments to earlier versions of this draft.

CRediT authors' statement

Both authors contributed equally on the conceptualisation, methodology, software, investigation, and funding acquisition for this project. Pablo Soto-Mota was in charge of the data curation, the formal analysis, visualisation and the writing of the original draft. Adrian Vargas-Lopez reviewed and edited the text. Both authors approved this draft.

References

Ashokkumar, A., Galaif, M., & Swann Jr, W. B. (2019). Tribalism can corrupt: Why people denounce or protect immoral group members. *Journal of Experimental Social Psychology*, 85, 103874. https://doi.org/10.1016/j.jesp.2019.103874

Bartling, B., & Özdemir, Y. (2023). The limits to moral erosion in markets: Social norms and the replacement excuse. *Games and Economic Behavior*, 138, 143-160. https://doi.org/10.1016/j.geb.2022.12.001

Bartling, B., & Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, 79(1), 67-87. https://doi.org/10.1093/restud/rdr023

Bandura, A. (2016). *Moral disengagement: How people do harm and live with themselves.* Worth publishers.

Bandura, A., Underwood, B., & Fromson, M. E. (1975). Disinhibition of aggression through diffusion of responsibility and dehumanization of victims. *Journal of research in personality*, 9(4), 253-269. https://doi.org/10.1016/0092-6566(75)90001-X

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2), 170-176. DOI: 10.1257/aer.97.2.170

Bellemare, C., Sebald, A., & Suetens, S. (2019). Guilt aversion in economics and psychology. *Journal of Economic Psychology*, 73, 52-59. https://doi.org/10.1016/j.joep.2019.05.002

Bénabou, R., Falk, A., & Tirole, J. (2018). *Narratives, imperatives, and moral reasoning* (No. w24798). National Bureau of Economic Research. Available at NBER: https://www.nber.org/papers/w24798

Behnk, S., Hao, L., & Reuben, E. (2022). Shifting normative beliefs: On why groups behave more antisocially than individuals. *European Economic Review*, *145*, 104116. https://doi.org/10.1016/j.euroecorev.2022.104116

Becker, R. J. (2013). World Cup 2026 now accepting bribes: A fundamental transformation of FIFA's World Cup bid process. *The International sports law journal*, 13, 132-147. https://doi.org/10.1007/s40318-013-0011-7

Bicchieri, C., Dimant, E., Gelfand, M., & Sonderegger, S. (2022). Social norms and behavior change: The interdisciplinary research frontier. *Journal of Economic Behavior & Organization.* https://doi.org/10.1016/j.jebo.2022.11.007

Bornstein, G., Kugler, T., & Ziegelmeyer, A. (2004). Individual and group decisions in the centipede game: Are groups more "rational" players?. *Journal of Experimental Social Psychology*, *40*(5), 599-605. https://doi.org/10.1016/j.jesp.2003.11.003

Brütt, K., Schram, A., & Sonnemans, J. (2020). Endogenous group formation and responsibility diffusion: An experimental study. *Games and Economic Behavior*, *121*, 1-31. https://doi.org/10.1016/j.geb.2020.02.003

Campos-Mercade, P. (2022). When are groups less moral than individuals?. *Games and Economic Behavior*, *134*, 20-36. https://doi.org/10.1016/j.geb.2022.03.009

Campos-Mercade, P. (2021). The volunteer's dilemma explains the bystander effect. Journal of Economic Behavior & Organization, 186, 646-661. https://doi.org/10.1016/j.jebo.2020.11.012

Charness, G., & Sutter, M. (2012). Groups make better self-interested decisions. *Journal of Economic Perspectives*, *26*(3), 157-176. DOI: 10.1257/jep.26.3.157

Danilov, A., Khalmetski, K., & Sliwka, D. (2018). Norms and guilt. *CESifo Working Paper Series No. 6999*, Available at http://dx.doi.org/10.2139/ssrn.3198152

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4, Pt.1), 377–383. https://doi.org/10.1037/h0025589

Duffy, J., & Tavits, M. (2008). Beliefs and voting decisions: A test of the pivotal voter model. *American Journal of Political Science*, *52*(3), 603-18. https://doi.org/10.1111/j.1540-5907.2008.00332.x

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature human behaviour*, 3(6), 554-559. https://doi.org/10.1038/s41562-019-0596-4

Engl, F. (2022). *A theory of causal responsibility attribution.* Available at SSRN: https://ssrn.com/abstract=2932769

Falk, A., Neuber, T., & Szech, N. (2020). Diffusion of being pivotal and immoral outcomes. *The Review of Economic Studies*, *87*(5), 2205-2229. https://doi.org/10.1093/restud/rdz064

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y. J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, 37(2), 663-677. https://doi.org/10.1002/hbm.23057

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*(4), 517–537. https://doi.org/10.1037/a0023304

Guerin, B. (2011). Diffusion of responsibility. *The encyclopedia of peace psychology*. https://doi.org/10.1002/9780470672532.wbepp084

Kocher, M. G., Schudy, S., & Spantig, L. (2018). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, *64*(9), 3995-4008. https://doi.org/10.1287/mnsc.2017.2800

Kugler, T., Kausel, E. E., & Kocher, M. G. (2012). Are groups more rational than individuals? A review of interactive decision making in groups. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(4), 471-482. https://doi.org/10.1002/wcs.1184

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin, 89*(2), 308–324. https://doi.org/10.1037/0033-2909.89.2.308

Luhan, W. J., Kocher, M. G., & Sutter, M. (2009). Group polarisation in the team dictator game reconsidered. *Experimental Economics*, *12*, 26-41. https://doi.org/10.1007/s10683-007-9188-7

Maggian, V. (2019). Negative externalities of cheating: An experiment with charities. In Dishonesty in behavioral economics (pp. 183-191). Academic Press. https://doi.org/10.1016/B978-0-12-815857-9.00012-1

McGloin, J. M., & Thomas, K. J. (2016). Incentives for collective deviance: Group size and changes in perceived risk, cost, and reward. *Criminology*, 54(3), 459-486. https://doi.org/10.1111/1745-9125.12111

McGloin, J. M., & Piquero, A. R. (2009). 'I Wasn't Alone': Collective behaviour and violent delinquency. *Australian & New Zealand Journal of Criminology*, 42(3), 336-353. https://doi.org/10.1375/acri.42.3.33

Muehlheusser, G., Roider, A., & Wallmeier, N. (2015). Gender differences in honesty: Groups versus individuals. *Economics Letters*, *128*, 25-29. https://doi.org/10.1016/j.econlet.2014.12.019

