# Why do oaths work?*

Sorravich Kingsuwankul† Chloe Tergiman‡ Marie Claire Villeval§

July 29, 2025

**Abstract**

We investigate the channels through which honesty oaths shape behavior and communication in settings with asymmetric information. We manipulate whether oath-taking is public and investigate three non-pecuniary costs of oath-breaking. Our results highlight the importance for oath-takers of image costs related to self-image, and image costs stemming from how other players perceive them, while suggesting a lesser concern for passive observation. Despite the significant impact of oaths on oath-takers, the mere knowledge that an oath has been signed is neither sufficient nor necessary to generate trust in oath-takers: instead, less informed players rely on their experience.

**Keywords:** Promise-keeping, Honesty Oaths, Common Knowledge, Deniability, Image Costs, Laboratory Experiment

**JEL codes:** C91, D83, D91

*Acknowledgements: We are grateful to N. Ali, B. Corgnet, L. Gangadharan, F. Gul, N. Jacquemet, R. Kerschbamer, C. Raymond, H. Schneider, V. Valero, A. Viriyavipart, and participants at presentations at the Applied Ethic and Morality Research Group at the University of Prague, the Credence Goods and Expert Markets Webinar, the Innsbruck Winter summit on (Un)Ethical Behavior in Markets, the IAER Microeconomics Workshop at Dongbei University, the PIER Research Workshop, Cornell University, Michigan Ross TO, Penn State University, Stanford University, Queen's University, the East Coast Conference, ASFEE in Montpellier, MBEES-MBEPS in Maastricht, the ESA World meeting in Lyon, the TIBER symposium in Tilburg, and the European Economic Association meeting in Barcelona for useful comments. We are thankful to Q. Thevenet for his assistance in programming the experiment. This research has benefited from the financial support of IDEXLYON from Université de Lyon (INDEPTH-IDEX/SBP/2018/03) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) operated by the French National Research Agency, and of the LABEX CORTEX(ANR-11-LABX-0042) of Université Claude Bernard Lyon 1, within the program "Investissements d'Avenir" (2019- ANR-LABX-02) operated by the French National Research Agency (ANR).

†School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. Tinbergen Institute, 1082 MS Amsterdam, The Netherlands. Email: s.kingsuwankul@vu.nl

‡**Corresponding Author.** Pennsylvania State University, Smeal College of Business, 334 Business Building, University Park, PA 16802, United States. Email: cjt16@psu.edu

§CNRS, Université Lumière Lyon 2, Université Jean-Monnet Saint-Etienne, emlyon business school, GATE, 35 rue Raulin, F-69007, Lyon, France. Email: villeval@gate.cnrs.fr

# 1 Introduction

Oaths exist in a myriad of real-world scenarios. For instance, in the legal system, individuals are required to take an oath before testifying in court. Similarly, within professions such as medicine, doctors take the Hippocratic Oath, a pledge to uphold ethical standards and prioritize patient well-being. Accountants and, in certain countries like the Netherlands and Belgium, even bankers, also swear oaths of ethical conduct. In the political sphere, leaders take inauguration oaths to signify their commitment to upholding the constitution of their countries and serving the interests of their constituents.

Despite the inherent information asymmetry in these situations — where only the oath-taker knows definitively if they've adhered to its principles — oaths are considered meaningful even though they lack the legal enforceability of contracts. They may instill an internalized feeling of moral duty, creating a "psychological contract" that is psychologically costly to break, in contrast to the external consequences of breaching legally binding contracts.

The main goal of this study is to evaluate the channels through which an oath can have an impact on behavior. While acknowledging that oaths have been shown to have demonstrable effects, understanding why individuals feel compelled to honor oaths and teasing apart the underlying mechanisms may allow one to tailor them to specific settings.

To evaluate the non-pecuniary costs of breaking an oath, we use a game that captures some of the important elements of the world outside the laboratory where oaths are used: asymmetric information and misaligned preferences. Our game consists of two players, A and B. The A player receives three cards at random, where each card has a 0.5 probability of displaying a star and otherwise is blank. Cards with a star represent a successful outcome and the fraction of cards that have a star on them determines the quality of the A player's portfolio. Information is asymmetric: only the A player sees the face of all three cards. The A player then sends a cheap-talk message about the number of stars to the B player. The B player, who is given an endowment, then decides whether to invest that endowment with the A player or not. Finally, Nature draws one of the A player's cards to determine payment for the B player. If the randomly drawn card has a star on it, the investment is a success and B earns a handsome payoff. Otherwise, it is a failure and B loses the investment. The A player's payment on the other hand is contingent on the investment decision, not its success, creating a conflict of interest between the two players when the state of the world is unfavorable (i.e., when there are few stars among the three cards). Our treatments vary whether A players have to sign an oath or not, and whether that promise to be honest is common knowledge.

In our setting, akin to real-life scenarios, lies can be either *detectable* or *deniable*. De-

tectable lies happen when the A player announces three stars after observing fewer than three stars, exposing the lie if Nature draws a blank card. Deniable lies happen when, for instance, the A player announces two stars after observing fewer than two stars, allowing for plausible deniability irrespective of Nature's draw.[1] Allowing for a richer set of lies allows us to more precisely understand the mechanism through which an oath can impact behavior: it allows us to separate moral costs related to other players being able to detect a lie, versus moral costs related to lies that only an informed audience or the A players would know have occurred.

We focus first on the behavior of the oath-taker: if an oath works, why does it work? We draw from the literature on preferences for promise-keeping and truth-telling (reviewed in the next section) and investigate the importance of three non-pecuniary costs of breaking a promise. The terminology we use to name these costs places us from the vantage point of the oath taker, and specifically relates to the three types of audiences that witness the oath. "Active-audience image costs" are costs that the oath-taker can incur because of how other players who witness the oath and subsequently take actions in the game (in our setting, the B players) may perceive them if they appear to have broken their oath.[2] "Passive-audience image costs" are costs that the oath-taker can incur because of how a *passive* audience — those who make no decisions in the game but who may form beliefs about the oath-taker after observing their actions — perceives them.[3] In our setting, the passive audience may be the experimenter. Lastly, "self-audience image costs" are those costs that may prevent the oath-taker from breaking a promise even when their actions are not observed by anyone else. These costs may arise from the discomfort of seeing oneself as possessing a low moral identity.[4]

After investigating how oaths increase the veracity of the oath-taker's announcement, we then study how they impact the B players: is knowledge of the oath necessary or sufficient to increase the credibility of the oath-taker's announcement? Or, when judging the truthfulness of the announcement of someone who has taken an oath, do people instead predominantly rely on their experience, independently of the knowledge that an oath has been taken?

[1] This game, known as the "Announcement Game," was first described in Tergiman and Villeval (2023) and stands out from others in that types of lies available (detectable or deniable) are richer than in other works, and the announcer chooses the degree of detectability endogenously, features that are closer to situations outside the laboratory.

[2] These costs may prevent an individual from breaking a promise if that individual does not want to be perceived by other active players as someone who would indeed not keep their word (see, *e.g.*, Dufwenberg and Dufwenberg, 2018; Gneezy et al., 2018; Abeler et al., 2019) or does not want to be seen as someone for whom it is easy to break a promise, that is, someone with a low private cost of doing so (see, *e.g.*, Kartik, 2009).

[3] This description of an audience is in line with that of Abeler et al. (2019).

[4] These costs may result from the fact that an oath encourages moral deliberation in oath-takers (de Bruin, 2016) or from an intrinsic preference for keeping one's words resulting from the internalization of a moral norm (Ellingsen and Johannesson, 2004). Our view of these costs aligns with Dufwenberg and Dufwenberg (2018) and Gneezy et al. (2020) who interpret self-image costs as costs coming from being one's own audience.

Our six between-subject treatments, run with 757 participants, re-match participants across periods and vary whether and how an oath is introduced.[5] In the ***No-Oath*** treatment, the A players do not swear any oath. This gives us a baseline for the distributions of the announcements by the A players and of the investment rates for each type of announcement by B players. In the ***Public-Oath*** treatment, A players swear a compulsory oath to be truthful, and this is common knowledge among all participants. Comparing data from the No-Oath and Public-Oath treatments allows us to directly measure the impact of such an oath on behavior. Overall, the fraction of fully truthful advisors goes from about 10% in the absence of an oath, to about 75% when a public oath is present. Public oath-taking has a dramatic impact on all types of lies: the fraction of A players who make detectable lies drops significantly from 71% to 12%, and the fraction who engage in deniable lies goes from 87% to 23%, a drop that is even more surprising since by definition these lies cannot be detected by B players.

After documenting the dramatic impact of the oath on A players' behavior, we investigate why the oath is so effective. When someone promises to tell the truth, which non-pecuniary costs explain promise-keeping? To evaluate the impact of active-audience image costs, we designed the ***Private-Oath*** treatment. In this treatment, A players swear an oath, but B players are kept in the dark regarding its existence, a fact that the A players are aware of. Comparing behavior in the Public-Oath and Private-Oath treatments allows us to isolate the impact of active-audience costs. Indeed, in both treatments, the oath-taker and the passive audience, in particular the experimenter, know that an oath has been sworn as well as the truth regarding the number of stars and the announcement made. The only difference between these treatments is whether the B player is aware of the oath, that is, whether active-audience costs can exist vis-à-vis an A player who breaks a promise. When the oath is privately taken, "only" about 40% of A players consistently tell the truth. Therefore, active-audience costs can account for the 35 percentage point increase in truth-tellers that we observe in the Public-Oath treatment, where about 75% of A players tell the truth at all times.

To evaluate the joint effect of self-audience and passive-audience costs, we compare the No-Oath and Private-Oath treatments. Indeed, both these treatments keep active-audience image concerns toward B players constant: the B players face identical instructions across these two treatments and, thus, have no reason to view the A players differently in the two treatments given a particular announcement and a particular card drawn by Nature. In other words, because signing the oath is a private event, it cannot lead to additional image

[5]The stranger matching protocol allows us to focus on the impact of oaths rather than on other elements that would arise if participants were in fixed pairs or if the history of play was known. In particular, this eliminates reputation and the possibility of direct punishment that would exist in the case of fixed matching.

costs induced by what B players would think about an A player who would break that oath. The two treatments instead differ in the following way: while in both treatments an A player who announces a number of stars that differs from the truth may incur a non-pecuniary cost of lying about the true number of stars, A players who lie in the Private-Oath treatment may incur additional costs: the self-audience and passive-audience costs associated with breaking a promise. Since in the No-Oath treatment, the fraction of truth-tellers is about 10% and is about 40% in the Private-Oath treatment, we find that self- and passive-audience image concerns can jointly contribute to 30% additional subjects moving to tell the truth. To better characterize the role of passive-audience image costs, we ran a ***Private-Oath***$^{++}$ treatment that replicated the Private-Oath treatment, but with greater salience of the fact that the experimenter was unable to identify each participant. The similarity of behavior in both the Private-Oath treatment and the Private-Oath$^{++}$ treatment provides evidence that the effect of the private oath on promise-keeping stems primarily from self-audience image costs, while passive-audience image costs only play a minor role, which is corroborated by the responses to a post-experimental questionnaire in the Private-Oath treatment.

Turning to the impact of oaths on the B players, we find that the public oath leads to higher earnings for the B players and that they interpret announcements more credibly than when there is no oath. The next two treatments aim to understand what role the oath *per se* plays in those results: are B players anticipating more honest behavior from A players because they took an oath? Or are they reacting to the truthfulness of announcements independently of the oath, that is, are they reacting to their own experience? To answer these questions we conducted the ***Necessary Condition*** treatment. In this treatment, all participants are B players who are paired with past A players from the Public-Oath treatment. They are informed that they received "an excerpt of the instructions" from the Public-Oath treatment, which are those very instructions except for the portion regarding the oath. Thus, the B players in the Necessary Condition treatment face the announcements from past A players who had taken the public oath but they are unaware of the existence of the oath. The B players from the Public-Oath and Necessary Condition treatments are thus asymmetrically informed about the oath but face the same set of A players who, as described above, were largely truthful. Despite this asymmetry, by the end of the sessions, we find little difference in the average behavior of B players in the two treatments. Therefore, the oath *per se* is not necessary to induce trust: truthful announcements lead to trust in these announcements even in a game in which deniable lies are easy to make and at the individual level difficult to detect.

If an oath is not necessary to generate trust, is it sufficient? To answer this question, we designed the ***Sufficient Condition*** treatment. In this treatment, all participants are B

players who are given the *full* instructions of the Public-Oath treatment, including the section about the oath. They are then told that in each period they will be randomly matched with a past A player chosen from a *subset* of all the past A players from that treatment. We built that subset so that the announcement behavior of the A players does not differ significantly from that of the A players in the No-Oath treatment who, as described above, were largely untruthful. Comparing investments between the Sufficient Condition treatment and the Public-Oath and No-Oath treatments informs us of how much importance knowledge of the oath has on B players and whether this impact is persistent. We find that the investment behavior in the Sufficient Condition treatment resembles that in the No-Oath treatment: knowledge of the oath is not a sufficient condition for trust to be established.

Finally, we complement our experimental findings through an online survey with residents in the Netherlands, where bankers have to swear a compulsory oath, and France, where such an oath does not exist. We find that public awareness about the oath is far from universal. This finding, together with the tendency of bankers to forget their oath (Weitzel and Kirchler, 2023), underscores the importance of oaths being common knowledge.

The remainder of this paper is as follows. Section 2 reviews the related literature. Section 3 outlines the experimental design and procedures. Section 4 reports the results of the experiment. Section 5 discusses these findings, suggests possible extentions as well as avenues for modeling, and concludes.

## 2 Related literature

Our main contribution is to the literature on preferences for promise-keeping, which has used pre-play communication to understand the impact of promises in social dilemmas and trust games. Typically, in these games, a player promises to take a given action if the first mover selects an option that may pay more to both players but that makes him or her vulnerable to the risk of earning less if the promise is broken. This literature shows that even under stranger matching protocols, individuals tend to keep their promises (Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2006; Serra Garcia et al., 2013; Ismayilov and Potters, 2016; Ederer and Stremitzer, 2017; Casella et al., 2018; Di Bartolomeo et al., 2019) and the effect persists over time (Ederer and Schneider, 2022). Vanberg (2008) demonstrates that promise-keeping is more likely driven by an intrinsic preference for keeping one's words – as advanced by Ellingsen and Johannesson (2004)– rather than the anticipation of the guilt experienced from letting down other players' expectations – as advanced by Charness and Dufwenberg (2006). However, the debate is not closed (Ederer and Stremitzer, 2017). Ismayilov and Potters (2016) argue that promise-keeping is associated with a desire for

internal consistency rather than a social obligation since promises create commitment even in settings where the promise is not known to the other party. Recent work by Saccardo and Serra-Garcia (2023) has highlighted the importance of cognitive environments in shaping moral behavior. They show that individuals actively choose contexts that either enable or limit their ability to distort beliefs self-servingly. In particular, some individuals demonstrate a clear demand for cognitive constraints that help commit them to moral behavior, reflecting an awareness of their potential for self-serving distortion. Our study complements this insight by examining how oath swearing, as a form of moral commitment, influences lying behavior through similar psychological mechanisms involving image and internalized moral standards. Sengupta and Vanberg (2023) show that people keep promises for fear of causing regret in others when the value of a foregone opportunity can be observed. In a context of promise competition, Lang and Schudy (2023) show that the transparency of promises reduces the frequency of promise-breaking but also their generosity.

We differ from these studies in that subjects in our experiment sign a compulsory honesty oath, and do not promise a particular behavior conditional on the choice of a given action by another player; it is a pledge regarding the truthful transmission of exogenous information provided to another player before she takes action. We exclude any instrumental individual reputational concerns since the players are randomly rematched in each round.[6] Finally, we go deeper into the investigation of the self-image costs of not keeping one's word by studying if A players refrain from lying even when B players are not aware that they took an oath. Note also that in contrast to our study, this literature does not consider that a passive audience external to the game (such as the experimenter) can induce additional moral costs in case of breaking a promise.

The literature on lying that has focused on individual decision-making in the absence of promises has explored the role of intrinsic preferences for truth-telling and reputational concerns. Based on a meta-analysis of studies using the die-under-the-cup paradigm (Fischbacher and Föllmi-Heusi, 2013) or its variants, Abeler et al. (2019) confront different models and show that one in which the decision to lie depends on the monetary benefits of lying, the intrinsic preference for truth-telling in the absence of an audience, and the reputational costs of being perceived as a liar by an outside audience can best fit the data.[7] Gneezy et al. (2018), Dufwenberg and Dufwenberg (2018), Khalmetski and Sliwka (2019), and Schudy et al. (2024) also highlight the importance of image concerns, in addition to a

[6]Regarding how the audience may play a role when promises are made, our study relates to Lang and Schudy (2023) who study political campaign promises in the lab. Their results suggest that the *size* of the audience may not matter: including an additional audience member who is a subject participant in the room does not impact behavior relative to when the only audience would be the experimenter.

[7]They note that inequality aversion, guilt aversion or conformity cannot rationalize the data.

fixed direct cost of lying.[8]

In our baseline setting, no oath is sworn and lying is widespread. The image costs that influence behavior in the aforementioned studies are thus not large enough to prevent players from telling lies in the absence of an oath. Nonetheless, while the literature described above has focused on individual decision-making tasks and therefore cannot consider active-audience image costs, the literature on lying provides a useful framework: we incorporate their notions of self- and audience-image costs into our analyses of the psychological costs that come with breaking an oath.

Finally, our study of course relates to the large literature that documents the impact of oaths. Table A1 in Online Appendix A provides an overview of the experimental economics literature on the effect of oath-taking (see Zickfeld et al. (2024) for a meta-analysis in psychology). The studies cited in the top panel in Table A1 show that the presence of an oath can lead to a reduction in lying in strategic settings, though the extent of this reduction depends on whether incentives are aligned or not and how the oath is implemented (compulsory versus non-compulsory, for example) (*e.g.*, Beck, 2021; Jacquemet et al., 2024). These studies, however, do not aim to understand how and which moral costs may impact the decision to uphold an oath. While we begin our own analyses by documenting the impact of an oath, our main objective goes beyond that: it is to understand *why* an oath can be impactful. Our original design allows us to identify the nature of various possible image costs, contributing to clarifying in which context and for whom an oath can have an impact.

# 3 Experimental design and procedures

We fist describe the general structure of the session, the treatment variations, then the procedure.