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22-27. https://doi.org/10.1016/j.jbef.2017.12.004

Panchanathan, K., Frankenhuis, W. E., & Silk, J. B. (2013). The bystander effect in an N-person dictator game. *Organisational Behavior and Human Decision Processes*, *120*(2), 285-297. https://doi.org/10.1016/j.obhdp.2012.06.008

Szymanski, S. (2016). Compromise or compromised? The bidding process for the award of the Olympic Games and the FIFA World Cup. in Transparency International (eds.) *Global Corruption Report: Sport*. Available at: https://doi.org/10.4324/9781315695709

Soraperra, I., Weisel, O., Kochavi, S., Leib, M., Shalev, H., & Shalvi, S. (2017). The bad consequences of teamwork. *Economics Letters*, *160*, 12-15. https://doi.org/10.1016/j.econlet.2017.08.011

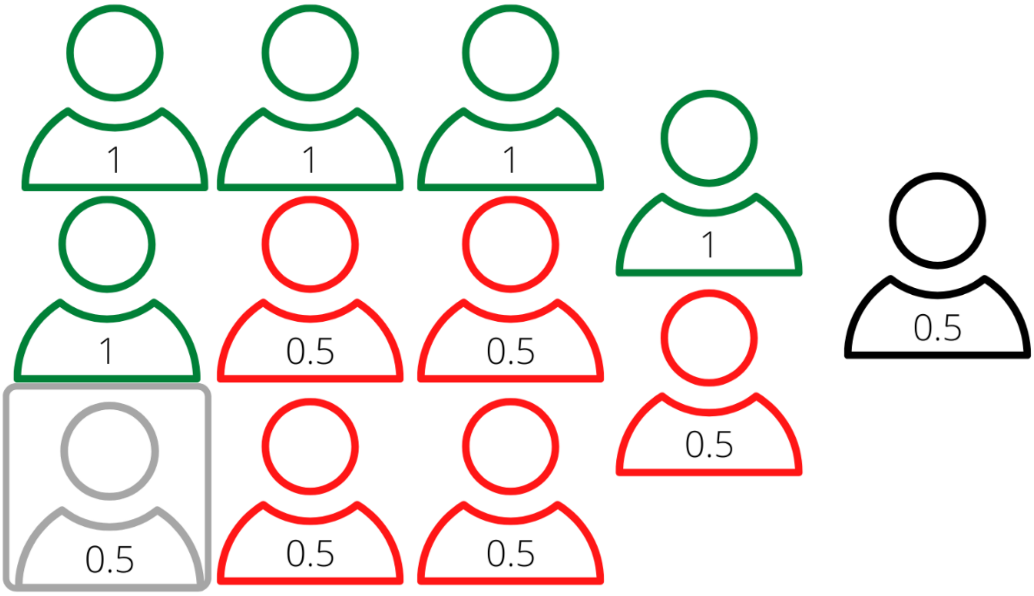
Thornberg, R., & Jungert, T. (2014). School bullying and the mechanisms of moral disengagement. *Aggressive behavior*, 40(2), 99-108. https://doi.org/10.1002/ab.21509

Appendix

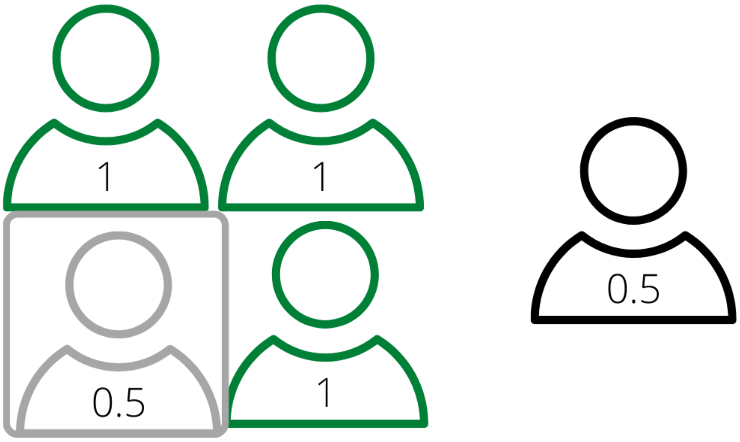
The red figure is a participant who voted No and received no additional bonus. The black figure represents the donation to the charity. In grey, there are two participants yet to vote.

A.1 Pivotal conditions

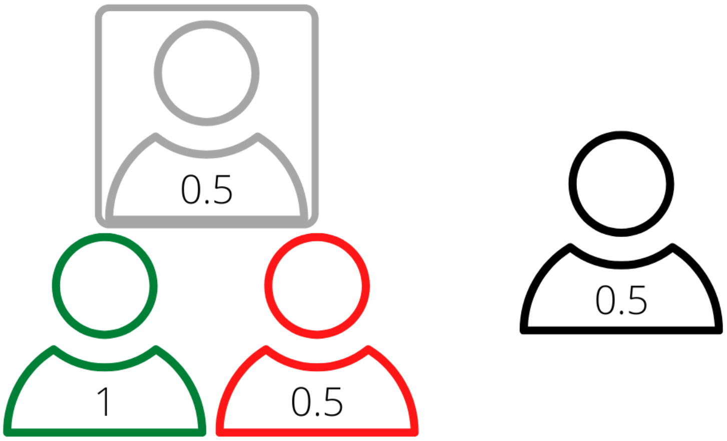
Size: 11 & Threshold: 6



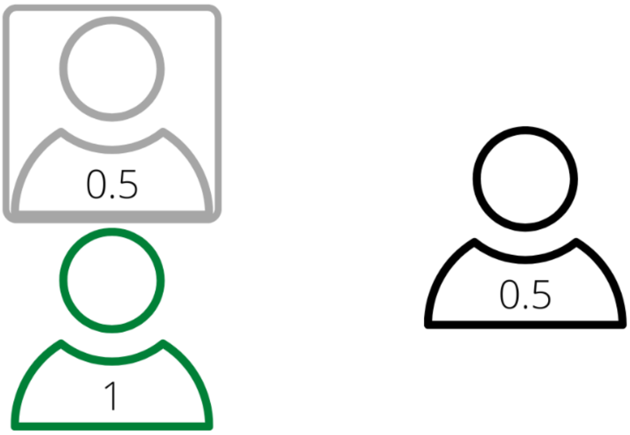
Size: 4 & Threshold: 4



Size: 3 & Threshold: 2



Size: 2 & Threshold: 2



A2. Experimental flow

A2. Experimental flow

A.3 Main Results: proportion tests on eliminating the donation

Proportion tests of voting *Yes* by hypothesis

Hypothesis		Mean 1 (SE)	Mean 2 (SE)	Diff. (SE)	z	One sided p-value
Effect of the size of the group	*S:2 & T:2 < S:4 & T:4*	0.808	0.861	-0.053	-2.131	0.017
		(0.018)	(0.016)	(0.025)		
	S:3 & T:2 < S:11 & T:6	0.763	0.808	-0.045	-1.652	0.049
		(0.02)	(0.019)	(0.027)		
Effect of Social Norms (SN)	*SN: 0.5 < SN:1*	0.786	0.834	-0.048	-2.626	0.004
		(0.014)	(0.012)	(0.018)		
	S:3 & T:2 < S:2 & T:2	0.763	0.808	-0.045	-1.637	0.051
		(0.02)	(0.018)	(0.027)		

Note: The hypotheses correspond to the ones mentioned in Section 3. *Mean 1* refers to the first condition mentioned in each hypothesis, while *Mean 2* corresponds to the second. We use one-sided p-values because our hypotheses are unidirectional.

A.4 Effects of predicted Yes votes: proportion tests

Proportion tests of predicted *Yes* votes

Hypothesis		**Mean 1 (SE)**	**Mean 2 (SE)**	**Diff. (SE)**	**t**	**One sided p-value**
Effect of the size of the group	*S:2 & T:2 < S:4 & T:4*	70.583 (0.865)	73.8628 (0.894)	-3.28 (1.244)	-2.637	0.004
	S:3 & T:2 < S:11 & T:6	67.858 (0.889)	68.766 (0.879)	-0.907 (1.25)	-0.726	0.234
Effect of Social Norms (SN)	*SN: 0.5 < SN:1*	68.308 (0.625)	72.220 (0.624)	-3.913 (0.883)	-4.430	<0.001
	S:3 & T:2 < S:2 & T:2	67.858 (0.889)	70.583 (0.865)	-2.725 (1.24)	-2.196	0.014

Note: *Mean 1* refers to the first condition mentioned in each hypothesis, while *Mean 2* corresponds to the second. We use one-sided p-values because our hypotheses are unidirectional.

A.5 Main Results about the size of the group on eliminating the donation

In this appendix, we analyse the effect of the size of the group on voting *Yes* using:

$$Accept_i = \beta_0 + \beta\ Exp.Group_i + \gamma\overrightarrow{Z_\iota} + \varepsilon_i$$

In this equation $Accept_i$ is a binary variable that takes the value of 1 when the participant voted *Yes* in the pivotal situation. $Exp.Group_i$ is a binary variable per each of the experimental conditions. In Column (1) we compare the conditions with social norms of 0.5, meaning that half of the participants who had already voted did it in favour of the generation of harm. In Column (2) we compare the conditions with social norm of 1. Finally, $\overrightarrow{Z_\iota}$ is the vector of controls we use to increase precision of our estimates.

		(1)		**(2)**	
		Social norm 0.5		**Social norm 1**	
		β/γ	**+/- 95% CI**	**β/γ**	**+/- 95% CI**
	S:11 & T:6	0.046	[-0.008,0.099]		
		One sided p-value: 0.048			
	S:4 & T:4			0.058	[0.01,0.11]
				One sided p-value: 0.008	
	Female	-0.029	[-0.083,0.024]	-0.071	[-0.12,-0.023]
			p=0.282		*p=0.004*
	Age	-0.003	[-0.006,-0.001]	-0.004	[-0.006,-0.002]
			p=0.003		*p <0.001*
Education	**High school**	-0.009	[-0.29,0.27]	0.276	[-0.18,0.73]
			p=0.948		*p=0.234*
	Bachelor's	-0.000	[-0.28,0.28]	0.289	[-0.17,0.74]
			p=0.999		*p=0.214*
	Master's	-0.059	[-0.34,0.22]	0.252	[-0.21,0.71]
			p=0.682		*p=0.282*
	Doctorate	-0.084	[-0.41,0.24]	0.274	[-0.20,0.75]
			p=0.610		*p=0.258*
	Constant	0.924	[0.63,1.22]	0.703	[0.24,1.16]
			p <0.001		*p=0.003*
	R^2 Adj.	0.014		0.028	
	N	896		905	

Note: We use robust standard errors. We present +/-95% confidence intervals in square brackets. We use one-sided p-values because our hypotheses are unidirectional. In Column (1) the baseline value for the experimental condition is *S:3 & T:2*. In Column (2) the baseline value for the experimental condition is *S:2 & T:2*.

A.6 Main Results about the social norms on eliminating the donation

In this appendix, we analyse the effect of the social norms on voting *Yes* using:

$$Y_i = \beta_0 + \beta \; Exp.Group_i + \gamma \overrightarrow{Z_\iota} + \varepsilon_i$$

In this equation Y_i is a binary variable that takes the value of 1 when the participant voted "Yes". $Exp.Group_i$ takes the value of 1 when the participant is either in a group with a perceived social norm of 1, Column (1), or in the condition "*S:3 & T:2*", Column (2). $\overrightarrow{Z_\iota}$ is the vector of covariates we also used in Appendix 6.

		(1)		(2)	
		SN 1 vs SN 0.5		*S:3 & T:2* vs *S:2 & T:2*	
		β/γ	+/- 95% CI	β/γ	+/- 95% CI
Social norms: 1		0.047	[0.011,0.083]		
	One sided p-value		*0.005*		
S:3 & T:2				-0.041	[-0.094,0.012]
	One sided p-value				*0.066*
Female		-0.049	[-0.085,-0.013]	-0.048	[-0.10,0.006]
	p-value		*0.007*		*0.080*
Age		-0.004	[-0.005,-0.002]	-0.003	[-0.005,-0.001]
	p-value		<0.001		0.004
Education	**High school**	0.088	[-0.15,0.33]	-0.079	[-0.31,0.15]
	p-value		*0.474*		*0.496*
	Bachelor's	0.097	[-0.15,0.34]	-0.094	[-0.32,0.13]
	p-value		*0.433*		*0.414*
	Master's	0.052	[-0.19,0.30]	-0.153	[-0.39,0.084]
	p-value		*0.680*		*0.205*
	Doctorate	0.056	[-0.21,0.32]	-0.129	[-0.41,0.15]
	p-value		*0.674*		*0.369*
Constant		0.862	[0.61,1.11]	1.052	[0.81,1.30]
	p-value		*<0.001*		*<0.001*
R^2 Adj.			0.021		0.016
N			1801		905

Note: We use robust standard errors. We present +/-95% confidence intervals in square brackets. We use one-sided p-values because our hypotheses are unidirectional. In Column (1) the baseline value for the experimental conditions with social norm 0.5. That is, the "*S:3 & T:2*" and "*S:11 & T:6*" conditions. In Column (2) the baseline value for the experimental condition is "*S:2 & T:2*".