## 3.1 General structure

We use the Announcement Game of Tergiman and Villeval (2023), where participants are assigned either the role of Player A or B. In each period, A and B start with an endowment of 30 and 100 tokens, respectively. At the beginning of the period, a random draw determines a set of three cards for each A player, each of which has an independent 0.5 probability of displaying a star, which indicates a successful project. This probability is known to both

[8]For Gneezy et al. (2018) the reputational cost depends on the probability that others perceive the individual as a liar. For Khalmetski and Sliwka (2019), growing image concerns increase the range of lies. For Dufwenberg and Dufwenberg (2018), belief-dependent lying costs are proportional to the size of the lies perceived by the audience. Schudy et al. (2024) approaches differently by measuring individual preference types based on intentions to be, and be seen, as honest.

A and B but the number of stars (0, 1, 2, or 3) is private information to Player A. After observing the number of stars, A sends a cheap-talk message to B regarding her number of stars and, thus, can misreport. Player B then decides whether or not to invest his endowment with Player A. Next, Nature selects one of Player A's three cards to determine whether the investment is a success (if the drawn card displays a star) or a failure (if it is blank).

Irrespective of B's decision, both A and B learn about the outcome of Nature's draw (but we never inform B players about the content of the three cards). The payoff of Player A depends on whether Player B invested or not: A earns 230 tokens if B invested, 30 otherwise. The payoff of B depends on the decision to invest and the outcome of Nature's draw. If B decides not to invest, he keeps his 100 tokens. If he decides to invest and a star is drawn, the investment is successful and he earns 300 tokens (100-100+300). If instead a blank card is drawn, the investment fails and he earns 30 tokens (100-100+30). Payoffs are realized, and the game ends.

In this game, truth-telling is not supported in equilibrium. In addition, Tergiman and Villeval (2023), who have studied this game in the absence of oaths, have shown that under random rematching both detectable and deniable lies are commonplace. This game and our implementation of it in the laboratory thus allows us to study in what way and why oaths can induce participants to be more truthful.

In the laboratory, roles were fixed for the whole session. Participants played this game 18 times with random and anonymous rematching between periods.[9] At the end of the session, the program randomly selected one period to count for payment for the Announcement Game.

The Announcement Game took place in the second part of the experiment. In part 1, that is prior to receiving the instructions for the Announcement Game, participants played eight practice periods of the "Truthful Announcement Game" where they all held the role of Player B, and where each participant was paired with a truth-telling computer. These periods aimed to help participants familiarize themselves with the setting and encounter all possible cases of the announcements corresponding to the probabilities used in the Announcement Game. After the eight practice periods, participants made five incentivized decisions: in period 9, participants decided whether or not to invest in each of the four possible scenarios (*i.e.,* 0, 1, 2, 3 stars), and in period 10, they made an investment decision without being informed of the number of stars. Investment decisions in periods 9 and 10 allow us to establish a benchmark of risk attitudes to study B players' behavior in the Announcement Game. Either period 9 or 10 was randomly selected to count for payment for the Truthful Announcement Game,

[9]To avoid any end-game effects, participants were only informed that they would play a minimum of 10 periods and a maximum of 30 periods, though the number of periods, 18, was pre-determined before the experiment.

without feedback until the end of the session.

After completing the Announcement Game, all participants answered incentivized questions about their beliefs depending on their role. The B players had to estimate the proportion of honest announcements for each number of stars announced to them during the session (*i.e.,* their first-order beliefs about the A players' honesty). The A players were asked to estimate both the B players' guesses about the proportion of honest announcements for each number of stars (*i.e.,* their second-order belief) and the proportion of honest announcements for each number of stars made by all A players in the session (*i.e.,* their first order belief). They earned a 0.50 Euro bonus if their answer in one (B players) or two (A players) randomly selected guesses was accurate within an interval of 5% point.

Finally, participants were told their earnings in each part and answered a questionnaire including socio-demographic information and questions on the motivation of choices in the experiment.

## 3.2 Treatments

Below we describe our six between-subject treatments. The full set of translated instructions are in Online Appendix B.

**No-Oath treatment:** In the No-Oath treatment, no oaths were taken by participants in the role of Player A.

**Public-Oath treatment:** In the Public-Oath treatment, participants in the role of Player A were required to swear the following oath: "I swear upon my honor that during this experiment I will behave honestly and I will always tell the truth."[10] The instructions for all participants stated that those who would be assigned the role of Player A would be asked to take an oath by which they would swear upon their honor that during the experiment they would behave honestly and always tell the truth. The oath-taking stage took place after participants discovered their role in the game. The A players swore the oath by typing in the text of the oath on their computers to facilitate its encoding in memory (see Naka and Naoi, 1995; Skinner et al., 1997).[11] The text for the oath and taking of the oath was

[10] As shown in Table A1 in Online Appendix A, the majority of previous experiments used voluntary oaths but in these studies, very few participants refused to sign it. We imposed that participants sign the oath to align ourselves with professional oaths that employees are typically required to sign when they are hired. In addition, it ensured that taking the oath was common knowledge. Participants were free to refuse participation in the experiment if they did not want to sign the oath; none did.

[11] The computer program checked that the words "I swear upon my honor", "honestly", and "truth" were correctly typed. We also checked the actual oaths typed by participants after the experiment. All were correctly typed, except for minor typos.

common knowledge. This treatment allows us to measure the overall impact of the oath on A players' likelihood of lying and on the frequency of detectable and deniable lies, as well as on B players' behavior.

**Private-Oath treatment:** In the Private-Oath treatment, the instructions handed out to all participants were identical to those of the No-Oath treatment. However, after discovering their role in the Announcement Game, all the participants in the role of Player A were asked to take the oath, while those in the role of Player B, who were not informed of the existence of the oath, were asked to type a neutral sentence during the oath-taking stage.[12] The A players were made aware that all participants in their role had to take the oath and were also informed of the asymmetry in information with respect to the B players. We compare behavior in the Private-Oath treatment with that in the No-Oath treatment to identify the importance of the self-image and passive audience-image costs of breaking an oath: since the instructions for the B players were identical across these two treatments, the impact on the reputation of the A players to the B players as it relates to breaking the oath is the same (and nil) in both those treatments, so that differences in behavior across these treatments can be assigned to the joint private and passive audience costs of breaking an oath.

**Private-Oath$^{++}$ treatment:** This treatment was designed to minimize the sense of being observed by a passive audience by making it clear to participants that the experimenter could not link their identities to their decisions. It was identical to the Private-Oath treatment except for the payment procedure, which was modified to further enhance anonymity. In the Private-Oath$^{++}$ treatment, participants were informed at the beginning of the session that at the end of this session, they would receive an envelope containing their payment in cash at their computer terminal. They learned that these envelopes (identified by the computer numbers) would be distributed by an independent person who would not be informed of the content of the experiment and would not have access to the experimental data. To preserve anonymity while maintaining accounting compliance, participants were informed that they would be asked to sign a "collective receipt", that is, a document including a table listing all payment amounts but without any corresponding computer numbers. Each participant would only be required to confirm that their payment amount appeared on the list without having to indicate which one it was. The experimenters did not participate in the payment procedure in any way, nor were they present once the experiment was launched. Participants knew they would not interact with the experimenter, and that it was impossible to link a

[12] This was done to avoid placing the B players in a situation in which they would wonder why other participants were typing while they were not.

payment or action to an identifiable person. This minimized the way participants might have felt observed or judged by a passive observer for breaking the oath.

**Necessary Condition treatment:** All participants in this treatment were assigned the role of Player B. They were told that the instructions they would receive would be an *excerpt* of the instructions that were given to participants from a past session. Those instructions were those from the Public-Oath treatment and were identical to them, except for the removal of the section on the oath.[13] After reading those instructions, participants learned that in each period, they were going to be randomly matched with A players from those past sessions and would face the announcement of the A player they would be matched with. In addition, their payment would be determined by Nature's draw for that particular A player in that period.[14,15] To ensure that participants had enough opportunity to learn from experience, we extended the number of periods in this treatment to 27; as in the other treatments, this number was unknown to the participants. Comparing investment decisions in this treatment to those in the Public-Oath treatment allows us to study whether the B players react mainly to the common-knowledge element of the oath or to their actual experience of the A players' honesty.

**Sufficient Condition treatment:** All participants in this treatment were assigned the role of Player B. The procedure of the Sufficient Condition treatment was similar to that of the Necessary Condition treatment, except that in part 2 the instructions explicitly integrated those used in the Public-Oath treatment including the part about the oath. Participants were then told that in each period they would be randomly matched with an A Player drawn from a *subset of A players* from these past sessions. Unbeknownst to the participants, this subset of A players was chosen such that the empirical distribution of announcements

[13] In the Public-Oath treatment, the reference to the honesty oath was placed at the very end of the instructions, and we simply removed that section in the Necessary Condition treatment. In our instructions, we were very explicit that the instructions they were to receive were an "*excerpt*" of the prior ones. On the reasons for not considering withholding information about treatment manipulations as deception, see Hey (1998); Hertwig and Ortmann (2008); Charness et al. (2022).

[14] In each period and for each B player independently, the program randomly selected one of the 1386 announcements made by the 77 A players who participated in the Public-Oath treatment and showed the B player that announcement. The card (star or blank) that determined the outcome of the investment in that period was the one actually drawn by Nature on that specific announcement in the past session. Thus, the Necessary Condition treatment preserved the same empirical distribution of the A players' announcements and Nature's draws as in the Public-Oath treatment. Moreover, due to a lack of time trends in the announcements by A players in the Public-Oath treatment, our randomization procedure does not create any period-related bias in this treatment. Using Kolmogorov-Smirnov test reveals no significant difference between the actual distribution of announcements in Necessary Condition treatment (*ex-post*) and that in Public-Oath treatment ($p = 1.000$).

[15] Naturally, one other difference is that the B players' decisions in this treatment impact the payoff of no one else. We acknowledge that participants may take others' payoffs into account. However, Tergiman and Villeval (2023) showed that with the parameters used, social preferences do not play a large role in the Announcement Game.

was as close as possible to that in the No-Oath treatment.[16] In other words, we created a situation where the B players in Sufficient Condition treatment faced the A players who took a public oath but behaved on average like those who did not take any oath. Comparing investment decisions in this treatment to those in the Public-Oath and No-Oath treatments allows us to examine whether the knowledge about the existence of the oath in itself is sufficient to persuade the B players to invest and whether this changes over time as they experience untruthful announcements.

## 3.3 Procedures

The experiment was run at GATE-Lab, Lyon, France. We ran a total of 36 sessions: 7 sessions with 152 participants in the No-Oath treatment (76 participants in each role), 7 sessions with 154 participants in the Public-Oath treatment (77 in each role), 7 sessions with 146 participants in the Private-Oath treatment (73 in each role), 7 sessions with 150 participants in the Private-Oath$^{++}$ treatment (75 in each role), 5 sessions with 77 participants in the Necessary Condition treatment (all B players), and 3 sessions with 78 participants in the Sufficient Condition treatment (all B players).[17] All 757 participants were recruited via HRoot (Bock et al., 2014). Participants were primarily students from local engineering and business schools.[18] The experiment was programmed using z-Tree (Fischbacher, 2007).

Upon arrival, participants randomly drew a ticket from an opaque bag which assigned them to terminals. Instructions were distributed at the beginning of the relevant part and read aloud. The average duration of a session was 80 minutes.

In all treatments, the instructions were delivered via audio recording. After privately answering any questions (which were rare and never related to the oath), the experimenter left the room. The computer program was then launched and monitored only by a technician. In all treatments except the Private-Oath$^{++}$ treatment, participants were informed in advance that their earnings would be paid by bank transfer rather than in cash. Payment information was entered by participants into a separate file that was sent directly to the institute's accountant and was never shared with the researchers.

Participants' earnings consisted of their payoffs from each part of the experiment: the payoff in the relevant scenario in the randomly chosen period (9 or 10) in the Truthful

[16] A Kolmogorov-Smirnov test reveals no significant difference between the distributions of announcements in Sufficient Condition treatment (*i.e.*, created by the subset of A players in Public-Oath) and that in the No-Oath treatment ($p = 1.000$).

[17] We based the number of observations on an ex-ante power calculation. Assuming a Type-I error rate of 0.05 and a medium effect size (Cohen's d = 0.60), targeting 77 observations per role and treatment would allow us to uncover the hypothesized effect of oath using a two-tailed Mann-Whitney ranksum test and achieve a power level of 0.95.

[18] See Tables C1 and C2 in Online Appendix C for summary statistics of the socio-demographic characteristics of the A and B players.

Announcement Game, the payoff from the randomly chosen round in the Announcement Game, and the payment from the belief elicitation. On average, participants earned 20 Euros (SD = 6.2), including a 7-Euro show-up fee.

# 4 Results

We begin by evaluating the overall impact of a public honesty oath on A players' behavior, considering both detectable lies, which happen when Player A sees fewer than three stars but announces three stars, and deniable lies, which happen when Player A announces two stars but sees fewer than two stars, or announces one star while seeing none.[19] Next, we explore why oaths lead to changes in behavior and document the importance of the non-pecuniary costs of breaking an oath. We then turn to the B players and explore whether change in investment behavior is due to the *knowledge* that an oath has been taken, or to the increase in honesty that the B players actually experience.

To analyze the data, and unless otherwise specified, we use two-tailed rank-sum tests and tests of proportions as our main statistical tool to evaluate differences across treatments, using one observation per subject.[20] We further support our claims via the use of econometric analyses, using mainly random-effects linear probability models with clustering at the individual level, all placed in Online Appendix C.

## 4.1 Impact of the public oath on lies

To evaluate the impact of the public oath on lies, we compare the No-Oath and the Public-Oath treatments.[21] Columns 1 and 2 of Table 1 display the frequency of detectable and deniable lies among the A players in the No-Oath and Public-Oath treatments.

Table 1 shows that the fraction of A players who always tell the truth increases significantly with the introduction of a public oath, going from 9.2% in the No-Oath treatment to 74.0% in the Public-Oath treatment ($p < 0.001$). Looking at the population of A players as a whole (top panel), it is clear that the reduction in lies after an oath is taken impacts both de-

[19] More precisely, regarding detectable lies, when fraudulently announcing three stars, the chances that the A player's lie is detected ranges from 100% if they actually saw zero stars, to 67% if they saw one star among the three cards, to 33% if they saw two stars among the three cards. Concerning deniable lies, Nature's draw will be consistent with the announcement, no matter whether the draw is of a blank card or of a card with a star, making the B player unable to detect the lie.

[20] We average choices within a subject and use this average choice as our unit of measure.

[21] Note that lying costs (independently of those incurred by breaking an oath) exist in all treatments. We assume here that lying costs and the costs of breaking an oath enter the utility in an additive and separable manner. If one is to assume a more complicated view that these costs interact and that the partials are non-zero, then our design can tease out the joint impact of lying and oath-breaking costs on the three non-pecuniary image costs that we consider.

tectable and deniable lies: the relative frequency of the former decreases from 25.2% to 3.9% ($p < 0.001$) and the frequency of the latter decreases from 57.4% to 13.8% ($p < 0.001$).[22]

*Table 1: Detectable and deniable lies in the No-Oath, Public-Oath, and Private-Oath treatments*

| Treatments | (1) No-Oath | (2) Public-Oath | (3) Private-Oath | (1-2) | (2-3) p-value | (1-3) |
|---|---|---|---|---|---|---|
| **All A players** | | | | | | |
| % Player A who always tell the truth | 9.2% | 74.0% | 39.7% | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| % Untrue announcements | 52.3% | 10.8% | 27.7% | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| *Relative frequency* | | | | | | |
| Detectable lies | 25.2% | 3.9% | 9.6% | $p < 0.001$ | $p = 0.002$ | $p < 0.001$ |
| Deniable lies | 57.4% | 13.8% | 37.9% | $p < 0.001$ | $p < 0.001$ | $p = 0.001$ |
| *Absolute frequency* | | | | | | |
| Detectable lies | 71.1% | 11.7% | 32.9% | $p < 0.001$ | $p = 0.002$ | $p < 0.001$ |
| Deniable lies | 86.8% | 23.4% | 56.2% | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ |
| **Among A players who lie at least once** | | | | | | |
| % Untrue announcements | 57.6% | 41.4% | 46.0% | $p = 0.010$ | $p = 0.397$ | $p = 0.004$ |
| *Relative frequency* | | | | | | |
| Detectable lies | 27.8% | 15.0% | 15.9% | $p = 0.007$ | $p = 0.593$ | $p = 0.003$ |
| Deniable lies | 63.2% | 53.2% | 62.8% | $p = 0.302$ | $p = 0.316$ | $p = 0.884$ |
| *Absolute frequency* | | | | | | |
| Detectable lies | 78.3% | 45.0% | 54.5% | $p = 0.004$ | $p = 0.479$ | $p = 0.008$ |
| Deniable lies | 95.7% | 90.0% | 93.2% | $p = 0.334$ | $p = 0.660$ | $p = 0.568$ |

*Notes*: Relative frequency statistics show the average frequency with which participants engage in any particular type of lie. Absolute frequency statistics show the proportion of A players who make any particular type of lie at least once. The relative frequency of detectable lies corresponds to how often A players who saw fewer than three stars announced three stars. The relative frequency of deniable lies corresponds to how frequently A players who saw fewer than two stars reported two stars or announced one star while seeing none. This table presents data both for all A players (top panel) and for the subset who make at least one lie (lower panel).

The same patterns arise if we look at the fraction of A players who engage in each type of lie at least once, as the "absolute frequency" data displayed in Table 1 shows.[23] In other words, the public oath has a large and statistically significant impact on lying, for both detectable and deniable lies. The result regarding the reduction of deniable lies is particularly striking, as those lies are precisely those that are impossible for the B players to detect.

[22]The Linear Probability Models reported in Tables C4 and C5 in Online Appendix C and estimating the probability of making a detectable and a deniable lie: a public oath causes a reduction in detectable and deniable lies by about 20 and 40 percentage points, respectively, compared to when there is no oath ($p < 0.001$ for both; model (1)).

[23]We note no differences in time trends across the two treatments: when comparing the difference-in-differences between the relative frequencies of detectable and deniable lies between the first and second half of the experiment across the two treatments, the p-values are 0.819 and 0.850, respectively. This is confirmed by the regressions reported in Tables C4 and C5, and by visual inspection of the period-by-period frequencies throughout the game, as displayed in Figures E1 and E2 in Online Appendix E.