A.7 Exploratory results: effects of social norms on predicted Yes votes

In this appendix, we analyse the effect of the social norms on the percentage of *Yes* votes participants predicted there would be in their situation:

$$Y_i = \beta_0 + \beta\ Exp.Group_i + \gamma\overrightarrow{Z_\iota} + \varepsilon_i$$

In this equation Y_i is the answer to the question "*If we offer this exact deal to 100 people, how many do you think would vote Yes*?" $Exp.Group_i$ takes the value of 1 when the participant is either in a group with a perceived social norm of 1, Column (1), or in the condition "*S:3 & T:2*", Column (2). $\overrightarrow{Z_\iota}$ is the vector of covariates we also used in Appendix 6.

		(1)		(2)	
		SN 1 vs SN 0.5		***S:3 & T:2* vs *S:2 & T:2***	
		$\boldsymbol{\beta / \gamma}$	**+/- 95% CI**	$\boldsymbol{\beta / \gamma}$	**+/- 95% CI**
Social norms: 1		3.837	[2.11,5.57]		
	One sided p-value		*<0.001*		
S:2 & T:2				2.623	[0.18,5.06]
	One sided p-value				*0.017*
Female		-2.545	[-4.27,-0.82]	-2.333	[-4.76,0.098]
	p-value		*0.004*		*0.06*
Age		-0.038	[-0.11,0.029]	0.038	[-0.053,0.13]
	p-value		*0.269*		*0.413*
Education	**High school**	4.102	[-6.43,14.6]	-1.559	[-15.1,12.0]
	p-value		*0.445*		*0.821*
	Bachelor's	5.523	[-4.97,16.0]	1.278	[-12.2,14.7]
	p-value		*0.302*		*0.852*
	Master's	3.179	[-7.51,13.9]	-2.036	[-15.9,11.8]
	p-value		*0.56*		*0.773*
	Doctorate	3.152	[-8.33,14.6]	-4.145	[-19.2,10.9]
	p-value		*0.59*		*0.590*
Constant		66.463	[55.7,77.2]	67.837	[54.0,81.7]
	p-value		*<0.001*		*<0.001*
R^2 Adj.			0.015		0.009
N			1801		905

Note: We use robust standard errors. We present +/-95% confidence intervals in square brackets. We use one-sided p-values because our hypotheses are unidirectional. In Column (1) the baseline value for the experimental conditions with social norm 0.5. That is, the "*S:3 & T:2*" and "*S:11 & T:6*" conditions. In Column (2) the baseline value for the experimental condition is "*S:2 & T:2*".

A.8 Exploratory results: effects of size of the group on predicted Yes votes

In this appendix, we analyse the effect of the size of the culprit group on the percentage of *Yes* votes participants predicted there would be in their situation:

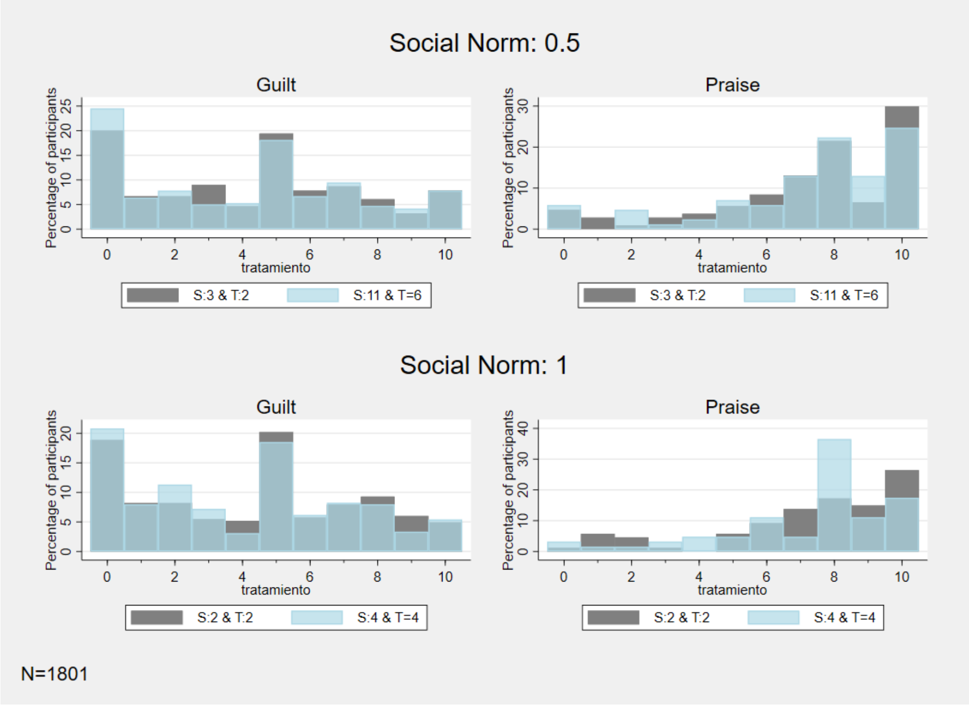
$$Y_i = \beta_0 + \beta\ Exp.Group_i + \gamma\vec{Z_\imath} + \varepsilon_i$$

In this equation Y_i is the answer to the question "*If we offer this exact deal to 100 people, how many do you think would vote Yes*?" $Exp.Group_i$ is a binary variable per each of the experimental conditions. In Column (1) we compare the conditions with social norms of 0.5, meaning that half of the participants who had already voted did it in favour of the generation of harm. In Column (2) we compare the conditions with social norm of 1. $\vec{Z_\imath}$ is the vector of covariates we also used in Appendix 6.