The oath also has an impact on those subjects who continue to lie despite it. Focusing on the subset of A players who lie at least once when there is no oath (which represents 91.8% of A players) and those who break the public oath (26.0% of A players), the oath changes both the intensity and the nature of lies (lower panel of Table 1). While lies are still widespread among the A players who do lie at least once, the frequency of lies drops from 57.6% without the oath to 41.4% with the oath ($p = 0.010$). Interestingly, the difference in frequency only comes from detectable lies: we note a sharp reduction in the relative frequency of detectable lies (27.8% *vs.* 15%, $p = 0.007$), whereas deniable ones remain high and are no different across treatments for this subset of A players (63.2% *vs.* 53.2%, $p = 0.302$). The same conclusion holds if we turn our attention to absolute frequencies (see bottom of Table 1).[24] This suggests that for the 26% of A players who lie under oath, the impact of the oath is only partial and targeted to lies that would obviously make plain that the oath has been broken.

## 4.2 Non-pecuniary costs of breaking an oath

We now ask why the public oath has such a large impact on A players' behavior. To answer this question, we turn to the non-pecuniary costs of breaking an oath. Players A may incur three types of non-pecuniary costs when breaking an oath, each of these arising from which particular audience witnesses the oath. The first costs we consider, which we call *active*-audience image costs, are those that relate to how the B player may perceive the A player if they believe the oath has been broken. The B player is "active" in the sense that the B player makes decisions in the game. The second are *passive*-audience image costs, which come from A Players' concerns regarding how an outside observer, such as the experimenter, may view them. The experimenter (say) is a "passive" audience in that, as stated Abeler et al. (2019), "it takes no action, but rather serves as a player who may hold beliefs about any of the subjects after observing the subjects' reports". The third are *self*-audience image costs, which come from how A players may feel about themselves after breaking an oath, irrespective of whether anyone else witnessed the oath or their behavior, as in Dufwenberg and Dufwenberg (2018) and Gneezy et al. (2020).

### Active-audience image costs of breaking an oath

The impact of the active-audience costs of breaking an oath can be extracted by comparing the Public-Oath and Private-Oath treatments. Indeed, across these two treatments, both the

[24] Models (3) in Tables C4 and C5 show that compared to the No-Oath treatment, detectable lies in the Public-Oath treatment are lower by 12 percentage points ($p = 0.091$), but only by 9 percentage points for deniable lies (and not significantly different).

self and passive-audience image costs of breaking an oath are held constant: in both cases, the A players and the passive observers are aware of the oath, the true number of stars, and the actions of the A players. However, the *active*-audience costs are different across these two treatments: in the Public-Oath treatment, the B players are aware of the oath and so, can form beliefs about the A player's behavior under oath; in the Private-Oath treatment, the B players are unaware of the oath and so, their beliefs about the A player honoring the oath have no role to play in the A players' decisions.

Columns 2 and 3 of Table 1 summarize the behavior of A players across these two treatments. Looking at all A players (top panel) reveals that when the oath is no longer public, the fraction of truth-tellers drops by 35 percentage points, going from 74.0% to 39.7% ($p < 0.001$). In other words, active-audience image costs account for about 35 of the 70 percentage point increase in honest A players after the introduction of a public oath relative to a situation in which no oath is present.[25]

### Self- and passive-audience image costs

The comparison between the No-Oath and Private-Oath treatments sheds light on the costs of breaking an oath that exist independently of how the other active players in the game see them. Indeed, the B players in the No-Oath and Private-Oath treatments face the same instructions since in the latter, the oath remains private to the A players. As a result, the image of Player A vis-à-vis Player B after a particular announcement and a particular payoff-determining draw cannot be impacted by beliefs that would be induced by knowledge of the oath. In other words, in the Private-Oath treatment, there exist no image costs of breaking the private oath vis-à-vis Player B, and any observed increase in honest behavior in the Private-Oath treatment relative to the No-Oath treatment is thus likely due to the rise in the self- and/or passive-audience image costs of breaking the oath.

**Joint impact of self- and passive-audience image costs:** Before teasing apart the self- and passive-audience costs, we first establish their joint impact. Columns 1 and 3 of Table 1 show that self- and passive-audience image costs of breaking the oath are high enough to fully impact 30% of participants, as the fraction of A players who tell the truth goes from roughly 10% in the No-Oath treatment to about 40% in the Private-Oath treatment ($p < 0.001$). In other words, the costs of making an untruthful announcement are high enough to move 10%

[25] We also note, as seen in the bottom panel of Table 1, that while the public oath sharply decreases the fraction of A players who continue to lie compared with when the oath is privately taken, we find no statistical differences between the A players who continue to make detectable lies in the Public-Oath treatment and those who continue to do so in the Private-Oath treatment, indicating that for these subjects any active-audience image costs induced by the oath are not sufficiently high to modify any behavior.

of subjects to tell the truth, while the self- and passive-audience image costs of breaking an oath are such that 30% additional subjects move to telling the truth.[26]

**Self- or passive-audience:** The honesty of the A players in the Private-Oath treatment may be driven by image concerns related to self as an audience (how they see themselves for breaking an oath) and/or a passive audience (how the experimenter may think of them for breaking an oath). To assess the magnitude of passive-audience image costs associated with the experimenter's observation, we use the data from the Private-Oath$^{++}$ treatment, where we substantially reduced the possible perception of experimenter scrutiny by removing the identifiability of decisions in the experiment. Across the Private-Oath and Private-Oath$^{++}$ treatments, we observe negligible differences in A players' behavior, suggesting that passive-audience image costs only play a minor role. Specifically, truth-telling in the Private-Oath$^{++}$ treatment is virtually identical to that in the Private-Oath treatment in all dimensions – whether it be in terms of overall lying, the absolute or relative frequency and the types of lies (see Table C6 in the Online Appendix C). Therefore, our evidence points to a negligible impact of experimenter scrutiny on oath-takers' behavior in our setting, and that the treatment differences between No-Oath and Private-Oath treatments were primarily driven by self-audience image concerns associated with breaking an oath.[27]

This is further corroborated by the post-experimental questionnaire in the Private-Oath treatment, which focuses on the truth-tellers' stated motivations.[28] None mentioned the experimenter explicitly (as it relates to how an experimenter may judge them). Four truth-tellers stated motivation (about 17% of answers) did reference an external person: two described being truth-tellers because "I was asked to," where the reasonably natural completion of this passive voice would be "by the experimenter;" the other two described not wanting to corrupt the research question (assuming that the experimenter wanted them to tell the truth). The modal stated motivation (roughly 42% of answers) instead references placing weight on their honor or character as it relates to the oath, with narratives such as "I swore

[26]The regressions reported in Tables C4 and C5 confirm the significant differences in the likelihood of making detectable and deniable lies between the No-Oath and the Private-Oath treatments (models 1 and 2). Among those who lied at least once (models 3 and 4), the likelihood of making a detectable lie is significantly lower in the Private-Oath treatment than in the No-Oath treatment (Table C4), but there is no significant treatment difference for deniable lies (Table C5).

[27]While no prior work has examined whether image concerns vis-à-vis a passive audience affect the decision to break a promise to be honest, existing research has studied how passive audiences influence lying. In non-strategic settings like the die-rolling task, studies have found that the presence of an experimenter can affect honesty (Jiang, 2013; Fischbacher and Föllmi-Heusi, 2013; Fries et al., 2021; Abeler et al., 2019). However, in strategic contexts like ours, the evidence points in the opposite direction. For example, van de Ven and Villeval (2015) show in a deception game that individuals do not lie less when a passive observer is introduced—even when anonymity is removed. A similar result is reported in Lang and Schudy (2023). Together, these findings suggest that in strategic interactions, image costs from passive audiences may play only a limited role.

[28]Twenty-nine A players were truth-tellers in this treatment, and all but four answered this part of the questionnaire.

an oath not to [lie] and my honor is more important than money," "Breaking the oath wasn't worth the additional gain," "It is important to me to keep my word," and similar statements.

In summary, we separate three non-pecuniary costs of breaking an oath linked to how oath-takers care about their image vis-à-vis themselves, an active audience, and a passive audience. Our data shows that while self- and active-audience image costs largely drive oath-takers' behavior, costs associated with how they are perceived by a passive observer play a minor role. Our data point to self-audience image costs independent of any outside audience judgment as being the primary motive for truth-telling under a private oath, whereas passive-audience image costs associated with experimenter scrutiny have minimal impact on oath adherence.[29] Overall, only an oath taken with common knowledge engages individuals with all costs – in our data, a public oath leads close to 75% of oath-takers to tell the truth, compared with only 40% when the oath is taken privately, and 10% when there is no oath.

## 4.3 How do oaths affect B players?

### Impact on investment and earnings

We now turn to the behavior of B players and look at the impact of the public oath both on overall investment and on investment after two- or three-star announcements. Table 2 shows the average investment rates as well as how the B players react to various levels of announcements in both the Announcement Game and the Truthful Announcement Game. Comparing behavior across these two games and across treatments allows us to understand how credible various announcements are and how the public oath impacts credibility and investment behavior, accounting for a player's baseline investment rates in the absence of asymmetric information.[30]

There are significantly more investments after two- and three-star announcements under the public oath than without the oath.[31] The investment rate after a three-star announcement increases by almost 20 percentage points after the introduction of a public oath, going from 77.8% to 97.7% ($p < 0.001$). The corresponding statistics for investments after a two-star announcement are 70.5% and 85.8% ($p < 0.001$). The data also show that the B players

[29] While our Private-Oath$^{++}$ treatment minimizes the experimenter scrutiny, we acknowledge that the hypothesis that participants care about an outside observer's beliefs about them. It is impossible to design a treatment in which the oath would be *fully* unknowable to the experimenter (even probabilitistically) and in which one could control for judgment by an outside audience such as God.

[30] As a complement, Online Appendix D provides additional information on the B players' end-of-experiment first-order beliefs about the credibility of the announcements.

[31] The Linear Probability model of the likelihood of investing reported in Online Appendix Table C7 indicates an increase in investment by 17 percentage points in the Public-Oath treatment compared to No-Oath ($p < 0.001$; model (1)).

*Table 2: Investment rates in the Announcement and Truthful Announcement Games*

| | | No-Oath | Public-Oath |
|---|---|---|---|
| 0 stars announced | % Announcement Game | 9.5%[1] (21) | 1.7% (59) |
| | % Truthful Ann. Game | 0.0% | 2.6% |
| 1 star announced | % Announcement Game | 14.5% (54) | 16.8% (77) |
| | % Truthful Ann. Game | 19.7% | 16.9% |
| 2 stars announced | % Announcement Game | 70.5% (76) | 85.8% (77) |
| | % Truthful Ann. Game | 98.7% | 98.7% |
| 3 stars announced | % Announcement Game | 77.8% (76) | 97.7% (72) |
| | % Truthful Ann. Game | 100.0% | 98.7% |
| Average investment rate | | 65.8% | 56.9% |
| Average earnings of B players | | 149.7 | 159.9 |

*Notes*: The table displays the fraction of times participants invested in the Announcement and Truthful Announcement Games. Numbers of participants faced with a particular announcement at least once are in parentheses. [1] While this percentage may seem relatively high, this represents a small number of participants, as only 21 B players saw a zero-star announcement in this treatment, and of these two invested.

find announcements more credible in the Public-Oath treatment relative to the No-Oath treatment. Indeed, the difference in investment rates between the Truthful Announcement and Announcement Games is significantly smaller in the Public-Oath than in the No-Oath treatment for two and three stars ($p < 0.001$ in both cases), showing that the gap closes when the public oath is introduced and, in parallel, that the B players are more suspicious about the honesty of A players in the baseline than in the Public-Oath treatment. As investment rates rise sharply with the introduction of the public oath, B players move closer to the choices they would have made if the information were symmetric. When three stars are announced, the gap across the Truthful Announcement Game and the Announcement Game narrows to one percentage point (97.7% in the Announcement Game versus 98.7% in the Truthful Announcement Game).[32]

Finally, since there are fewer two- and three-star announcements under oath than in the baseline, overall average investment levels are slightly lower under oath, but the proportion of "good investments" compensates for this drop and leads to higher earnings for B players. Indeed, as announcements are more frequently truthful and credible, the B players avoid investing when the truth is zero stars, and, at the same time, are more likely to invest

[32]Comparing these investment rates with investment rates under the Truthful Announcement Game in the absence of any information (period 10 in part 1), we see that, in fact, under the No-Oath treatment two-star announcements are treated as uninformative, while three-star announcements do carry some information ($p = 0.550$ and $p < 0.001$, respectively – signrank tests), while both types of announcements under Public-Oath are treated as containing information ($p = 0.005$ and $p < 0.001$, respectively – signrank tests).

when the actual number of stars is two or three.[33] This leads to an increase in the average earnings of B players from 149.7 tokens without the oath to 159.9 tokens under a public oath ($p = 0.016$).[34]

### Why do B players invest more under oath?

We now ask what role knowledge of the oath plays in the behavior of B players: are the B players reacting to their experience with more truthful A players, or are they reacting to the knowledge of the oath *per se*? In other words, we ask whether knowledge of the oath is *necessary* for B players to trust the announcements and whether that knowledge is *sufficient* for B players to think the announcements are credible.

**A necessary condition?** To answer the first question, we turn to the Necessary Condition treatment. In this treatment, all participants are B players. These B players are faced with the announcements from the A players in the Public-Oath treatment (who behaved predominantly honestly), without being aware that those A players had sworn an oath. Columns (1) and (2) in Table 3 place side-by-side the behavior of B players in the Necessary Condition and Public-Oath treatments, focusing on the last nine periods of play to allow participants the opportunity to learn, particularly in the Necessary Condition treatment.[35]

Table 3 shows that average investment rates are no different across these two treatments (56.1% in Necessary Condition *vs.* 56.0% in Public-Oath, $p = 0.860$) (see also the non-significant Wald test at the bottom of Table C7 in Online Appendix C).[36] The gap in investment rates after zero-, one-, two- or three-star announcements across treatments is always small in magnitude and statistically not significant: investment choices in the Necessary Condition treatment are no different than those from the B players who were aware of the oath. In addition, while the investment rate decreases over time when there is no oath,

[33] The proportions of "good investments" (*i.e.,* made when the announcement is truthful) on two- and three-star announcements in the Public-Oath treatment are 83.3% and 83.8%, respectively. The corresponding figures in the No-Oath treatment are 50.2% and 33.2%, which are significantly lower ($p < 0.001$ for both).

[34] For Player A the lower investment rate under oath naturally leads to a significant drop in their average earnings from 161.6 tokens in No-Oath to 143.9 tokens in Public-Oath ($p < 0.001$), and 145.1 tokens in Private-Oath ($p = 0.005$), with no significant difference between the two oath treatments ($p = 0.608$).

[35] Recall that in the Public-Oath treatment, there are a total of 18 periods so that we focus on the last half of the game, and in the Necessary Condition treatment a total of 27 so that we focus on the last third. Our conclusions are robust to focusing on the second half of all periods in Necessary Condition, and are largely unchanged if we instead restrict the analysis to the first 18 periods to match the number of periods between the Public-Oath and Necessary Condition treatments (the only difference is that, while close in magnitude, there is a weak difference in investment rates across the two sets of data when three stars are announced: 92.1% *vs.* 97.7%, $p = 0.076$). Conclusions are similar if we instead look at all the periods in each treatment without any restriction.

[36] The fact that investment decisions are similar across the Public-Oath and Necessary Condition treatments, even though they only impact B players' payoff in the latter, indicates that in our setting social preferences may be outweighed by other considerations.

*Table 3: Investment rates in the Necessary Condition, Public-Oath, and Private-Oath treatments*

| | Necessary Cond. (1) | Public-Oath (2) | Private-Oath (3) | *p-value* *(1) vs. (2)* | *(1) vs. (3)* |
|---|---|---|---|---|---|
| 0 stars announced | 1.1% | 2.3% | 1.6% | $p = 0.966$ | $p = 1.000$ |
| 1 star announced | 11.3% | 15.9% | 10.5% | $p = 0.455$ | $p = 0.487$ |
| 2 stars announced | 84.7% | 82.8% | 70.4% | $p = 0.388$ | $p < 0.001$ |
| 3 stars announced | 95.2% | 97.2% | 88.8% | $p = 0.587$ | $p = 0.068$ |
| Average investment rate | 56.1% | 56.0% | 56.9% | $p = 0.860$ | $p = 0.513$ |

*Notes*: The table shows the fraction of times B players invested in the Announcement Game, looking at the last nine periods to allow participants to gain experience.

such a time trend is not observed in the Public-Oath or Necessary Condition treatments.[37]

These data have two implications. First, the increase in investment after two- and three-star announcements in the Necessary Condition treatment compared to the No-Oath treatment is robust to facing a small fraction of lies. Recall that just over a quarter of A players lie in the Public-Oath treatment despite having sworn not to. In other words, although the B players in the Necessary Condition treatment do face some dishonest behavior, they still invest at very high rates. Second, making the oath common knowledge is not a necessary condition for its impact to be effective on B players: trusting announcements can be learned even in a setting with asymmetric information and the possibility of deniable lies. This is further supported by the data from the Private-Oath treatment where B players are unaware of the oath (as is the case in the Necessary Condition treatment) but experience different levels of truthfulness (as the A players in Private-Oath are significantly less honest than those that B players face in the Necessary Condition treatment). The comparison of columns (1) and (3) in Table 3 shows that investment rates in the last nine periods in the Necessary Condition treatment are higher than in the Private-Oath treatment after two- and three-star announcements ($p < 0.001$ and $p = 0.068$, respectively).[38] In other words, knowledge about the oath in and of itself is not necessary for trust to develop.

[37] This can be seen from comparisons between investment rates in the first nine periods and the last nine (57.9% *vs.* 56% in Public-Oath; 55.4% *vs.* 56.1% in Necessary Condition). The difference is insignificant in the Public-Oath treatment ($p = 0.409$, signrank test), as well as in the Necessary Condition treatment ($p = 0.977$, signrank test), whereas there is a declining trend in the No-Oath treatment (69.2% *vs.* 62.4%, $p = 0.009$, signrank test). Table C8 in Online Appendix C reports random-effects regressions of the probability of investing on a time trend, with clustering at the individual level. The *p*-value of the period variable is equal to 0.843 in Public-Oath and 0.532 in Necessary Condition, whereas it is equal to 0.022 in No-Oath.

[38] Table C7 in Online Appendix C reports the estimates of random-effect Linear Probability models of the probability of investing, considering all periods and with the No-Oath treatment as the reference category. Whereas the coefficient of the Private-Oath treatment dummy is not significant, the coefficient of the Necessary Condition treatment dummy is significant (0.166, $p < 0.001$) and does not differ ($p = 0.894$) from the coefficient of the Public-Oath treatment dummy (0.171, $p < 0.001$).

**A sufficient condition?** If knowledge of the oath is not necessary, is it sufficient to convince B players to trust the announcements? To answer this question, we use data from the Sufficient Condition treatment. In this treatment, all participants were B players, were given the instructions of the Public-Oath treatment (including the oath), and were told that they would face announcements from a *subset* of A players in the Public-Oath treatment. The subset of A players was selected so that the distribution of announcements would closely resemble the distribution of announcements in the No-Oath treatment. Table 4 shows investment behavior in the last nine periods of play in the Sufficient Condition, Public-Oath, and No-Oath treatments.

*Table 4: Investment rates in the Sufficient Condition, Public-Oath and No-Oath treatments*

| | Sufficient Cond. (1) | Public-Oath (2) | No-Oath (3) | *p-value* *(1) vs. (2)* | *(1) vs. (3)* |
|---|---|---|---|---|---|
| 0 stars announced | 0.0% | 2.3% | 0.0% | $p = 1.000$ | $p = 1.000$ |
| 1 star announced | 6.6% | 15.9% | 14.8% | $p = 0.064$ | $p = 0.302$ |
| 2 stars announced | 64.1% | 82.8% | 65.9% | $p = 0.002$ | $p = 0.947$ |
| 3 stars announced | 80.2% | 97.2% | 76.1% | $p < 0.001$ | $p = 0.513$ |
| Average investment rate | 65.1% | 56.0% | 62.4% | $p = 0.010$ | $p = 0.430$ |

*Notes*: The table shows the fraction of times B players invested in the Announcement Game, looking at the last nine periods to allow participants to gain experience.

B players' behavior in the Sufficient Condition treatment differs significantly from that in the Public-Oath treatment but matches that of the No-Oath treatment. This is the case if we look at the overall average investment rates, but also if we look at investment rates after particular announcements (see the last two columns that show the $p$-values comparing treatments). These patterns show that knowledge of the oath is not sufficient to lead B to trust A. Instead, the B players are influenced by their own experience: when they know about the existence of an oath but realize that the A players have perjured themselves, they become less likely to invest: although the B players in the Sufficient Condition treatment are aware of the oath as those in the Public-Oath treatment, they behave as B players in the No-Oath treatment.[39]

Does knowledge that an oath has been taken then have no impact at all? What we have shown above is that if it does have an impact, it certainly is not persistent over time. However, it could be that knowledge of the oath *per se* changes initial attitudes and beliefs.

[39]The investment rate also tends to decrease over time in the Sufficient Condition treatment (from 77.2% in the first nine periods to 65.1% in the last nine periods, $p < 0.001$), as in the No-Oath treatment, and in contrast with the Public-Oath treatment where it is stable, as previously shown. In the random-effects Linear Probability models of the probability of investing reported in Table C8 in the Online Appendix, the $p$-value of the Period variable is <0.001 in Sufficient Condition, whereas it is equal to 0.022 in No-Oath and 0.843 in Public-Oath).

While we are very cautious in making claims regarding what may be going through our subjects' minds, our conjecture, and what our data suggest, is that initial behavior is not necessarily influenced by the knowledge that an oath has been taken. We support this claim by comparing first-period behavior, that is before any experience of play, across the Sufficient Condition and Necessary Condition treatments, which differ only in whether B players knew of the oath. We find no difference in how they react to announcements (the lowest $p$-value in all pairwise comparisons in a test of proportions for each type of announcement is 0.106). Had knowledge of the oath been a sufficient condition to believe it would lead to higher levels of truthfulness, then we would have seen higher investment rates for both two- and three-star announcements in the Sufficient Condition treatment.[40]

In summary, from Player B's standpoint, an oath is neither a sufficient nor a necessary condition for trust to be established, and the impact that we observe in the Public-Oath treatment stems from the difference in A's behavior that it results in, which in itself changes B's behavior.

# 5 Conclusion

In this paper, we have examined the role of distinct categories of image costs for oath-breaking in studying the efficacy of oaths on honest behavior. Our results highlight the importance of image costs related to self-perception and reputational costs stemming from how players perceive each other, while suggesting a lesser influence of image costs linked to passive observation. That passive observation does not seem to play a large role in our game stands in contrast to the literature on lying in individual decision-making tasks, but is aligned with lying in sender-receiver games, suggesting that in strategic settings, other image costs dominate.

Perhaps surprisingly, our results also reveal that despite significantly influencing the behavior of the oath-taker, the mere knowledge of an oath being taken is neither necessary nor sufficient for those interacting with the oath-taker to believe in its influence. Instead, those who interact with the oath-takers adopt a cautious "wait-and-see" approach, relying on experiential evidence.

While we remain cautious of the generalizability of our findings, we see relevance to real-world scenarios where oaths are being used. Our research suggests that to enhance the

[40]This is further supported by similar comparisons across the Private-Oath and Public-Oath treatments, as well as between the Public-Oath and Sufficient Condition treatments. In particular, there is no difference in how the B players react to two-star announcements in the first period, showing that experience is necessary for trust to be established.

effectiveness of an oath, one can leverage the psychological costs associated with violating it. Self-image costs may be influenced by how the oath is administered, such as through a *formal* ceremony, and their salience may be reactivated by periodic reminders. For example, Weitzel and Kirchler (2023) have shown that financial advisors in the Netherlands (who have to sign a compulsory oath when entering the profession) change their behavior when reminded about the oath, possibly indicating that self-image costs dissipate over time.

What we term "active-audience" image costs could be heightened with increased public awareness of the oath by people who would be interacting with the oath-takers. This could come from publicizing the oath, particularly in contexts where oaths may be relatively novel. This element of common knowledge is not trivial to establish. For instance, in countries like the Netherlands—where such oaths are mandatory in the financial sector—this aspect is often missing. To illustrate, we conducted a short online survey on Prolific with 198 Dutch residents. Only 53.5% were either certain or thought it likely that such an oath exists, despite ad campaigns that aim to increase the awareness of the oath. Further, among the 53.5% it's unclear whether these responses reflect actual awareness of the oath or simply an assumption that bankers must take one, given their professional responsibilities. To tease apart these interpretations, we ran a similar survey in France, where no such oath exists. There, 28% of 197 respondents believed bankers are required to swear an oath. This comparison suggests that a meaningful share of the Dutch responses may reflect generic assumptions rather than genuine knowledge of the oath's existence.[41] While general trust in bankers is low in the survey, there is, however, a positive link between being aware of the oath and trust in bankers (Spearman's correlation coefficient $= 0.15$, $p = 0.035$).

Our study establishes key distinctions in the types of costs associated with oath-breaking, and thus opens several avenues for future empirical and theoretical research. For instance, evaluating the effectiveness of oath swearing across different cultural contexts would be valuable, paralleling the global examination of honesty conducted by Cohn et al. (2019). Such cross-cultural investigations could reveal whether the costs associated with breaking an oath vary systematically across societies, potentially shaping the efficacy and applicability of oaths as moral commitments. Alternatively, a better understanding of the magnitude of these costs and how they evolve over time may help predict when—and for whom—they outweigh the benefits of breaking an oath.

In the theoretical domain, developing more precise models and better understanding the structure of the image costs of breaking an oath would be valuable. While our experiment was not designed to distinguish between alternative modeling approaches, our data nonetheless provides some insight into their possible form (see Online Appendix G for more details).

[41] See Online Appendix F for the instructions, detailed results of the surveys, and related discussion.

For example, data patterns we observe suggest that breaking the oath, even when it remains private information, may involve a cost that discourages maximal lying after promising not to lie at all. This type of cost differs both from a simple fixed cost that would be independent of the kind of lie that is made and from a cost that would depend on the size of the lie and that has been studied in the absence of an oath (see Gneezy et al., 2018). Additionally, the oath does not more strongly reduce "highly detectable" lies relative to "less detectable" ones, despite the former being associated with greater expected gains. This suggests that the oath's effectiveness may not hinge on fear of being seen breaking it (since the probability of detection is higher for the former than for the latter), while it is consistent with the willingness of not being seen as someone for whom it is easy to break a promise. Modeling the costs of breaking a promise opens up a broad space of theoretical possibilities. Our findings may help motivate future efforts to formalize the mechanisms behind oath-keeping and deepen our understanding of promise-related behavior.

# References


Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for Truth-Telling. *Econometrica 87*(4), 1115–1153.

Beck, T. (2021). How the honesty oath works: Quick, intuitive truth telling under oath. *Journal of Behavioral and Experimental Economics 94*, 101728.

Bock, O., I. Baetge, and A. Nicklisch (2014). hroot: Hamburg Registration and Organization Online Tool. *European Economic Review 71*, 117–120.

Casella, A., N. Kartik, L. Sanchez, and S. Turban (2018). Communication in context: Interpreting promises in an experiment on competition and trust. *Proceedings of the National Academy of Sciences of the USA 115*(5), 933–938.

Charness, G. and M. Dufwenberg (2006). Promises and partnership. *Econometrica 74*(6), 1579–1601.

Charness, G., A. Samek, and J. van de Ven (2022). What is considered deception in experimental economics? *Experimental Economics 25*, 385–412.

Cohn, A., M. A. Maréchal, D. Tannenbaum, and C. L. Zünd (2019). Civic honesty around the globe. *Science*, eaau8712.

de Bruin, B. (2016). Pledging integrity: Oaths as forms of business ethics management. *Journal of Business Ethics 136*(1), 23–42.

Di Bartolomeo, G., M. Dufwenberg, S. Papa, and F. Passarelli (2019). Promises, expectations and causation. *Games and Economic Behavior 113*, 137—-146.

Dufwenberg, M. and M. A. Dufwenberg (2018). Lies in disguise – a theoretical analysis of cheating. *Journal of Economic Theory 175*, 248 – 264.

Ederer, F. and F. Schneider (2022). Trust and promise over time. *American Economic Journal: Microeconomics 14*(3), 304–320.

Ederer, F. and A. Stremitzer (2017). Promises and expectations. *Games and Economic Behavior 106*, 161—-178.

Ellingsen, T. and M. Johannesson (2004). Promises, threats and fairness. *The Economic Journal 114*(495), 397–420.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Fischbacher, U. and F. Föllmi-Heusi (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association 11*(3), 525–547.

Fries, T., U. Gneezy, A. Kajackaite, and D. Parra (2021). Observability and lying. *Journal of Economic Behavior & Organization 189*, 132–149.

Gneezy, U., A. Kajackaite, and J. Sobel (2018). Lying Aversion and the Size of the Lie. *American Economic Review 108*(2), 419–453.

Gneezy, U., S. Saccardo, M. Serra-Garcia, and R. van Veldhuizen (2020). Bribing the Self. *Games and Economic Behavior 120*, 311–324.

Hertwig, R. and A. Ortmann (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior 18*(1), 59–92.

Hey, J. (1998). Experimental economics and deception: A comment. *Journal of Economic Psychology 19*(3), 397–401.

Ismayilov, H. and J. Potters (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics 19*(2), 382—-393.

Jacquemet, N., S. Luchini, J. F. Shogren, and A. Zylbersztejn (2024). Commitment to the truth creates trust in market exchange: Experimental evidence. *Games and Economic Behavior 148*, 279–295.

Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior & Organization 93*, 328–336.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies 76*(4), 1359–1395.

Khalmetski, K. and D. Sliwka (2019). Disguising lies—image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics 11*(4), 79–110.

Lang, M. and S. Schudy (2023). (Dis)honesty and the value of transparency for campaign promises. *European Economic Review 159*, 104560.

Naka, M. and H. Naoi (1995). The effect of repeated writing on memory. *Memory & Cognition 23*(2), 201–212.

Saccardo, S. and M. Serra-Garcia (2023). Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment. *American Economic Review 113*(2), 396–429.

Schudy, S., S. Grundmann, and L. Spantig (2024). Individual preferences for truth-telling. *CESifo Working Paper Series 11521*.

Sengupta, A. and C. Vanberg (2023). Promise keeping and reliance damage. *European Economic Review 152*, 104344.

Serra Garcia, M., E. E. C. van Damme, and J. J. M. Potters (2013). Lying about what you know or about what you do? *Journal of the European Economic Association 11*(5), 1204–1229.

Skinner, C. H., T. McLaughlin, and P. Logan (1997). Cover, copy, and compare: A self-managed academic intervention effective across skills, students, and settings. *Journal of Behavioral Education 7*(3), 295–306.

Tergiman, C. and M. C. Villeval (2023). The Way People Lie in Markets: Detectable Vs. Deniable Lies. *Management Science 69*(6), 3340–3357.

van de Ven, J. and M. C. Villeval (2015). Dishonesty under scrutiny. *Journal of the Economic Science Association 1*, 86–99.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations. *Econometrica 76*(6), 1467–1480.

Weitzel, U. and M. Kirchler (2023). The Banker's Oath And Financial Advice. *Journal of Banking & Finance 148*, 106750.

Zickfeld, J. H., S. T. S. Karg, S. S. Engen, A. S. R. Gonzalez, J. Michael, and P. Mitkidis (2024). Committed (dis)honesty: A systematic meta-analytic review of the divergent effects of social commitment to individuals or honesty oaths on dishonest behavior. *Psychological Bulletin 150*(5), 586–620.

# ONLINE APPENDIX

## A Online Appendix: Literature

*Table A1: Overview of previous experimental studies on oath-taking on various behaviors*

| Reference | Task | Common knowledge | Compulsory | Findings |
|---|---|---|---|---|
| *Strategic settings* | | | | |
| Jacquemet et al. (2018) | Coordination game | No | No | Positive impact on coordination rates. |
| Jacquemet et al. (2019) | Deception game | No | No | For selfish lies in loaded frame, the fraction of liars reduced from 35% in No Oath to 16.7% in (voluntary) Oath. Neutral frame does not reduce lies. |
| Beck (2021) | Deception game | Yes | Yes | Fraction of liars reduced from 42% in No Oath to 20% in (compulsory) Oath. No impact on the size of lies. |
| Jacquemet et al. (2021) | Deception game | No | No | For selfish lies, fraction of liars goes from 32.8% in No Oath to 14.3% in (voluntary) Oath. No impact on Pareto lies. |
| Koessler et al. (2021) | Public good game | Yes | Both | Positive impact of both mandatory and voluntary oaths on contributions, but the effect of the voluntary oath fades away more quickly than when it is mandatory. |
| Davis and Jaber-Lopez (2022) | Binary social dilemma game | No | Both | Impact of both mandatory and voluntary oaths, but crowding-out effect by non oath takers when it is voluntary. |
| Hergueux et al. (2022) | Public good game | No | No | Positive impact on contributions, according to social types. |
| Koessler (2022) | Public good game | Yes | No | Positive impact on contributions. More pledges when the majority of the group does so. |
| Weitzel and Kirchler (2023) | Audit study | Yes | Yes | Positive impact of reminders of the bankers' oath on real financial advisors' loan recommendations. |
| Jacquemet et al. (2024) | Trust game | No | No | Oath increases cooperation only with preplay communication. The effect of oath is equivalent to that of a mild fine. |
| *Individual decision-making* | | | | |
| Carlsson et al. (2013) | Contingent valuation | - | No | Decrease of the shares of null and very high willingness-to-pay. |
| Jacquemet et al. (2013) | Second-price auctions | - | No | Decrease of the shares of null and very high willingness-to-pay. |
| Krüger (2016) | Coin cheating task | - | Yes | Both positive and negative effects depending on gender |
| Jacquemet et al. (2017) | Voting referenda | - | No | Decrease of the hypothetical bias. |
| Heinicke et al. (2019) | Number cheating task | - | Yes | Negative impact on extreme lies, no impact on partial lies, in both gain and loss frames. |
| Koessler et al. (2019) | Tax evasion game | - | No | No impact on compliance except when combined with non-financial rewards. |
| Schild et al. (2019) | Mind game | - | Yes | Negative impact when lies are unobservable, no impact when lies are observable. |
| Beck et al. (2020) | Die cheating task | - | Yes | Negative impact on both extreme and partial lies. |
| Jacquemet et al. (2020) | Tax evasion game | - | No | No impact on extreme lies, negative impact on partial lies. |
| Kemper et al. (2020) | Discrete choices | - | Yes | Decrease of hypothetical bias in the estimation of willingness-to-pay. |
| Prima et al. (2020) | Asset reporting | - | Yes | No impact on lies. |
| Jacquemet et al. (2021) | Coin cheating task | - | No | Negative impact on extreme lies, no impact on partial lies. |
| Peer and Feldman (2021) | Reporting perform. | - | Yes | Long-term negative impact on lies. |
| Akin (2022) | Coin cheating task | - | Yes | Negative effects on undeserved applications. |
| Babin et al. (2022) | Reporting eye color | - | No | Negative impact on extreme lies, no impact on partial lies. |
| Babin and Chauhan (2023) | Coin cheating task | - | No | Negative impact on both plausible and implausible misreporting. |
| Cagala et al. (2024) | Exam-taking | - | Yes | Cheating doubles. |
| Cagala et al. (2024) | Chip cheating task | - | Yes | No impact of commitment requests on lies in both offline and online settings. |
| Zickfeld et al. (2025) | Tax evasion game | - | No | Positive impact on tax compliance depending on context. |

# B Online Appendix: Instructions [Translated from French]

### NO-OATH, PRIVATE-OATH, AND PRIVATE-OATH$^{++}$ TREATMENTS

*[The printed instructions in Private-Oath treatment and Private-Oath$^{++}$ treatment are the same as in the No-Oath treatment. However, the A players in the Private-Oath treatments discovered about the oath on their computer screens at the beginning of part 2. During the oath-taking stage, A players had to type a sentence by which they swore on their honour that, during this experiment, they would be honest and always tell the truth. Meanwhile, the B players were asked to type a neutral sentence: "The Nobel Prize in Economics for the year 2021 has been awarded to three American economists."]*

**Preliminary instructions** *[Only for the Private-Oath$^{++}$ treatment]*

We are testing a new payment method for participants in experiments at GATE-Lab.

At the end of the session, you will receive an envelope containing your earnings in cash. This envelope will be given to you by an independent person who will never be informed of the content of the experiment, will not know the contents of your envelope, and will not have access to the experimental data.

After taking your cash from the envelope, you can leave the experimental room one by one. Please leave the instructions, the computer tag, and the envelope at your seat.