		(1)		**(2)**	
		Predicted *Yes* votes		**Predicted *Yes* votes**	
		$\boldsymbol{\beta/\gamma}$	**+/- 95% CI**	$\boldsymbol{\beta/\gamma}$	**+/- 95% CI**
S:11 & T:6		0.978	[-1.48,3.43]		
	One sided p-value		*0.217*		
S:4 & T:4				3.558	[1.11,6.01]
	One sided p-value				*0.002*
Female		-2.932	[-5.38,-0.48]	-2.113	[-4.55,0.32]
	p-value		*0.019*		*0.089*
Age		-0.045	[-0.14,0.049]	-0.032	[-0.13,0.064]
	p-value		*0.348*		*0.513*
Education	**High school**	7.973	[-4.23,20.2]	-3.973	[-23.4,15.5]
	p-value		*0.200*		*0.689*
	Bachelor's	8.739	[-3.39,20.9]	-1.610	[-21.0,17.8]
	p-value		*0.158*		*0.871*
	Master's	7.006	[-5.42,19.4]	-5.03	[-24.7,14.6]
	p-value		*0.269*		*0.616*
	Doctorate	3.26	[-10.6,17.1]	-1.691	[-22.2,18.8]
	p-value		*0.644*		*0.871*
Constant		63.037	[50.4,75.7]	75.575	[56.0,95.1]
	p-value		*<0.001*		*<0.001*
R^2 Adj.			0.005		0.01
N			896		905

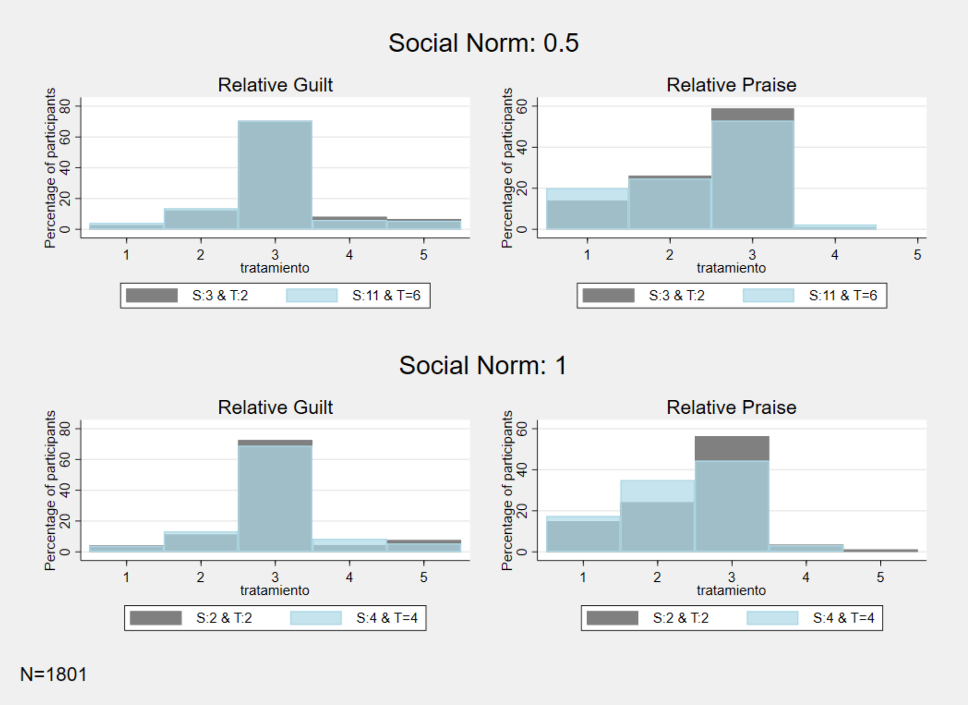
Note: We use robust standard errors. We present +/-95% confidence intervals in square brackets. We use one-sided p-values because our hypotheses are unidirectional. In Column (1) the baseline value for the experimental condition is *S:3 & T:2*. In Column (2) the baseline value for the experimental condition is *S:2 & T:2*.

A.9 Feelings of guilt and praise by experimental condition



Note: Histograms by experimental conditions of answers to the question: "*From 1 to 10, how morally responsible do you feel for eliminating (making) the donation? (0: not at all, 10: totally).*" Participants who voted "Yes" in the pivotal question answered the version with eliminating the donation (guilt), and those who voted "No" answered the version with making the donation (praise).

A.10 Feelings of relative guilt and praise by experimental condition



Note: Histograms by experimental conditions of answers to the question: "*How morally responsible for eliminating (making) the donation do you feel in comparison to the other participants who voted like you?*" With answers from 0 (*I feel much more responsible*) to 5 (I feel much less responsible) Participants who voted "Yes" in the pivotal question answered the version with eliminating the donation (guilt), and those who voted "No" answered the version with making the donation (praise).

A.11 Exploratory analysis of proportion of acceptance by type of power

As we explain in section 3, we collected incentivised answers from all possible group situations in our experiment. Each group situation varied in the number of participants that had voted before and the proportion that had accepted (rejected) the 0.5 USD bonus for voting in favour of eliminating the donation. As we mention in the methods section, each participant voted in five group situations. One pivotal and four non-pivotal. The four non-pivotal situations were selected randomly from the pool of 85 possible situations. We informed participants that we would pay them according to one of their choices, but we did not disclose which one. Our objective was to implement, without deception, the pivotal situations that were useful for our analysis.

Notice that in each group-situation, the individual's decision has a different type of consequence for the joint decision. For example, participants in a group of the type *S:11 & S:6* who are the first to vote have full uncertainty whether their vote will contribute to eliminate or save the donation. In contrast, participants in that same group who know that more than six others have voted *Yes* know that their vote is irrelevant and the donation will be eliminated either way. We call *type of power* the relevance of a vote in a particular group situation. We can classify the types of power in all group situations as:

Irrelevant vote type 1 The donation will be eliminated no matter the participants' vote

Irrelevant vote type 2 The donation will be made no matter the participant's vote.

Trigger vote By voting *Yes*, the participant ensures that the donation will be eliminated. However, voting *No* is not enough to guarantee that it will be saved.

Veto vote By voting *No*, the participant makes sure that the donation will be saved. However, voting *Yes* does not guarantee that it will be eliminated.

Pivotal vote By voting *No*, the participant makes sure that the donation will be saved. In addition, by voting *Yes,* the participant guarantees that it will be eliminated. In other words, it has simultaneously veto and trigger power.

Contribution vote The participant's vote is not enough either to trigger the elimination or to ensure it.