Before you leave GATE-Lab, you will see a table in the reception room showing only the earnings distributed during the session, presented in random order (i.e., with no mention of the computer numbers or other information). All you will have to do is sign a statement that one of the amounts shown on this list corresponds to the earnings you have received, but you do not have to indicate which one. While guaranteeing compliance with accounting rules, this procedure ensures that no one (neither the experimenters nor the person handing out the envelopes) will ever be able to link your decisions during the experiment, your earnings and your identity.

**Welcome** *[Common to all treatments]*

Hello and welcome to an experiment on decision-making. Please turn off your phone and put it away. You are not allowed to communicate with other participants during the whole experiment, otherwise you will be excluded from the session and any potential earnings.

During this session you can earn money. The amount of money you will earn depends on your decisions and the decisions of other participants in the session. Please read the instructions carefully. All your decisions and responses will be anonymous.

This session consists of two successive parts. The amount you earn at the end of this session is the sum of your earnings in the different parts plus your participation fee of 5 Euros. During the session we will not talk in Euros but in tokens. The conversion rate from tokens to Euros is as follows:

100 tokens = 4 Euros

You will be paid via bank transfer by the CNRS. To do this, at the end of the session we will ask for your IBAN number. We will also ask you to send us a bank statement (containing the IBAN) in pdf format with the title "your first name-your last name" to the following e-mail address: gatelab[at]gate.cnrs.fr. Due to administrative and banking delays, the transfer may take up to approximately two weeks. We promise that your bank details will only be used to make the transfer. This information will be kept separate from the files containing your decisions and the

researchers who process your decisions will not have access to your bank details. *[This paragraph was omitted in the Private-Oath$^{++}$ treatment.]*

You will now find out the instructions for Part 1. The instructions for Part 2 will be explained at the end of Part 1.

**Part 1** *[Common to all treatments]*

This part consists of 10 periods. The first 8 periods are training rounds and nothing you decide during these 8 periods counts towards determining your actual earnings. For these 8 training periods the earnings stated are therefore hypothetical.
The only periods that can count towards your earnings in this part are the 9th and 10th periods. At the end of the session, the program will draw period 9 or period 10 and the earnings in the drawn period will constitute your earnings for this part. Each of these two periods has 50 chances out of 100 to be drawn.

**Description of the task**
In each period, you receive an initial endowment of 100 tokens and you must decide whether to keep these tokens as your earnings for the period or to invest them all in an investment project. If this project is successful, you earn 3 times the number of tokens invested, that is 300 tokens. If the project is not successful, you earn 30 tokens.

**Description of the investment project**
In each period, 3 cards appear on your screen, face down. Each card can have a star ($\star$) or be blank. Each card has 50 chances out of 100 to have a star and 50 chances out of 100 to be blank. These chances are independent for each card.
Thus, your three cards can have a total of zero stars (which happens with 12.5 chances out of 100), a total of 1 star (which happens with 37.5 chances out of 100), a total of 2 stars (which happens with 37.5 chances out of 100), or a total of 3 stars (which happens with 12.5 chances out of 100).

You must press the "Reveal" button to reveal your three cards. In each case, the cards appear in the following format:

After revealing the cards, you must choose whether or not to invest your 100 tokens.
After your choice, the program draws one of your three cards, each card having the same chance of being drawn (so each card has 1 chance out of 3 to be selected).

- If the card drawn has a star and you have invested, the project is a success and you earn 300 tokens (*i.e.,* endowment of 100 – investment of 100 + earnings of 300).
- If the card drawn is blank and you have invested, the project is not a success and you earn 30 tokens (*i.e.,* endowment of 100 – investment of 100 + earnings of 30).
- If you did not invest your tokens, you keep your initial endowment of 100 tokens and therefore earn 100 tokens.

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program from the three cards has a star or not.

Remember that in these 8 training periods, these earnings are hypothetical and nothing you decide in these periods counts towards determining your actual earnings.

To make sure that you have encountered all possible cases, we have in advance chosen cases that correspond to the probabilities announced in the task description.

**Period 9**
If drawn at the end of the session, this period determines your actual earnings for this part. The rules and task are the same as in the previous 8 periods. The only difference is in the way you must make your investment choice.

Your screen will show 3 cards, face down. Each card has 50 chances out of 100 to have a star. In this period, you have to make a decision in each of the following 4 scenarios. Would you invest in the project or not if the program announced that among the three cards there are :

Scenario a) 0 stars?

Scenario b) 1 star?

Scenario c) 2 stars?

Scenario d) 3 stars?

Once you have answered these questions, the program will inform you of the total number of stars among your three cards.

**Your earnings**
If this period 9 is drawn, your earnings are determined by the answer to the scenario that applies. That is, the one that corresponds to the total number of stars among your three cards. For example, suppose that the three cards hide a total of two stars; in this case, your decision in scenario (c) applies. Another example, suppose the three cards hide a total of three stars; in this case, your decision in scenario (d) applies.

The program then draws one of your three cards at random.

- If you have invested in the project and the card drawn has a star, then you earn 300 tokens (endowment of 100 - investment of 100 + earnings of 300).
- If you have invested in the project and the card drawn is blank, then you win 30 tokens (endowment of 100 - investment of 100 + earnings of 30).
- If you have not invested, you earn the 100 tokens of your initial endowment.

As you can see, the principle is the same as in the 8 training periods, but here you make a decision in each possible scenario.
Since only one of your responses will count towards your earnings if this period is drawn, when you make your decision in each scenario it is in your best interest to treat each scenario as if it were the one that actually counts towards your earnings for this part.

**Period 10**
If drawn at the end of the session, this period determines your actual earnings for this part. As in the previous periods your screen will show 3 cards, face down. Each card has 50 chances out of 100 to have a star. You must again decide whether or not to invest in the project. However, unlike the previous periods, you only have to make one decision without being informed of the number of cards with a star. Only at the end of the session will you be informed by the programme of the total number of stars among your three cards if this period is selected for payment.
Once you have made your decision, the program will then draw one of the three cards at random.

**Your earnings**
If period 10 is drawn for payment, your earnings are determined as follows:

- If you have invested and the card drawn has a star, then you earn 300 tokens.
- If you have invested and the card drawn is blank, then you earn 30 tokens.
- If you did not invest, you earn the 100 tokens from your initial endowment.

You will be informed of the period drawn (9 or 10), the card drawn and your earnings in that round at the end of the session.

Please read these instructions again. If you have any questions, please raise your hand or press the red button on the side of your desk. We will come and answer your questions in private immediately.

**Part 2** *(Instructions distributed after completion of part 1)*

In this part, each of you will be given a role, either 'A' or 'B'. Half of the participants have a role of A and the other half have a role of B. Your role remains the same for the entirety of part 2: you will never change it.
Part 2 has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.

In each period, each of you is paired such that there is one participant A and one participant B in each pair. You will never know the identity of the participant you are paired with. At the beginning of each period, you are randomly re-matched with a new participant. It is unlikely that you will be paired with the same participant two periods in a row.

**Your task in each period**
Participant A: Participant A sees three cards on his/her screen, face down. Each card can have a star ($\star$) or be blank. Each card has 50 chances out of 100 to have a star and 50 chances out of 100 to be blank. These chances are independent for each card. Thus, participant A can have *a total of* 0 stars, 1 star, 2 stars or 3 stars. Participant A can see how many stars he/she has by pressing the "Reveal" button.
Participant A's task is to then announce his/her total number of stars to participant B with whom he/she is matched with. Participant B cannot see Participant A's cards at any time.

Participant B: Participant B receives an initial endowment of 100 tokens and he/she is only informed of the announcement of Participant A with whom he/she is paired about his/her number of stars. Participant B has to decide whether he/she wants to keep his/her tokens or invest them in Participant A's project.

**Determination of earnings**

If you are Participant A: You earn a fixed amount of 30 tokens, plus 200 tokens if Participant B has invested in your project.
If you are Participant B: Once you have made your investment decision, the program draws one of Participant A's three cards at random.

- If you have invested in the project and the card drawn has a star, the project is a success and you earn 300 tokens (*i.e.,* endowment of 100 - investment of 100 + earnings of 300).
- If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens (*i.e.,* endowment of 100 - investment of 100 + earnings of 30).
- If you did not invest, you earn the 100 tokens of your initial endowment.

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.

**Your screen**
At the end of each period, your screen will show a history of what happened in the previous periods. Specifically, you will see four types of information:

1. your announcements in previous periods, if you are Participant A; or the announcements of the different Participants A you were paired with, if you are Participant B;
2. whether the card drawn in previous periods had a star or not, regardless of your role and decision;
3. whether you invested in the previous periods, if you are Participant B; or whether the Participants B you were matched with invested or not, if you are Participant A.
4. your potential earnings in each previous period.

For each of you, **only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part**. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.
Remember that **the pairs are re-matched in each period**.

[***ADDITIONAL PARAGRAPH IN PUBLIC-OATH TREATMENT ONLY***]

**Oath**
At the beginning of part 2, **Participants A will have to take an oath about the truthfulness of their announcements in part 2**. By this oath, they will swear on their honour that during this experiment they will be honest and always tell the truth.

*[After the instructions and role assignment, the A players in Public-Oath treatment swore the oath by typing it on their computer. The B players did not type anything during this stage. After all A players took the oath, every participant was informed that the oath had been taken.]*

- - -

Please read these instructions again. If you have any questions, please raise your hand or press the red button. We will come and answer your questions in private immediately.

- - -

**End of the instructions in part 2**

*[At the end of part 2 but before informing participants of their final gains, we elicited beliefs (displayed on computer screen only), depending on their role in the Announcement Game. For the B players, we asked the following question about their beliefs concerning the honesty of the A players in their session.]*
Please estimate below the percentage of true announcements among all the announcements you received in part 2. Enter integer values between 0 and 100. At the end of the session the program will draw one of these questions. If your answer is correct, plus or minus 5%, you will earn an extra 0.50€.

In your opinion, what percentage of announcements were true when Participant A announced:

- 0 stars: _%
- 1 star: _%
- 2 stars:_%
- 3 stars:_%

*[For the A players, we asked two questions (in randomized order at an individual level). One elicited their beliefs about the honesty of the A players in their session, which reads as follows:]*
Please estimate below the percentage of true announcements among all announcements made in part 2 by all participants A in this session. Enter integer values between 0 and 100. At the end of the session the program will draw one of these questions. If your answer is correct, plus or minus 5%, you will earn an extra 0.50€.

In your opinion, what percentage of announcements were true when participants A in this session announced:

- 0 stars: _%
- 1 star: _%
- 2 stars:_%
- 3 stars:_%

*[The other question asked about their beliefs about what the B players think about their honesty, which reads as follows:]*
Your participant B is answering the following question: "In your opinion, what percentage of announcements were true when participant A told you?".
What do you think your participant B's answer was to these questions?
Enter integer values between 0 and 100. At the end of the session the program will draw one of these questions. If your answer is correct, plus or minus 5%, you win an extra 0.50€.

- 0 stars: _%
- 1 star: _%
- 2 stars:_%
- 3 stars:_%

*[After the belief elicitations, participants were informed about their final gains (flat payment and additional payoffs from part 1, part 2, belief questions). They then completed a final questionnaire consisting of the following information:]*

- Age
- Gender
- Status
- School
- *[Player A only]* In part 2, if sometimes you did not announce the correct number of stars or if you announce the correct number of stars all the time, can you explain why you did so? If you have different reasons, can you rank them?
- *[Player B only]* In part 2, if sometimes you have invested or not in Participant A's project, can you explain why? If you have different reasons, can you rank them.
- What do you think this experiment is testing?
- If you had to leave a piece of advice for a successor in your role, what would it be?

- - -

## NECESSARY CONDITION TREATMENT

*[The written instructions for part 1 of the experiment are the same as in the other treatments. The following instruction was given to participants for part 2.]*

### Part 2

Some weeks ago, participants from the same subjects pool as you took part in an experimental session. These participants performed the same task as you in part 1, under exactly the same conditions as you did. Please read below an extract from the instructions they were given for part 2 (text in italics). The supplementary instructions that only concern you will be given after.

***Extract from the instructions given to past participants***

*In this part, each of you will be given a role, either 'A' or 'B'. Half of the participants have a role of A and the other half have a role of B. Your role remains the same for the entirety of part 2: you will never change it.*
*Part 2 has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.*

*In each period, each of you is paired such that there is one participant A and one participant B in each pair. You will never know the identity of the participant you are paired with. At the beginning of each period, you are randomly re-matched with a new participant. It is unlikely that you will be paired with the same participant two periods in a row.*

***Your task in each period***
*Participant A: Participant A sees three cards on his/her screen, face down. Each card can have a star (⋆) or be blank. Each card has 50 chances out of 100 to have a star and 50 chances out of 100 to be blank. These chances are independent for each card. Thus, participant A can have* a total of *0 stars, 1 star, 2 stars or 3 stars. Participant A can see how many stars he/she has by pressing the "Reveal" button.*

*Participant A's task is to then announce his/her total number of stars to participant B with whom he/she is matched with. Participant B cannot see Participant A's cards at any time.*

*Participant B: Participant B receives an initial endowment of 100 tokens and he/she is only informed of the announcement of Participant A with whom he/she is paired about his/her number of stars. Participant B has to decide whether he/she wants to keep his/her tokens or invest them in Participant A's project.*

***Determination of earnings***
*If you are Participant A: You earn a fixed amount of 30 tokens, plus 200 tokens if Participant B has invested in your project.*

*If you are Participant B: Once you have made your investment decision, the program draws one of Participant A's three cards at random.*

- *If you have invested in the project and the card drawn has a star, the project is a success and you earn 300 tokens (i.e., endowment of 100 - investment of 100 + earnings of 300).*
- *If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens (i.e., endowment of 100 - investment of 100 + earnings of 30).*

- *If you did not invest, you earn the 100 tokens of your initial endowment.*

*Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.*

***Your screen***
*At the end of each period, your screen will show a history of what happened in the previous periods. Specifically, you will see four types of information:*

1. *Your announcements in previous periods, if you are Participant A; or the announcements of the different Participants A you were paired with, if you are Participant B;*
2. *Whether the card drawn in previous periods had a star or not, regardless of your role and decision;*
3. *Whether you invested in the previous periods, if you are Participant B; or whether the Participants B you were matched with invested or not, if you are Participant A.*
4. *Your potential earnings in each previous period.*

*For each of you,* ***only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part****. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.*

*Remember that* ***the pairs are re-matched in each period****.*

***End of the extract***

**Supplementary instructions for participants in today's session**

The previous instructions apply to you in this part as well but there are some differences from the previous extract:

1. **All participants in today's session are assigned to <u>the role B</u>**; you will keep this role throughout the part.
2. Participants in **role A are not present in this session**: they have already participated in a previous session with other participants in role B.
3. At the beginning of each period, you will be matched with a participant A from a previous session randomly selected by the program.
4. Before you decide whether to invest, you will be informed of the announcement made by this participant A from a previous session after the initial draw of his/her three cards.
5. After deciding whether or not to invest, you will be informed of the card that was randomly drawn by the program from the three cards of this participant A from a previous session. This card drawn will determine your earnings for the period if you have decided to invest.
6. Your investment decisions will only affect your earnings and not the earnings of the participants A from previous sessions with whom you are matched.

**Summary**
This part has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.

In each period you will be informed about the number of stars (0, 1, 2 or 3) announced by a participant A from a previous session in a period drawn at random. You will then have to decide whether or not to invest in the project of participant A from the previous session.

- If you have invested in the project and the card drawn is a star, the project is a success and you earn 300 tokens.
- If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens.
- If you have not invested, you earn the 100 tokens of your initial endowment.

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.

For each of you, **only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part**. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.

Remember that **in each period the program randomly selects a decision** (*i.e.,* an announcement made after an initial draw of three cards) **from all the decisions made by participants A in previous sessions. You are re-matched with a new participant A in each period** and it is very unlikely that the program will select the same participant A for you two periods in a row.

- - -

Please read these instructions again. If you have any questions, please raise your hand or press the red button. We will come and answer your questions immediately in private.

- - -

**End of the instructions in part 2**

*[In the Necessary Condition treatment, at the end of part 2 but before informing participants of their final gains, we elicited beliefs (displayed on computer screen only), depending on their role in the Announcement Game like in the other treatments. However, after the belief elicitation about the proportion of honest announcements, participants in this treatment, who all had the role of B, were asked to compare their responses to their initial expectations about Player A's honesty before starting the Announcement Game.]*

We remind you below of your estimates of the proportions of true announcements received in part 2. Please let us know how well these estimates match your initial expectations before you began part 2. Specifically, indicate for each scenario whether your estimates are:

- More or less equal to your initial expectations

- Higher than your initial expectations (before the part started, you expected a lower proportion of true announcements)
- Lower than your initial expectations (before the part started, you expected a higher proportion of true announcements)

Your estimate of the proportion of truthful announcements are ...

- 0/1/2/3 stars (corresponding estimate shown here):
  - More or less equal your initial expectations
  - Higher than your initial expectations ...
  - Lower than your initial expectations ...

- - -

**SUFFICIENT CONDITION TREATMENT**

*[The written instructions for part 1 of the experiment are the same as in the other treatments. In part 2, like in the Necessary Condition treatment, participants were first given the extract of the instructions used in Public-Oath treatment. Note that the extract part for the Sufficient Condition treatment included information about the oath, as in the Public-Oath treatment.]*

**Part 2**

Some weeks ago, participants from the same subjects pool as you took part in an experimental session. These participants performed the same task as you in part 1, under exactly the same conditions as you did. Please read below an extract from the instructions they were given for part 2 (text in italics). The supplementary instructions that only concern you will be given after.

***Extract from the instructions given to past participants***

*In this part, each of you will be given a role, either 'A' or 'B'. Half of the participants have a role of A and the other half have a role of B. Your role remains the same for the entirety of part 2: you will never change it.*
*Part 2 has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.*

*In each period, each of you is paired such that there is one participant A and one participant B in each pair. You will never know the identity of the participant you are paired with. At the beginning of each period, you are randomly re-matched with a new participant. It is unlikely that you will be paired with the same participant two periods in a row.*

***Your task in each period***
*Participant A: Participant A sees three cards on his/her screen, face down. Each card can have a star (⋆) or be blank. Each card has 50 chances out of 100 to have a star and 50 chances out of 100 to be blank. These chances are independent for each card. Thus, participant A can have* a total of *0 stars, 1 star, 2 stars or 3 stars. Participant A can see how many stars he/she has by pressing the "Reveal" button.*

*Participant A's task is to then announce his/her total number of stars to participant B with whom he/she is matched with. Participant B cannot see Participant A's cards at any time.*

*Participant B: Participant B receives an initial endowment of 100 tokens and he/she is only informed of the announcement of Participant A with whom he/she is paired about his/her number of stars. Participant B has to decide whether he/she wants to keep his/her tokens or invest them in Participant A's project.*

***Determination of earnings***
*If you are Participant A: You earn a fixed amount of 30 tokens, plus 200 tokens if Participant B has invested in your project.*

*If you are Participant B: Once you have made your investment decision, the program draws one of Participant A's three cards at random.*

- *If you have invested in the project and the card drawn has a star, the project is a success and you earn 300 tokens (i.e., endowment of 100 - investment of 100 + earnings of 300).*

- *If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens (i.e., endowment of 100 - investment of 100 + earnings of 30).*

- *If you did not invest, you earn the 100 tokens of your initial endowment.*

*Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.*

***Your screen***

*At the end of each period, your screen will show a history of what happened in the previous periods. Specifically, you will see four types of information:*

1. *Your announcements in previous periods, if you are Participant A; or the announcements of the different Participants A you were paired with, if you are Participant B;*

2. *Whether the card drawn in previous periods had a star or not, regardless of your role and decision;*

3. *Whether you invested in the previous periods, if you are Participant B; or whether the Participants B you were matched with invested or not, if you are Participant A.*

4. *Your potential earnings in each previous period.*

*For each of you,* ***only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part****. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.*

*Remember that* ***the pairs are re-matched in each period****.*

***Oath***

*At the beginning of part 2,* ***Participants A will have to take an oath about the truthfulness of their announcements in part 2****. By this oath, they will swear on their honour that during this experiment they will be honest and always tell the truth.*

***End of the extract***

**Supplementary instructions for participants in today's session**

The previous instructions apply to you in this part as well but there are some differences from the previous extract:

1. **All participants in today's session are assigned to <u>the role B</u>**; you will keep this role throughout the part.

2. Participants in **role A are not present in this session**: they have already participated in a previous session with other participants in role B. The excerpt of the instructions you read above was for these past participants. Before starting the part, **all participants in role A had to take an oath on honor** that during the experiment they would behave honestly and would always tell the truth. These participants in role A had to type the text of the oath on their computer and this was made known to the participants in role B.

3. **At the beginning of each period, you will be matched with a participant A, randomly selected from a subset of these participants A.**

4. Before you decide whether to invest, you will be informed of the announcement made by this participant A from a previous session after the initial draw of his/her three cards.

5. After deciding whether or not to invest, you will be informed of the card that was randomly drawn by the program from the three cards of this participant A from a previous session. This card drawn will determine your earnings for the period if you have decided to invest.

6. Your investment decisions will only affect your earnings and not the earnings of the participants A from previous sessions with whom you are matched.

**Summary**
This part has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.

In each period you will be informed about the number of stars (0, 1, 2 or 3) announced by a participant A from a previous session in a period drawn at random. You will then have to decide whether or not to invest in the project of participant A from the previous session.

- If you have invested in the project and the card drawn is a star, the project is a success and you earn 300 tokens.

- If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens.

- If you have not invested, you earn the 100 tokens of your initial endowment.

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.

For each of you, **only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part**. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.

Remember that **in each period you will face a decision** (i.e., an announcement made after an initial draw of three cards) **made by one participant A, chosen at random from a subset of the participants A from previous sessions. You are rematched to a new participant A each period** and it is unlikely that the program will select the same participant A for you two periods in a row.

- - -

Please read these instructions again. If you have any questions, please raise your hand or press the red button. We will come and answer your questions immediately in private.

- - -

**End of the instructions in part 2**

*[In the Sufficient Condition treatment, at the end of part 2 but before informing participants of their final gains, we elicited beliefs (displayed on computer screen only), depending on their role in the Announcement Game like in the other treatments. After the belief elicitation about the proportion of honest announcements, participants in this treatment were asked to compare their responses to their initial expectations about the A players' honesty before starting part 2. Note that the wording of the question was slightly adapted between Necessary Condition and Sufficient Condition treatments.]*

We will ask you questions about the possible differences between your initial expectations (after reading the instructions but before starting the part) and your experience during the experiment regarding the reliability of the announcements received from participants A.

- When the announcement was 0 stars:
  - My experience has shown me that these announcements were LESS RELIABLE than my expectations (the 0-star announcements were less truthful than expected)
  - My experience has shown me that these announcements were MORE RELIABLE than I expected (the 0 announcements were more truthful than expected)
  - My experience has shown me that the reliability of these announcements was NEARLY EQUAL to my expectations
- When the announcement was 1 star:
  - My experience has shown me that these announcements were LESS RELIABLE than my expectations (the 1-star announcements were less truthful than expected)
  - My experience has shown me that these announcements were MORE RELIABLE than I expected (the 1-star announcements were more truthful than expected)
  - My experience has shown me that the reliability of these announcements was NEARLY EQUAL to my expectations
- When the announcement was 2 stars:
  - My experience has shown me that these announcements were LESS RELIABLE than my expectations (the 2-star announcements were less truthful than expected)
  - My experience has shown me that these announcements were MORE RELIABLE than I expected (the 2-star announcements were more truthful than expected)
  - My experience has shown me that the reliability of these announcements was NEARLY EQUAL to my expectations
- When the announcement was 3 stars:
  - My experience has shown me that these announcements were LESS RELIABLE than my expectations (the 3-star announcements were less truthful than expected)
  - My experience has shown me that these announcements were MORE RELIABLE than I expected (the 3-star announcements were more truthful than expected)
  - My experience has shown me that the reliability of these announcements was NEARLY EQUAL to my expectations

- - -

# C Online Appendix: Tables

*Table C1: Summary statistics of A players*

| | (1) No-Oath | | (2) Public-Oath | | (3) Private-Oath | | (4) Private-Oath$^{++}$ | | (1-2) | (1-3) | (2-3) | (3-4) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | *p-value* | | | |
| Age (Years) | 21.34 | 3.62 | 20.90 | 1.84 | 20.92 | 2.13 | 21.19 | 2.32 | 0.626 | 0.656 | 0.968 | 0.188 |
| Male (%) | 0.66 | 0.48 | 0.44 | 0.50 | 0.44 | 0.50 | 0.55 | 0.50 | 0.007** | 0.007** | 0.969 | 0.188 |
| Business school (%) | 0.50 | 0.50 | 0.61 | 0.49 | 0.40 | 0.49 | 0.45 | 0.50 | 0.170 | 0.208 | 0.009** | 0.490 |
| Number of observations | 76 | | 77 | | 73 | | 75 | | | | | |

*Notes*: This table summarizes the socio-demographic characteristics of participants in the role of Player A for each treatment. The *p-value* reported using ranksum test for age and tests of proportion for gender (coded 1 as male, 0 otherwise) and business school (coded 1, 0 otherwise). Regression analyses controlled for any effects of these characteristics. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

*Table C2: Summary statistics of B players*

| | (1) No-Oath | | (2) Public-Oath | | (3) Private-Oath | | (4) Private-Oath$^{++}$ | | (5) Necessary Cond. | | (6) Sufficient Cond. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Age (Years) | 20.82 | 2.03 | 20.60 | 1.32 | 21.42 | 2.11 | 21.63 | 2.48 | 21.71 | 1.85 | 20.99 | 1.84 |
| Male (%) | 0.61 | 0.49 | 0.38 | 0.49 | 0.49 | 0.50 | 0.44 | 0.50 | 0.44 | 0.50 | 0.64 | 0.48 |
| Business school (%) | 0.54 | 0.50 | 0.66 | 0.48 | 0.48 | 0.50 | 0.49 | 0.50 | 0.48 | 0.50 | 0.36 | 0.48 |
| Number of observations | 76 | | 77 | | 73 | | 75 | | 77 | | 78 | |

*Notes*: This table summarizes the socio-demographic characteristics of participants in the role of Player B for each treatment. The *p-values* are reported in the next table.

*Table C3: Between-treatment comparisons of socio-demographics of B players*

| | No-Oath *vs.* Public-Oath | No-Oath *vs.* Private-Oath | Public-Oath *vs.* Private-Oath | Public-Oath *vs.* Necessary Cond. | Private-Oath *vs.* Necessary Cond. | Public-Oath *vs.* Sufficient Cond. | No-Oath *vs.* Sufficient Cond. | Private-Oath *vs.* Private-Oath$^{++}$ |
|---|---|---|---|---|---|---|---|---|
| Age (Years) | $p = 0.935$ | $p = 0.024$ | $p = 0.007$ | $p < 0.001$ | $p = 0.126$ | $p = 0.212$ | $p = 0.330$ | $p = 0.610$ |
| Male (%) | $p = 0.005$ | $p = 0.169$ | $p = 0.150$ | $p = 0.413$ | $p = 0.527$ | $p = 0.001$ | $p = 0.647$ | $p = 0.517$ |
| Business school (%) | $p = 0.121$ | $p = 0.464$ | $p = 0.024$ | $p = 0.023$ | $p = 0.990$ | $p < 0.001$ | $p = 0.024$ | $p = 0.866$ |

*Note*: This table reports the p-values from ranksum tests (for age) and tests of proportions (for gender and school) comparing the summary statistics of Player B across treatments. Regression analyses controlled for any effects of these characteristics.

Tables C4 and C5 present the results of random-effects Linear Probability models (GLS) in which the dependent variable is the decision to make a detectable or a deniable lie, respectively. Models (1) and (2) consider all the A players. Models (3) and (4) are restricted to the A players who lied at least once in part 2.

Alternatively to models (3) and (4), we estimated two-step Heckman models, estimating in the first step the probability of making a lie and, in the second step, the probability of making a detectable or a deniable lie, using gender as the identifying variable. However, since the IMR was not significant, meaning that there is no need to correct for a selection bias, we omit these regressions and only report the GLS models.

*Table C4: Probability of making detectable lies*

| Dependent variable: | All A players | | Lied at least once | |
|---|---|---|---|---|
| Detectable lie | (1) | (2) | (3) | (4) |
| *No-Oath* | *Ref.* | *Ref.* | *Ref.* | *Ref.* |
| Public-Oath | -0.201*** | -0.210*** | -0.121 | -0.166* |
| | (0.038) | (0.049) | (0.071) | (0.080) |
| Private-Oath | -0.147*** | -0.153** | -0.112* | -0.125* |
| | (0.040) | (0.051) | (0.052) | (0.062) |
| Period | 0.001 | 0.001 | 0.002 | 0.001 |
| | (0.001) | (0.004) | (0.002) | (0.004) |
| Male | 0.064* | 0.064* | 0.058 | 0.058 |
| | (0.030) | (0.030) | (0.048) | (0.048) |
| Age | -0.008* | -0.008* | -0.010* | -0.010* |
| | (0.003) | (0.003) | (0.004) | (0.004) |
| Business School | -0.015 | -0.015 | -0.008 | -0.008 |
| | (0.030) | (0.030) | (0.047) | (0.047) |
| Public-Oath X Period | | 0.001 | | 0.005 |
| | | (0.004) | | (0.005) |
| Private-Oath X Period | | 0.001 | | 0.001 |
| | | (0.004) | | (0.005) |
| Constant | 0.382*** | 0.387*** | 0.430*** | 0.441*** |
| | (0.088) | (0.094) | (0.108) | (0.113) |
| Number of observations | 3551 | 3551 | 2096 | 2096 |
| Number of clusters | 226 | 226 | 133 | 133 |

*Notes*: This table presents the coefficients from random-effects Linear Probability models (GLS) in which the dependent variable is the decision to make a detectable lie (*i.e.*, announcing three stars, conditional on observing less than three stars), for all A players (models (1) and (2)) and those who lied at least once (models (3) and (4)). Independent variables in model (1) include treatment dummies (with No-Oath treatment as the reference category), period, a male dummy, age (in years), and a dummy for being a student at the business school. Model (2) includes interaction terms between the treatment dummies and the period variable. Models (3) and (4) correspond to models (1) and (2) respectively for the subset of A players who lied at least once. Standard errors are clustered at the individual level. This analysis excludes the Private-Oath$^{++}$ treatment. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

*Table C5: Probability of making deniable lies*

| Dependent variable: | All A players | | Lied at least once | |
|---|---|---|---|---|
| Deniable lie | (1) | (2) | (3) | (4) |
| *No-Oath* | *Ref.* | *Ref.* | *Ref.* | *Ref.* |
| Public-Oath | -0.406*** | -0.424*** | -0.094 | -0.252* |
| | (0.055) | (0.065) | (0.086) | (0.121) |
| Private-Oath | -0.171** | -0.242*** | 0.013 | -0.135 |
| | (0.063) | (0.072) | (0.061) | (0.086) |
| Period | 0.007*** | 0.004 | 0.012*** | 0.004 |
| | (0.002) | (0.004) | (0.003) | (0.005) |
| Male | 0.109* | 0.109* | 0.034 | 0.036 |
| | (0.049) | (0.049) | (0.058) | (0.057) |
| Age | 0.008 | 0.008 | 0.007 | 0.007 |
| | (0.007) | (0.007) | (0.006) | (0.006) |
| Business School | -0.007 | -0.007 | 0.031 | 0.034 |
| | (0.047) | (0.047) | (0.056) | (0.056) |
| Public-Oath X Period | | 0.002 | | 0.017 |
| | | (0.005) | | (0.009) |
| Private-Oath X Period | | 0.007 | | 0.016* |
| | | (0.005) | | (0.007) |
| Constant | 0.277 | 0.307 | 0.325* | 0.401* |
| | (0.160) | (0.163) | (0.153) | (0.157) |
| Number of observations | 2039 | 2039 | 1179 | 1179 |
| Number of clusters | 226 | 226 | 133 | 133 |

*Notes*: This table presents the coefficients from random-effects Linear Probability models (GLS) in which the dependent variable is the decision to make a deniable lie (*i.e.*, announcing one or two stars conditional on observing zero or one star, respectively) for all A players (models (1) and (2)) and those who lied at least once (models (3) and (4)). Independent variables in model (1) include treatment dummies (with No-Oath treatment as the reference category), period, a male dummy, age (in years), and a dummy for being a student at the business school. Model (2) includes interaction terms between the treatment dummies and the period variable. Models (3) and (4) correspond to models (1) and (2), respectively, for the subset of A players who lied at least once. Standard errors are clustered at the individual level. This analysis excludes the Private-Oath$^{++}$ treatment. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

*Table C6: Detectable and deniable lies in the Private-Oath and Private-Oath$^{++}$ treatments*

| Treatments | Private-Oath | Private-Oath$^{++}$ | p-value |
|---|---|---|---|
| **All A players** | | | |
| % Player A who always tell the truth | 39.7% | 38.7% | $p = 0.895$ |
| % Untrue announcements | 27.7% | 29.2% | $p = 0.734$ |
| *Relative frequency* | | | |
| Detectable lies | 9.6% | 10.4% | $p = 0.629$ |
| Deniable lies | 37.9% | 38.1% | $p = 0.896$ |
| *Absolute frequency* | | | |
| Detectable lies | 32.9% | 36.0% | $p = 0.689$ |
| Deniable lies | 56.2% | 58.7% | $p = 0.758$ |
| **Among A players who lie at least once** | | | |
| % Untrue announcements | 46.0% | 47.6% | $p = 0.645$ |
| *Relative frequency* | | | |
| Detectable lies | 15.9% | 16.9% | $p = 0.593$ |
| Deniable lies | 62.8% | 62.2% | $p = 0.934$ |
| *Absolute frequency* | | | |
| Detectable lies | 54.5% | 58.7% | $p = 0.691$ |
| Deniable lies | 93.2% | 95.7% | $p = 0.609$ |

*Notes*: Relative frequency statistics show the average frequency with which participants engage in any particular type of lie. Absolute frequency statistics show the proportion of A players who make any particular type of lie at least once. The relative frequency of detectable lies corresponds to how often A players who saw fewer than three stars announced three stars. The relative frequency of deniable lies corresponds to how frequently A players who saw fewer than two stars reported two stars or announced one star while seeing none. This table presents data both for all A players (top panel) and for the subset who make at least one lie (lower panel).

*Table C7: Determinants of the investment decision (two- and three-star announcements)*

| Dependent variable: Investment decision | (1) | (2) |
|---|---|---|
| *No-Oath* | *Ref.* | *Ref.* |
| Public-Oath | 0.171*** | 0.096* |
| | (0.035) | (0.038) |
| Private-Oath | 0.049 | -0.001 |
| | (0.040) | (0.041) |
| Necessary Cond. | 0.166*** | 0.101* |
| | (0.039) | (0.041) |
| Sufficient Cond. | 0.071 | 0.065 |
| | (0.039) | (0.040) |
| Period | -0.005*** | -0.004*** |
| | (0.001) | (0.001) |
| Male | 0.026 | 0.030 |
| | (0.024) | (0.024) |
| Age | 0.000 | 0.001 |
| | (0.006) | (0.006) |
| Business | -0.011 | -0.017 |
| | (0.025) | (0.025) |
| Index of lies detected | | -0.685*** |
| | | (0.107) |
| % of blank cards on 2 Stars | | -0.064* |
| | | (0.027) |
| Constants | 0.755*** | 0.835*** |
| | (0.131) | (0.132) |
| Number of observations | 6146 | 5552 |
| Number of clusters | 381 | 381 |
| *p-values* from Wald tests | | |
| Necessary Cond. *vs.* Public-Oath | 0.894 | 0.893 |
| Sufficient Cond. *vs.* Public-Oath | 0.006 | 0.404 |

*Notes*: This table presents the coefficients from random-effects Linear Probability models (GLS) in which the dependent variable is the decision to invest, conditioned on receiving a two- or three-star announcement (excluding Private-Oath$^{++}$ treatment). Independent variables in model (1) include treatment dummies (with No-Oath treatment as the reference category), period, a male dummy, age (in years), and a dummy for being a student at the business school. Model (2) includes controls for the percentage of the time a lie has been detected up to the period ('Index of detection') and the percentage of blank cards on two-star announcements up to the period ('% of blank cards on 2 Stars'). Note that interaction terms between the treatments and these two control variables are not significantly different from zero; thus, we do not report this additional model. Standard errors are clustered at the individual level. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

*Table C8: Time trends of investment decision by treatment (all announcements)*

| Dependent variable | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Investment decision | No-Oath | Public-Oath | Private-Oath | Private-Oath$^{++}$ | Necessary Cond. | Sufficient Cond. |
| Period | -0.006* | -0.000 | -0.001 | -0.002 | 0.001 | -0.007*** |
| | (0.003) | (0.002) | (0.002) | (0.003) | (0.001) | (0.001) |
| Constants | 0.718*** | 0.574*** | 0.583*** | 0.645*** | 0.540*** | 0.818*** |
| | (0.032) | (0.027) | (0.034) | (0.034) | (0.028) | (0.027) |
| Number of observations | 1368 | 1386 | 1314 | 1350 | 2079 | 2106 |
| Number of clusters | 76 | 77 | 73 | 75 | 77 | 78 |

*Notes*: This table presents the coefficients from random-effects Linear Probability models (GLS) in which the dependent variable is the decision to invest. The only independent variable is the period. Standard errors are clustered at the individual level. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

# D Online Appendix: Beliefs

We elicited the beliefs of participants at the end of the sessions. The following tables report descriptive statistics and non-parametric tests on these beliefs.