In this appendix, we make an effort of analysing the effect of the group situation on the acceptability of collaborating in causing harm (i.e. voting *Yes* to the elimination of the donation to charity). We consider this analysis as exploratory and we caution interpreting these results as causal. The reasons derive from the fact that, to answer our research question, we designed the experiment to analyse only pivotal decisions. As we explain in section 3, this responds to the necessity of disentangling the effect of the size of the group and social norms from other mechanisms such as aversion to pivotality and replacement logic[9]. The reasons are threefold. First, in the experiment we varied two things: the configuration of the group (number of members and threshold), and information about the behaviour of previous voters. However, it is likely that this information became less credible after answering more than one case[10]. Second, since all cases where similar in structure it is possible that participants were more inattentive to the last cases in comparison with the first. Finally, it is possible that participants suspected the objective of the experiment after noticing that only a few elements in the cases were changing.

We make a within-subject analysis of the effect of the power implied in each group situation on in favour of the elimination of the donation. We use the following specification.

$$Accept_{i,k} = \beta_0 + \beta\ Power_{i,k} + \gamma_1 \overrightarrow{Z_\iota} + \gamma_2 GroupConf_{i,k} + \gamma_2 ExpGroup_k + \varepsilon_i$$

In this equation $Accept_{i,k}$ is a binary variable that takes the value of 1 when the participant i voted "Yes" in the situation k. $Power_{i,k}$ is a binary variable that takes the value of 1 when situation k implies a particular type of power. $GroupConf_{i,k}$ is the group configuration, that is the number of participants in the group and its threshold in situation k. Since all participants answered first the pivotal situation, we include $ExpGroup_k$, which is a binary variable that takes de value of one for the pivotal situation answered by the participant. $\overrightarrow{Z_\iota}$ is the set of controls that we mention in the table. In the table, the baseline power is *Irrelevance 1*, meaning that the damage will not happen independently of the participants' decision.

[9] Other experiments are adequate to compare other types of power. For example, Falk and coauthors (2022) compare situations that we classify as *Pivotal* to *Irrelevant type 1* to answer their research question about aversion to pivotality. Our exploratory results, presented in this appendix, confirm their observations.

[10] For example, in the pivotal situation in the group *S:4 & S:4* participants received information that three others had voted before them and all voted *Yes*. It is possible that this information changed the credibility of a case when they received the opposite information.

Table A.11.1 Proportion of acceptance by type of power

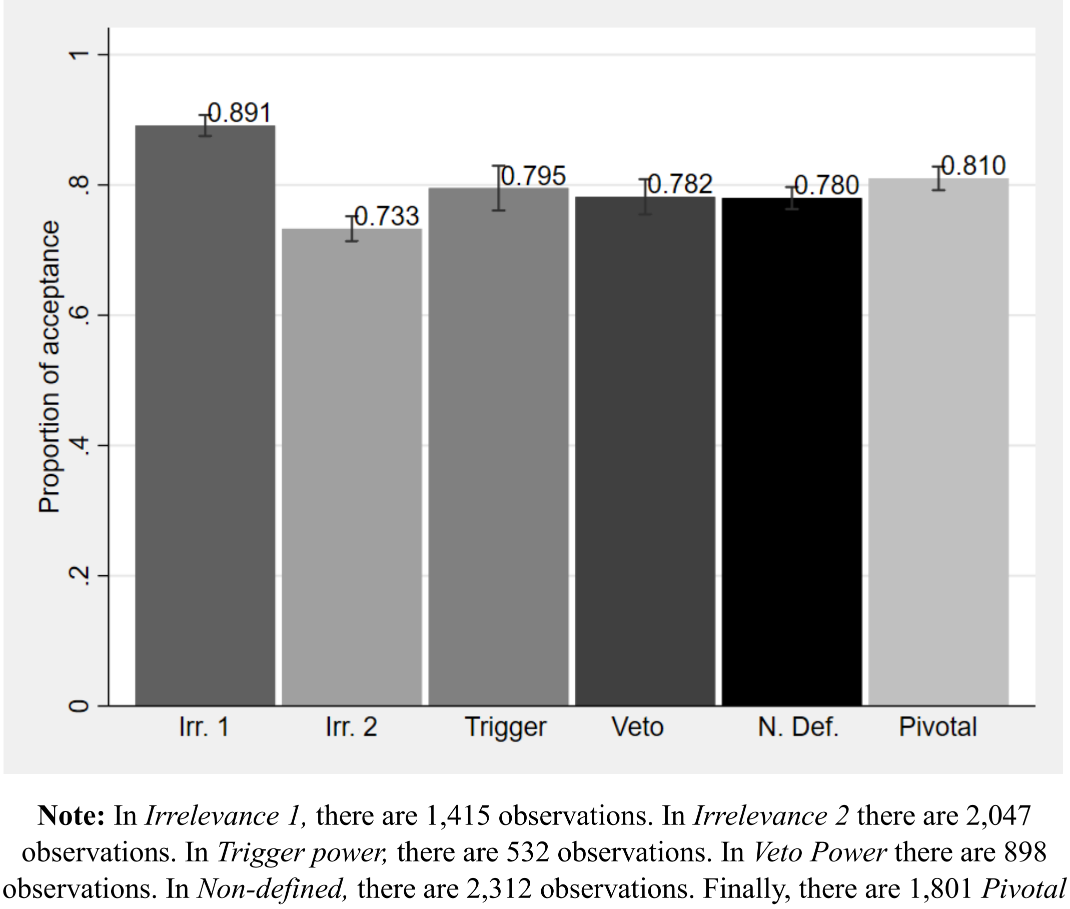
		(1)		(2)	
		Voted "Yes"		Voted "Yes"	
		β	**+/- 95% CI**	***β***	**+/- 95% CI**
Power	**Irrelevant 2**	-0.158	[-0.18,-0.13]	-0.172	[-0.20,-0.14]
	p-value		*<0.001*		*<0.001*
	Trigger	-0.096	[-0.13,-0.058]	-0.099	[-0.14,-0.061]
	p-value		*<0.001*		*<0.001*
	Veto	-0.109	[-0.14,-0.078]	-0.126	[-0.16,-0.093]
	p-value		*<0.001*		*<0.001*
	N. Def.	-0.111	[-0.13,-0.088]	-0.112	[-0.14,-0.088]
	p-value		*<0.001*		*<0.001*
	Pivotal	-0.081	[-0.11,-0.057]	-0.104	[-0.13,-0.074]
	p-value		*<0.001*		*<0.001*
Threshold	**T=2**			-0.040	[-0.084,0.0051]
	p-value				*0.082*
	T=4			-0.005	[-0.046,0.037]
	p-value				*0.820*
	T=11			-0.045	[-0.085,-0.0049]
	p-value				*0.028*
Female				-0.040	[-0.056,-0.023]
	p-value				*<0.001*
Age				-0.002	[-0.0027,-0.0013]
	p-value				*<0.001*
Education				-0.005	[-0.098,0.088]
	p-value				*0.914*
	Bachelor's			0.002	[-0.091,0.094]
	p-value				*0.972*
	Master's			-0.044	[-0.14,0.050]
	p-value				*0.359*
	Doctorate			0.007	[-0.095,0.11]
	p-value				*0.895*
Pivotal situation	**S:3-T:2**			-0.004	[-0.030,0.022]
	p-value				*0.746*
	S:4-T=4			0.067	[0.043,0.092]
	p-value				*<0.001*
	S:11-T:6			0.034	[0.0088,0.059]
	p-value				*0.008*
Constant		0.891	[0.87,0.91]	1.014	[0.91,1.12]
	p-value		*<0.001*		*<0.001*
R^2 Adj.			0.014		0.028
N			9005		9005