The B players were asked to estimate the proportion of truthful announcements for each number of stars made to them in part 2 (*i.e.,* their first-order beliefs about the honesty of A players). Belief data is notoriously noisy, and this portion of our data is no exception. We note that our belief data are nonetheless often qualitatively consistent with the actions that we observe among our subjects. For example, even though the point estimates do not match actual investment rates, we see in Table D1 that in the No-Oath, Private-Oath, and Sufficient Condition treatments B players are more distrustful when they are faced with a higher announcement than when they are faced with a lower announcement (and indeed in these treatments the A players are in large part dishonest). In contrast, in the Public-Oath and Necessary Condition treatments, which are the treatments in which B players generally experience truthful announcements, the credibility of announcements is more stable across announcements than in the treatments in which the A players are dishonest. The data in Table D2 also shows that the cross-treatment comparisons often match the patterns we observe in actual investment rates, for example, with subjects trusting three-star announcements significantly more in Public-Oath than in No-Oath (first column), or that there are no significant differences in how much they trust a three-star announcement across the No-Oath and Sufficient Condition treatments (second last column).

The within-treatment data on the first-order beliefs of A players on the truthfulness of other A players' announcements are also very noisy (see Table D3). However, the cross-treatment comparisons of these beliefs align with what would be expected given behavior. For example, as Table D4 shows, beliefs on truthfulness levels of zero-star announcements are no different across treatments. In that same table, we also see that there are large and significant cross-treatment differences in beliefs about the truthfulness of three-star announcements: it is the highest in Public-Oath, followed by Private-Oath and then No-Oath.

Finally, the second-order beliefs of A players regarding what they think the B players believe the proportion of truthful announcements by number of stars is (see Tables D5 and D6) generally follow the same patterns as the first order beliefs of A players: for example, the A players believe that B players will believe a three-star announcement more in the Public-Oath than in the Private-Oath, and more in the Private-Oath than in the No-Oath treatment.

*Table D1: First-order beliefs of B players*

| Announcements | No Oath | Public-Oath | Private-Oath | Private-Oath$^{++}$ | Necessary Cond. | Sufficient Cond. |
|---|---|---|---|---|---|---|
| 0 Stars | 81.00 | 86.68 | 81.52 | 44.79 | 79.25 | 77.84 |
| 1 Star | 82.69 | 88.30 | 78.29 | 68.69 | 75.64 | 80.66 |
| 2 Stars | 53.83 | 80.82 | 62.05 | 58.56 | 71.51 | 60.78 |
| 3 Stars | 44.26 | 82.53 | 60.64 | 56.48 | 74.12 | 49.65 |

*Note*: This table reports Player B's first-order beliefs about the proportion of truthful announcements, by number of stars announced and by treatment.

*Table D2: Between-treatment comparisons of first-order beliefs of B players*

| | No-Oath *vs.* Public-Oath | No-Oath *vs.* Private-Oath | No-Oath *vs.* Necessary Cond. | No-Oath *vs.* Sufficient Cond. | Public-Oath *vs.* Private-Oath | Public-Oath *vs.* Necessary Cond. | Private-Oath *vs.* Necessary Cond. | Public-Oath *vs.* Sufficient Cond. | Private-Oath *vs.* Private-Oath$^{++}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 Stars | $p = 0.716$ | $p = 0.684$ | $p = 0.979$ | $p = 0.998$ | $p = 0.917$ | $p = 0.614$ | $p = 0.532$ | $p = 0.773$ | $p < 0.001$ |
| 1 Star | $p = 0.239$ | $p = 0.040$ | $p = 0.002$ | $p = 0.187$ | $p = 0.071$ | $p = 0.011$ | $p = 0.294$ | $p = 0.730$ | $p = 0.097$ |
| 2 Stars | $p < 0.001$ | $p = 0.034$ | $p < 0.001$ | $p = 0.038$ | $p < 0.001$ | $p = 0.009$ | $p = 0.015$ | $p < 0.001$ | $p = 0.441$ |
| 3 Stars | $p < 0.001$ | $p = 0.002$ | $p < 0.001$ | $p = 0.175$ | $p < 0.001$ | $p = 0.023$ | $p = 0.010$ | $p < 0.001$ | $p = 0.420$ |

*Note*: This table reports the p-values from rank-sum tests comparing the 1st-order beliefs of Players B across treatments.

*Table D3: First-order beliefs of A players*

| Announcements | No-Oath | Public-Oath | Private-Oath | Private-Oath$^{++}$ |
|---|---|---|---|---|
| 0 Stars | 63.62 | 75.90 | 69.77 | 67.21 |
| 1 Star | 63.16 | 79.30 | 70.16 | 68.81 |
| 2 Stars | 53.72 | 84.55 | 69.12 | 69.60 |
| 3 Stars | 50.99 | 87.91 | 72.77 | 71.72 |

*Note*: This table displays the first-order beliefs of Players A about the proportion of truthful announcements by the other As, by the number of stars announced, and by treatment.

*Table D4: Between-treatment comparisons of first-order beliefs of A players*

| | No-Oath *vs.* Public-Oath | No-Oath *vs.* Private-Oath | Public-Oath *vs.* Private-Oath | Private-Oath *vs.* Private-Oath$^{++}$ |
|---|---|---|---|---|
| 0 Stars | $p = 0.124$ | $p = 0.281$ | $p = 0.690$ | $p = 0.864$ |
| 1 Star | $p = 0.011$ | $p = 0.160$ | $p = 0.314$ | $p = 0.791$ |
| 2 Stars | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p = 0.999$ |
| 3 Stars | $p < 0.001$ | $p < 0.001$ | $p = 0.009$ | $p = 0.783$ |

*Notes*: This table reports the p-values from rank-sum tests comparing the first-order beliefs of Players A across treatments.

*Table D5: Second-order beliefs of A players*

| Announcements | No-Oath | Public-Oath | Private-Oath | Private-Oath$^{++}$ |
|---|---|---|---|---|
| 0 Stars | 56.07 | 65.19 | 53.07 | 50.68 |
| 1 Star | 56.36 | 68.21 | 55.45 | 54.52 |
| 2 Stars | 57.32 | 79.71 | 59.14 | 66.21 |
| 3 Stars | 48.88 | 88.16 | 67.71 | 68.36 |

*Note*: This table displays the second-order beliefs of Players A, that is, what they think Player B believes about the proportion of truthful announcements, by number of stars announced, and by treatment.

*Table D6: Between-treatment comparisons of the second-order beliefs of A players*

| | No-Oath *vs.* Public-Oath | No-Oath *vs.* Private-Oath | Public-Oath *vs.* Private-Oath | Private-Oath *vs.* Private-Oath$^{++}$ |
|---|---|---|---|---|
| 0 Stars | $p = 0.226$ | $p = 0.939$ | $p = 0.162$ | $p = 0.838$ |
| 1 Star | $p = 0.095$ | $p = 0.784$ | $p = 0.046$ | $p = 0.944$ |
| 2 Stars | $p < 0.001$ | $p = 0.603$ | $p < 0.001$ | $p = 0.065$ |
| 3 Stars | $p < 0.001$ | $p = 0.001$ | $p < 0.001$ | $p = 0.822$ |

*Note*: This table reports the p-values from rank-sum tests comparing the second-order beliefs of Players A across treatments.

# E Online Appendix: Figures

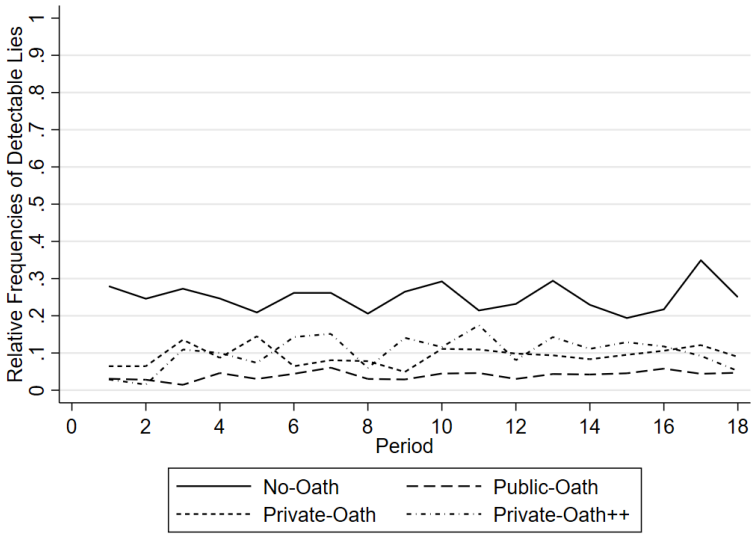


*Figure E1: Evolution of the relative frequencies of detectable lies, by period and treatment*

*Note*: The figure displays the evolution of the relative frequency of detectable lies over time, by treatment, including only the observations in which these lies were possible, that is, excluding the observations in which three stars were observed.

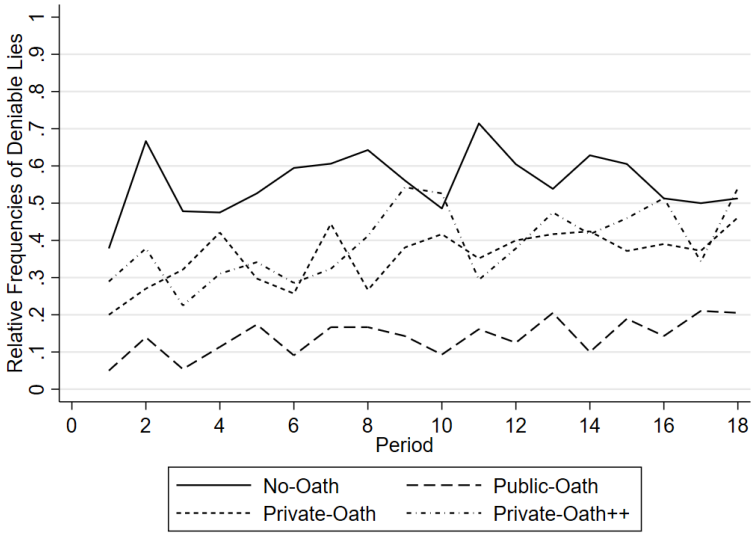


*Figure E2: Evolution of the relative frequencies of deniable lies, by period and treatment*

*Note*: The figure displays the evolution of the relative frequency of deniable lies over time, by treatment, including only the observations in which these lies were possible, that is, excluding the observations in which two or three stars were observed.

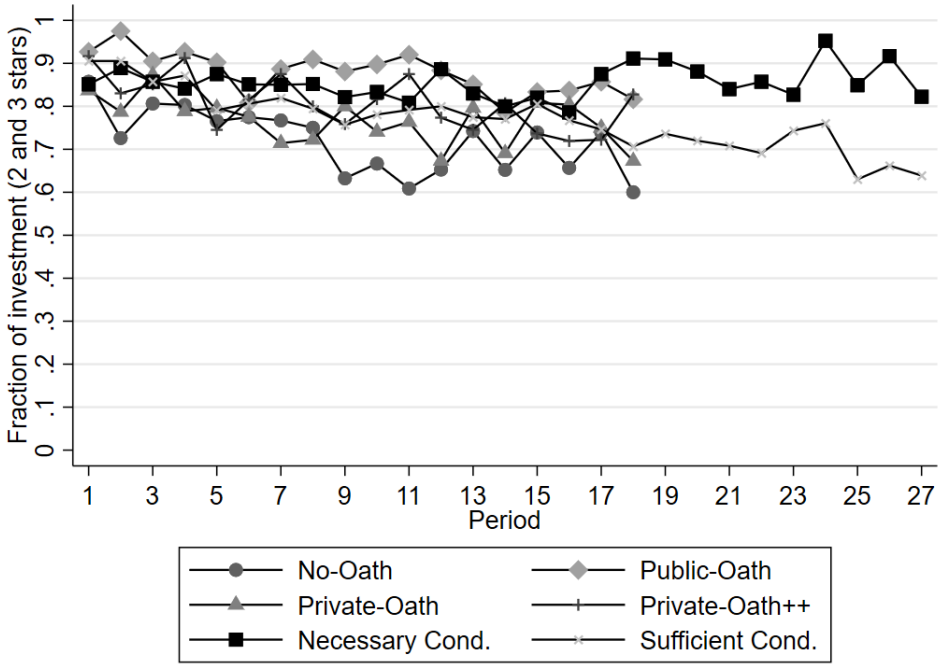


*Figure E3: Evolution of the fraction of investment on 2 and 3 stars, by period and treatment*

*Note*: The figure displays the evolution of the fraction of investments on two- and three-star announcements over time, by treatment. There were 18 periods in the No-Oath, Public-Oath, and Private-Oath treatments, and 27 periods in the Necessary Condition and Sufficient Condition treatments.

# F Online Appendix: Survey on Prolific

We report in subsection F.1 the questionnaire used in the survey conducted in December 2022 via Prolific with 395 individuals residing in the Netherlands (198) and France (197). The survey was collected via Google Forms in Dutch and French respectively, and translated back to English. We analyze the responses in subsection F.2.

## F.1 Questionnaire

**Consent**
This study includes a short survey conducted by Chloe Tergiman (Pennsylvania State University), Marie Claire Villeval (CNRS, University of Lyon) and Sorravich Kingsuwankul (University of Lyon). It takes about 4 minutes to complete the survey.
Participation is voluntary. You can withdraw from the survey at any time without giving a reason. This will not affect the relationship between you, the researchers, or Prolific in any way. All aspects of the study are confidential and anonymous. We will not ask for your name or any information that may identify you during the study. A report of this survey may be submitted for publication, but all information will be used in an aggregate form.
Upon completion of this survey, you will receive a fixed fee of 1€. If you have specific questions about the survey, you can contact us by e-mail at kingsuwankul[at]gate.cnrs.fr
If you agree with the above terms, please select "I agree" to continue with the survey.

- I agree *(Continue)*
- I do not agree *(Exit survey)*

**Please enter your Prolific ID:**
*(Input box here)*

**Do you live in the Netherlands/France?**

- Yes *(Continue)*
- No *(Exit survey)*

Q1. Could you say in general that most people can be trusted or that you have to be very careful when dealing with people?

- Most people can be trusted.
- You have to be very careful.

Q2. Please indicate on a 5-point scale from 1 "I don't trust at all" to 5 "I trust completely", how much you trust each of these groups:

- Doctors
- Members of the parliament
- Judges

- Firefighters
- Bankers

Q3A. *(If reported trust in bankers in Q2 is low: 1-3)* In your response to the previous question, you have indicated that **bankers cannot be completely trusted**. Can you explain the reason(s)? You can choose up to three reasons.

- Banking culture is not trustworthy.
- I/someone I know had a bad experience with bankers/banks.
- The banking industry attracts more dishonest employees than other sectors.
- Bankers' unethical behavior is difficult to detect.
- Unethical behavior by employees is usually not punished by the banks and can sometimes even be rewarded by bonuses.
- Bankers do not prioritize the customers' interests over those of the bank.

Is there another reason, not mentioned above, why you think bankers cannot be completely trusted? If not, you can skip this question.

Q3B. *(If reported trust in bankers in Q2 is high: 4 or 5)* In your response to the previous question, you have indicated that **bankers can generally be trusted**. Can you explain the reason(s)? You can choose up to three reasons.

- Scandals are rare and don't reflect the behavior of most bankers.
- I have had a good experience with bankers/banks so far.
- There are enough safeguards to protect clients.
- Unethical behavior by employees is usually punished by the banks.
- Bankers generally have to place the customers' interests over those of the bank.

Is there another reason, not mentioned above, why you think bankers can generally be trusted? If not, you can skip this question.

Q4. How often do you seek financial advice from your banker or financial advisor?

- I do not have a financial advisor
- Less than once a year
- Once a year
- A few times a year
- Every month or more often

Q5. Do you know whether bankers in the Netherlands/France are required to swear a professional oath of good conduct?

- I am sure that bankers do not have to swear an oath
- I am not sure, but it is unlikely that bankers have to swear an oath
- I am not sure, but it is likely that bankers have to swear an oath
- I am sure that bankers have to swear an oath

Q6. Do you think a bankers' oath should be compulsory in all European countries?

- Yes *(Go to Q7A.)*
- No *(Go to Q7B.)*

Q7A. You responded 'Yes' to the previous question. Why do you think a bankers' oath should be compulsory in all European countries?

- Because it disciplines the bankers
- Because it makes customers' interest more salient to the bankers
- For other reasons

You have responded 'For other reasons' to the previous question. Please explain why you think a bankers' oath should be compulsory in all European countries.

Q7B. You responded 'No' to the previous question. Why do you a banker's oath should not be made compulsory in all European countries?

- Because people can rely on their own experience to trust bankers or not
- Because an oath does not do anything against dishonest bankers
- For other reasons

You have responded 'For other reasons' to the previous question. Please explain why you think a bankers' oath should not be compulsory in all European countries.

Q8. What is your gender?

- Male
- Female
- Other

Q9. In what year were you born? (*e.g.* 1970)

Q10. What is the highest level of education you have obtained? If you are currently enrolled, please select the highest level of education you have attained.

- High school or below

- Bachelor
- Master or above

Q11. What is your current status?

- Student
- Employee
- Self-employed
- Retired
- Unemployed

Q12. What is your annual personal net income (your salary after income tax, social security contributions, and pension contributions have been deducted)?

- $<$ €15000
- €15000 - €29999
- €30000 - €44999
- €45000 - €59999
- €60000 and more

Q13. The next question is about the following problem. In questionnaires like ours, some participants sometimes do not read the questions carefully and just click quickly on the questionnaire. This means that many random answers jeopardize the results of surveys. To show that you have read our questions carefully, give "bankers" as your answer to the following question. What is the profession indicated?

- Firefighters
- Bankers
- Doctors
- Teachers
- Police officers

**End of the survey**

You have completed the survey. We thank you for your time. Important: Please complete the following 3 steps to register your survey response and receive your payment.

1. Write down this survey code: *(Survey completion code here)*
2. Click on 'Submit' on this page to register your responses. If you do not complete this step, we will not receive your information and cannot reward you.
3. Enter the survey code into your Prolific account to register your submission.

## F.2 Results of the survey and discussion

**Low trust in bankers** Using the survey from the Netherlands as our main case, we find low average levels of trust when it comes to the banking sector: when asked to rate their trust on a scale of 1 ('I don't trust at all') to 5 ('I trust completely'), the average rating for bankers is 2.55 (which is barely higher than in the French sample - 2.40 -, although there is no such compulsory oath in France). Bankers rank at the bottom compared with members of parliament (who barely do better with an average of 2.61), judges (3.87), doctors (3.96), and firefighters (4.62). Our survey points to several reasons for this lack of trust: respondents believe bankers do not prioritize clients' interests, that the banking culture is not trustworthy, and that unethical behavior is hard to detect (and therefore to punish). All summary statistics are provided in Table F1.

**Low awareness of the oath** Is the Dutch bankers' oath then of no impact? Our survey shows that despite an oath being compulsory, the awareness of it is limited. In fact, only 53.5% of our respondents were either sure or thought it is likely that such an oath exists. Whether these responses are the result of knowledge of the actual oath or simply the result of the respondents thinking that bankers "obviously" have to swear an oath given their positions is uncertain. To be able to speak to this point, we conducted a similar survey in France, where such an oath does not exist. In that survey about 28% of respondents still believe it is the case. Extrapolating this to the Dutch survey hints that perhaps the answers of a non-negligible proportion of the 53% mentioned above do not stem from the "actual" knowledge of the oath. Responses to the other questions are much more in line with the surveys.

**A correlation between oath awareness and trust in bankers** The results of the survey complement our laboratory findings. While generally trust is low, there is, however, a positive link between being aware of the oath and trust in bankers. Respondents in the Netherlands who are aware of the oath hold a higher level of trust in bankers (Spearman's correlation coefficient $= 0.15$, $p = 0.035$). Our laboratory results show that being aware of the oath is not a necessary condition for the trust since participants in the role of Player B (who can be thought of as investors) learned from their experience. This element is also reflected in our survey, as about 60% of the respondents with high trust in bankers indicated that it was because of the good experiences they have had with their bankers and banks.

In all, there are two takeaways from the field survey. First, the awareness of bankers' oath is far from being universal, even when such a practice has been legally implemented country-wide and many discussions have been conducted both in the media and at the parliament. Second, the banking profession suffers from trust issues, but they may be alleviated as a large majority of people in the survey support the implementation of a bankers' oath. Indeed, Table F1 shows that 83% of the respondents are in favor of a compulsory bankers' oath in the European Union (this is also the case for 80% of the respondents in the French sample). 60% of those who are in favor of such an oath believe that it will help in making their interests more salient, and one-third think that it would discipline bankers. The field experiment of Weitzel and Kirchler (2023), using reminders of the oath, shows that these beliefs are in fact well grounded.

*Table F1: Overview of responses from the Prolific survey in the Netherlands and France*

| Country | Netherlands N=198 | France N=197 |
|---|---|---|
| *Trust attitudes and level of awareness* | | |
| Most people can be trusted | 69% | 45% |
| Mean trust in different groups (St.Dev.) | | |
| Doctors | 3.96 (0.77) | 3.96 (0.75) |
| Members of the parliament | 2.61 (0.87) | 2.22 (0.81) |
| Judges | 3.87 (0.83) | 3.37 (0.95) |
| Firefighters | 4.62 (0.56) | 4.46 (0.68) |
| Bankers | 2.55 (0.96) | 2.40 (0.95) |
| Three most selected reasons for ... | | |
| *Low trust in bankers* | n=162 | n=174 |
| Bankers do not prioritize customers' interest | 78% | 84% |
| Banking culture is not trustworthy | 60% | 65% |
| Unethical behaviors are hard to detect | 49% | 40% |
| *High trust in bankers* | n=36 | n=23 |
| I have had good experiences | 64% | 83% |
| There are enough safeguards | 67% | 39% |
| Scandals are rare | 53% | 61% |
| *Awareness* | | |
| I am sure that bankers do not take an oath | 2.53% | 20.81% |
| I am not sure but it is unlikely that bankers take an oath | 43.94% | 51.27% |
| I am not sure but bankers likely take an oath | 36.36% | 22.84% |
| I am sure that bankers take an oath | 17.17% | 5.08% |
| Spearman's correlation between awareness and trust in bankers | 0.15 ($p = 0.036$) | 0.02 ($p = 0.782$) |
| *Should bankers' oath should be compulsory in EU?* | | |
| Yes because ... | n=165 | n=158 |
| It disciplines bankers | 34% | 37% |
| It makes customers' interest salient | 60% | 57% |
| Other reasons | 6% | 6% |
| No because ... | n=33 | n=39 |
| People can rely on experience | 3% | 13% |
| Oath does not do anything | 94% | 79% |
| Other reasons | 3% | 8% |
| *Socio-economic characteristics* | | |
| Male | 49% | 52% |
| Mean age (St.Dev.) | 36 (9.2) | 37 (10.8) |
| *Education* | | |
| High school or below | 12.63% | 1.02% |
| Bachelor | 49.49% | 10.66% |
| Master or above | 37.88% | 88.32% |
| *Status* | | |
| Student | 7.58% | 11.17% |
| Employee | 76.77% | 55.33% |
| Self-employed | 8.59% | 22.34% |
| Retired | 0.51% | 6.60% |
| Unemployed | 6.57% | 4.57% |
| *Frequency of financial advice* | | |
| No financial advisor | 59.60% | 25.38% |
| Less than once a year | 23.74% | 50.76% |
| Once a year | 11.62% | 18.27% |
| A few times a year | 5.05% | 5.58% |
| Every month or more often | - | - |
| *Annual personal net income* | | |
| < €15000 | 17.17% | 27.92% |
| €15000 - €29999 | 22.73% | 36.55% |
| €30000 - €44999 | 33.33% | 21.32% |
| €45000 - €59999 | 17.68% | 9.14% |
| €60000 or more | 9.09% | 5.08% |

*Notes*: The table reports the descriptive statistics of the responses to the Prolific survey conducted with residents in the Netherlands and France. For awareness of the oath, it is coded as 1 if the respondent indicated that he or she is sure or it is likely that Dutch (or French) bankers take an oath of conduct.

# G Online Appendix: What Modeling Choices Might the Data Suggest

**Insights into modeling active-audience image costs:** These image costs are those that relate to the beliefs that Player B may hold regarding the type of Player A or their actions. Using the lying literature as a guide, one approach is to model them as a function of the B player's belief that the oath has been broken (the A player feels judged on their action). This parallels the models of Gneezy et al. (2018), Abeler et al. (2019), and Khalmetski and Sliwka (2019). The second approach is to think of these costs as coming from a desire to not be perceived as someone for whom it is easy to break an oath (the A player feels judged as a person), paralleling the model of Kartik (2009).

If these image costs are specific to not wanting to be detected as having broken an oath, we should observe that the common-knowledge element of the oath would more sharply reduce "highly detectable" lies compared with "less detectable" ones, since the chances of being identified as a "confirmed oath-breaker" are higher when one announces three stars after observing none than after announcing three stars after observing two, for example. We find that this is not the case: relative to no oath, the reduction in extreme lies caused by the public oath is not statistically different from those in high-risk lies ($p = 0.755$) and lower-risk lies ($p = 0.814$), even though the lower the true number of stars, the higher the benefit of making a detectable lie (recall that the investments in the Public-Oath treatment are monotonic in the number of stars announced). In fact, the absolute probability of making a detectable lie in the Public-Oath treatment is statistically independent of the true number of stars.[1] Therefore, the patterns in our data are incompatible with this modeling choice. While we cannot of course know how difficult a decision to lie is for a participant (morally speaking), active-audience image costs stemming from the oath may in fact come from not wanting to be seen as someone for whom it is *easy* to break an oath rather than simply someone who has or may have broken an oath.

**Insights into modeling self-audience image costs:** We start by considering two possible models of costs: fixed costs (the cost incurred by breaking the oath does not depend on "how badly" the oath was broken), and costs that vary with the size of the lie that was made when breaking it (for example, it may be less costly to break an oath if one is to announce three stars when the truth is two than when the truth is one).

We first highlight that we cannot use data from consistent truth-tellers to better understand these costs, and so focus on subjects who have lied at least once in the Private-Oath treatment.[2] We focus on data from the Private-Oath treatment as it excludes active-audience image costs of breaking an oath and hence allows us to focus *solely* on self-audience image costs.

Fixed costs of breaking the oath are incompatible with the patterns we see in our data: indeed, if the costs of breaking an oath were fixed, then there should be no difference in terms of their impact on detectable and deniable lies. Instead, we see a drop in the former and no difference in the latter: the private oath reduces detectable lies compared with the absence of an oath (27.8% versus 15.9%; $p = 0.003$) but has no impact on deniable lies (63.2% versus 62.8%; $p = 0.884$).

A second possibility is that instead of a fixed cost, the self-audience image costs of breaking the oath, $c$, depend solely on the "size" of the lie, that is the distance between the message $m$ and the truth $\tau$: $c = f(|m - \tau|)$ where $f$ is increasing. We can largely rule this modeling choice out as well. Indeed, let us consider two subjects: one who is facing one star and considering announcing three stars, and the other who is facing zero stars and considering

[1] The probabilities of announcing three stars when the truth is zero, one, or two are 13.2%, 13.0%, and 17.3%, respectively; these are not statistically different as the smallest pairwise $p$-value for signrank tests is 0.563.

[2] In addition, almost all such subjects in the No-Oath treatment who make detectable lies also make deniable ones. This rules out that subjects who lie in the Private-Oath treatment would only be those who exclusively make detectable lies.

announcing two stars. Using our empirical values for the probabilities of investment at various announcement levels, the net benefits of breaking the oath relative to telling the truth are $150 - f(2)$ and $137 - f(2)$, respectively. Thus, we should find that the probability of announcing three stars after seeing one star is greater than the probability of announcing two stars after seeing none. However, we find the opposite, as the probability of the former is 14.5% and the latter 53.7% ($p < 0.001$, signrank test). Thus, fixed costs or costs that depend *only* on the size of the lie cannot explain the patterns in our data.

Our data suggest that an additional mechanism may be at play: subjects simply do not want to lie maximally if they have sworn to not lie at all.[3] Indeed, the probability of announcing three stars when seeing zero is not significantly different from the probability of announcing three stars when seeing one or two (15.5%, 14.5%, 18.5%, respectively; the smallest pairwise $p$-value for signrank tests is 0.172), although telling a maximal lie would yield a higher return than telling the truth when the truth is zero stars than when it is one or two (see Section 4.3). In addition, supporting the argument that "lying maximally" costs are due to breaking the oath, we point out that in the No-Oath treatment, maximal lies are twice as common, and the probabilities of lying maximally in the presence of an oath (whether private or public) are significantly lower than those in the No-Oath treatment (the largest $p$-value is 0.038), showing that the oath itself changes behavior.

The data on the subgroup of A players who lie at least once (bottom panel of Table 1) further supports this modeling. Among these liars, there is no difference in deniable lies across our three treatments, suggesting that none of the image costs we consider play a large role. On the other hand, detectable lies (by our design, maximal lies) are statistically fewer under oath relative to no oath – with no difference between Public-Oath and Private-Oath. This implies that self-image costs of breaking an oath is sufficient in discouraging from lying blatantly, while the costs associated with judgment by the B players do not further suppress maximal lies. Thus, the observed pattern is consistent with costs of breaking an oath being related to breaking it maximally since only detectable lies are impacted.

[3]Recall that in the Private-Oath treatment, the B players are unaware of the oath. Thus, the self-audience image costs of breaking an oath and lying maximally are not related to the active-audience image costs of being identified as someone who lies despite signing an oath.

# References for Online Appendix

## References


Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for Truth-Telling. *Econometrica 87*(4), 1115–1153.

Akin, Z. (2022). Playing the victim behavior: An experimental study. Technical report, American University in Dubai.

Babin, J. J. and H. S. Chauhan (2023). Show no quarter: combating plausible lies with ex-ante honesty oaths. *Journal of the Economic Science Association 9*, 66–76.

Babin, J. J., H. S. Chauhan, and F. Liu (2022). You Can't Hide Your Lying Eyes: Honesty Oaths and Misrepresentation. *Journal of Behavioral and Experimental Economics 98*, 101880.

Beck, T. (2021). How the honesty oath works: Quick, intuitive truth telling under oath. *Journal of Behavioral and Experimental Economics 94*, 101728.

Beck, T., C. Buhren, B. Frank, and E. Khachatryan (2020). Can honesty oaths, peer interaction, or monitoring mitigate lying? *Journal of Business Ethics 163*(3), 467–484.

Cagala, T., U. Glogowsky, and J. Rincke (2024). Detecting and Preventing Cheating in Exams: Evidence from a Field Experiment. *Journal of Human Resources 59*(1), 210–241.

Cagala, T., U. Glogowsky, J. Rincke, and S. Schudy (2024). Commitment Requests Do Not Affect Truth-Telling in Laboratory and Online Experiments. *Games and Economic Behavior 143*, 179–190.

Carlsson, F., M. Kataria, A. Krupnick, E. Lampi, Löfgren, P. Qin, T. Sterner, and S. Chung (2013). The truth, the whole truth, and nothing but the truth–a multiple country test of an oath script. *Journal of Economic Behavior & Organization 89*, 105–121.

Davis, B. J. and T. Jaber-Lopez (2022). Do voluntary commitment mechanisms improve welfare? The effect of mandatory and voluntary oaths in a social dilemma. *Bulletin of Economic Research 75*(2), 525–540.

Gneezy, U., A. Kajackaite, and J. Sobel (2018). Lying Aversion and the Size of the Lie. *American Economic Review 108*(2), 419–453.

Heinicke, F., S. Rosenkranz, and U. Weitzel (2019). The effect of pledges on the distribution of lying behavior: An online experiment. *Journal of Economic Psychology 73*, 136–151.

Hergueux, J., N. Jacquemet, S. Luchini, and J. F. Shogren (2022). Leveraging the Honor Code: Public Goods Contributions under Oath. *Environmental and Resource Economics 81*(3), 591–616.

Jacquemet, N., A. James, S. Luchini, J. Murphy, and J. Shogren (2021). Do truth-telling oaths improve honesty in crowd-working? *PLoS ONE 16*.

Jacquemet, N., A. James, S. Luchini, and J. F. Shogren (2017). Referenda Under Oath. *Environmental and Resource Economics 67*(3), 479–504.

Jacquemet, N., R.-V. Joule, S. Luchini, and J. F. Shogren (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management 65*, 110–132.

Jacquemet, N., S. Luchini, A. Malézieux, and J. F. Shogren (2020). Who'll stop lying under oath? Empirical evidence from tax evasion games. *European Economic Review 124*, 103369.

Jacquemet, N., S. Luchini, J. Rosaz, and J. F. Shogren (2019). Truth Telling Under Oath. *Management Science 65*(1), 426–438.

Jacquemet, N., S. Luchini, J. Rosaz, and J. F. Shogren (2021). Can we commit future managers to honesty? *Frontiers in Psychology*, 2785.

Jacquemet, N., S. Luchini, J. F. Shogren, and A. Zylbersztejn (2018). Coordination with communication under oath. *Experimental Economics 21*(3), 627–649.

Jacquemet, N., S. Luchini, J. F. Shogren, and A. Zylbersztejn (2024). Commitment to the truth creates trust in market exchange: Experimental evidence. *Games and Economic Behavior 148*, 279–295.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies 76*(4), 1359–1395.

Kemper, N., J. Popp, and R. Nayga (2020). A query theory account of a discrete choice experiment under oath. *European Review of Agricultural Economics 47*(3), 1133–1172.

Khalmetski, K. and D. Sliwka (2019). Disguising lies—image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics 11*(4), 79–110.

Koessler, A.-K. (2022). Pledges and how social influence shapes their effectiveness. *Journal of Behavioral and Experimental Economics 98*, 101848.

Koessler, A.-K., L. Page, and U. Dulleck (2021). Public cooperation statements. *Journal of Economic Interaction and Coordination 16*(4), 747–767.

Koessler, A.-K., B. Torgler, L. P. Feld, and B. S. Frey (2019). Commitment to pay taxes: Results from field and laboratory experiments. *European Economic Review 115*, 78–98.

Krüger, M. P. (2016). Can an oath reduce dishonesty and overconfidence for bankers? Technical report, Mimeo.

Peer, E. and Y. Feldman (2021). Honesty pledges for the behaviorally-based regulation of dishonesty. *Journal of European Public Policy*, 1–21.

Prima, R. A., S. Feeny, R. Hoffmann, and E. Satriawan (2020). Threats, Pledges, and Asset Misreporting: A Framed Field Experiment in Indonesia. *TNP2K Working Paper 57-e/2020*.

Schild, C., D. W. Heck, K. A. Ścigała, and I. Zettler (2019). Revisiting REVISE: (Re)Testing unique and combined effects of REminding, VIsibility, and SElf-engagement manipulations on cheating behavior. *Journal of Economic Psychology 75*, 102161.

Weitzel, U. and M. Kirchler (2023). The Banker's Oath And Financial Advice. *Journal of Banking & Finance 148*, 106750.

Zickfeld, J. H., K. A. Ścigała, C. T. Elbæk, J. Michael, M. H. Tønnesen, G. Levy, S. Ayal, I. Thielmann, L. Nockur, E. Peer, V. Capraro, R. Barkan, S. Bø, Bahník, D. Nosenzo, R. Hertwig, N. Mazar, A. Weiss, A.-K. Koessler, R. Montal-Rosenberg, S. Hafenbrädl, Y. A. Nielsen, P. Kanngiesser, S. Schindler, P. Gerlach, N. Köbis, N. Jacquemet, M. Vranka, D. Ariely, J. B. Martuza, Y. Feldman, M. Białek, J. K. Woike, Z. Rahwan, A. Seidl, E. Chou, A. Kajackaite, S. Schudy, U. Glogowsky, A. Z. Czarna, S. Pfattheicher, and P. Mitkidis (2025). Effectiveness of ex ante honesty oaths in reducing dishonesty depends on content. *Nature Human Behaviour 9*(1), 169–187.