Note: We use robust standard errors. We present +/-95% confidence intervals in square brackets. In column (1), we do not add any control, while in column (2) we add socioeconomic covariates, the pivotal situation they answered first, and the threshold in the group of each decision as controls.

Table A.11.1 shows that in all types of power, participants voted less in favour of eliminating the donation than in power called *Irrelevance 1* (the donation will be eliminated no matter the participants' vote). This is partially congruent with our theoretical model. On the one hand, the model predicts that participants with Irrelevance 1 power should have a larger proportion of participants taking the selfish option. However, it also that all voters with an irrelevant type of power should vote yes if the monetary incentive is positive, which does not happen for the *Irrelevance 2* power (the donation will be made no matter the participants' vote). Also being female and younger were related to voting *No* in each group situation.

In figure A.11.1 we present in a graphic way these results. The number of observations in each bar varies since there are different number of group situation in each configuration.

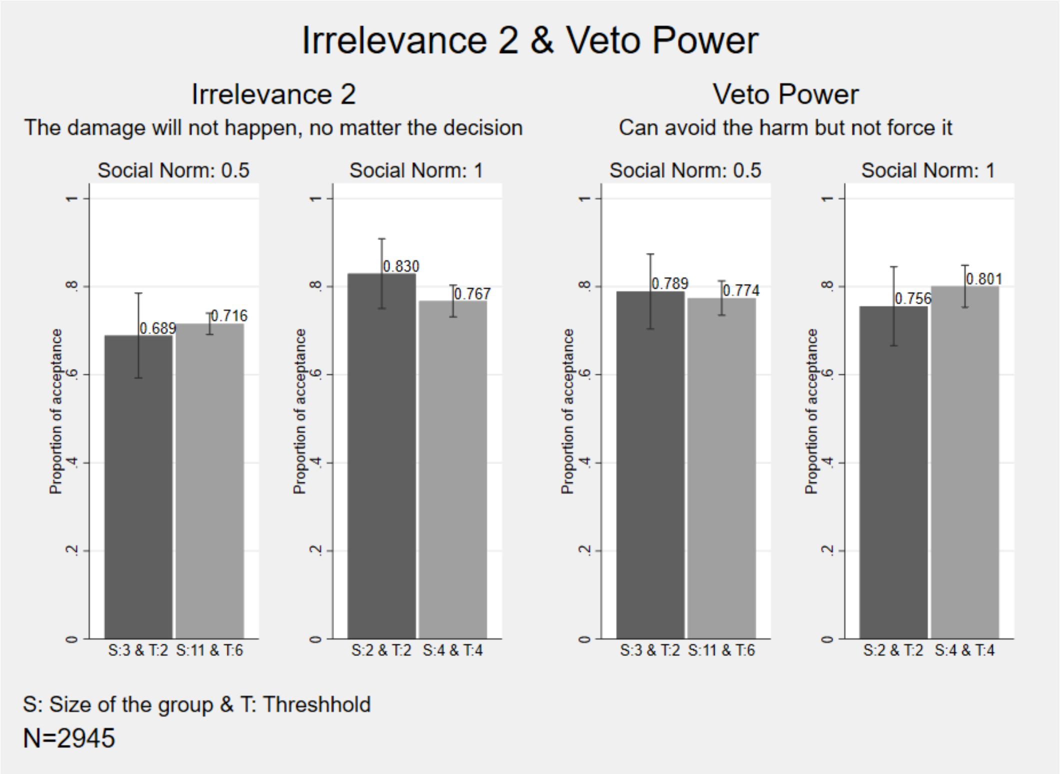
Figure A.11.1 Proportion of *Yes* votes by type of power



Note: In *Irrelevance 1,* there are 1,415 observations. In *Irrelevance 2* there are 2,047 observations. In *Trigger power,* there are 532 observations. In *Veto Power* there are 898 observations. In *Non-defined,* there are 2,312 observations. Finally, there are 1,801 *Pivotal* observations. The error bars show the +/-95% confidence intervals.

To further analyse our data, we focus on the two types of power that, besides *pivotality*, are present in all four group configurations: *Irrelevance 2* and *Veto Power*. This is a natural occurrence. For example, in the groups with configuration *S:2 & T:2* and *S:4 & T:4,* no participant has *Irrelevance 1* power (the donation will be eliminated no matter the participants' vote). Figure A.11.2 shows comparisons across the four group configurations. We do not observe any statistically significant difference.

Figure A.11.2 Proportion of *Yes* votes in Irrelevant (type 2) and Veto power cases



Note: In *Irrelevance 2,* there are 2,047 observations. In *Veto Power,* there are 898 observations. Finally, there are 1,801 *Pivotal* observations. The error bars show the +/-95% confidence intervals.

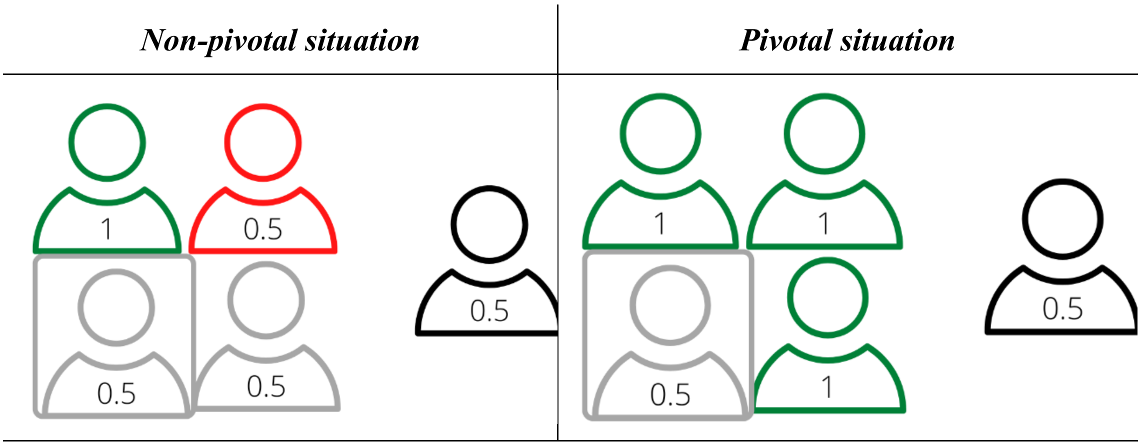
Future research could focus on designing experiments to analyse how different types of power affect collaboration in causing harm in group decisions. A starting possibility would be implementing an experiment with all types of power for a single group configuration. However, since our paper demonstrates that the culprit group size matters, it would be relevant to later interact those results with other group configurations.

A. 12 Instructions to participants

Each participant answered first one pivotal group situation and then four non-pivotal group situations. There are 85 potential group situations. The values of S, T, n_a (number of participants that voted Yes) and n_r (number of participants that voted No) presented in this supplementary material depend on the group configurations presented in Table 1.

Participants read a description of the situation and we presented an image to clarify it. The left panel of the following figure corresponds to the group configuration *S:4 & T4*. The green figure represents a participant who voted *Yes* and received a 0.5 USD additional bonus. The red figure is a participant who voted *No* and received no additional bonus. The black figure represents the donation to the charity. In grey, there are two participants yet to vote. The grey figure inside the square is the participant answering this situation. In the right panel, we present the pivotal situation of that configuration.

Group situations – examples



Non-pivotal situation

Pivotal situation

Note: This figures are simmilar to the ones used in the experiments' instructions. Appendix 1 presents the four figures for the pivotal situations.

Informed consent letter

Procedures

This study consists of two parts. In the first part, you will make 5 decisions that could have implications for your payments and a donation to a charity. In the second part, you will answer questions about those decisions and about yourself. You will be given instructions on your screen before every part and every decision. Please make sure always to read the instructions carefully. Some questions control that you read and understand the instructions.

Participation

Participation in this research study is entirely voluntary. You have the right to withdraw at any time or refuse to participate entirely without jeopardy to future participation in other studies conducted by us.

Confidentiality

Your ProlificID will be used solely for the purpose of making the payments for participation in the study. All data obtained from you will be anonymised after the payments have been made. Anonymised data will be analysed and made available in open science repositories after the study's conclusion.

Payment

Your payment for participating in this project consists of a 1.3 USD participation fee and a first bonus of 0.5 USD that you are guaranteed to get. This sums 1.8 USD that you will get for your answers. You can get a second 0.5 USD bonus during the study. One of the 5 decisions that you will make in the first part of the study will be implemented and will have real consequences on whether you or a charity get the extra bonus. Therefore, you should answer all decisions considering that they could be implemented. Your payment for participating in the study will be sent to you shortly after completing the task. Your bonus will be paid using the bonus system within three weeks.

Questions about the research project

If you have questions regarding this study, you may contact: thechoicelab@nhh.no

Consent

Please select ACCEPT in the box below if you have understood the above and wish to participate in this study. [I accept] [I don't accept]

Remember

Your initial payment for participating in this project consists of a 1.3 USD participation fee and a bonus of 0.5 USD that you are guaranteed to get. One of the decisions that you will make in the first part of the study will be implemented and will have real consequences on whether you or a charity get an additional bonus of 0.5 USD. Therefore, you should answer all decisions considering that they could be implemented.

Part I. Group situation questions

You are part of a group of S individuals. You are going to decide as a group by voting. The decision is about eliminating a 0.5 USD donation to a charity. If T or more vote *Yes*, we will eliminate a donation of 0.5 USD to a charity, and we will use that money for something else. However, if less than T participants vote *Yes,* we will donate the money to a charity. All participants are offering the following:

- If you vote **"Yes"**, you will be paid 0.5 USD added to your initial bonus independently of the vote of the rest. That means that you will receive a final bonus of 1 USD.
- If you vote **"No"**, you will not receive any money added to your initial bonus independently of the vote of the rest.

N participants have answered before you and of them n_a voted *Yes* and n_r voted *No*. This means that if $T - n_a$ participants vote *Yes*, the donation will be eliminated. $T - N$ participants will vote after you. We will not reveal your identity to the other participants. The participants who voted before you will never know your decision.

Additional text for those who are not the final participant to decide in the group:
You will not know the decision of those who will vote after you.

In the image, you can see the situation so far. In green, you can see the participants who voted YES, with their final bonus. In red, you can see the participants who voted NO with their final bonus. The figures in gray represent the participants who are yet to vote. You are represented by the gray figure in the square. The figure in black represents the donation to the charity.

[Insert image with a group state defined by of S, T, N, n_a, n_r]

Q1: How do you decide to vote?

[Yes (receive 0.5 USD as an additional bonus)]

[No (do not receive 0.5 USD as an additional bonus)]

Part II. Exploratory and background questions

In the previous section of this study, you were presented with the following scenario:

[Description of the pivotal group-state answered by the participant]

You voted Yes/No. Therefore, the donation to charity will be made/eliminated.

Q2: If we offer this exact deal to 100 people, how many do you think would vote "*Yes*"?

Options: (Slider with options from 0 to 100)

Q3: From 0 to 10, how morally responsible do you feel for making/eliminating the donation? [0: I'm not responsible at all, 10: I have the full responsibility]

Options: [from 0 to 10]

Q4: How morally responsible for making/eliminating the donation you feel in comparison to the other participants who voted like you?

Options: (1: I feel much more responsible) (2: I feel slightly more responsible) (3: I feel equally responsible) (4: I feel slightly less responsible) (5: I feel much less responsible)

Q5: What is your gender?

Options: (Male) (Female) (Other)

Q6: How old, in years, are you?

Options: (Number entry box)

Q7: What is the highest level of education that you have finished so far?

Options: (Less than high school) (High school) (Bachelor's degree) (Master's degree) (Doctorate degree)

Q8: How would you describe your political opinions? (0: extreme left, 10: extreme right)

Options: (0) (1) (2) (3) (4) (5) (6) (7) (8) (9) (10)

Figures

Tables