cbess
Centre for Behavioural and Experimental Social Science

University of East Anglia

2019-02

EXPLORING IMAGE MOTIVATION IN PROMISE KEEPING – AN EXPERIMENTAL INVESTIGATION

Kevin Grubiak

www.uea.ac.uk/cbess

Exploring Image Motivation in Promise Keeping – An Experimental Investigation[*]

Kevin Grubiak[†]

May 13, 2019

Abstract

This paper reports an experiment designed to investigate the role of image concerns in promise keeping. The task employed allows to shed light on the relevance of both *social*-image and *self*-image concerns. Whereas in the former case, behavior is expected to depend on how *others* perceive a given action, in the latter case what matters is how actions reflect on a decision-maker's *self*-perception. We observe strong evidence of social-image concerns in treatments which feature ex-ante opportunities for promise exchange. Ruling out alternative explanations, our results are consistent with subjects exhibiting an aversion to being perceived as a promise breaker by others. Surprisingly, subjects seem not to anticipate social-image concerns to be present in others. Our test of self-image concerns yields a null result: there is no evidence suggesting that subjects in our experiment engaged in self-deception to evade their promise-induced commitments. This resilience can be interpreted as corroborating evidence of the strength of promises. Our results shed light on the conditions under which promises can be expected to facilitate successful relationships based on trust.

Keywords: Trust; Communication; Promises; Image Concerns; Beliefs

JEL Classification: C91; D03; D82; D83

[*] I would like to thank my supervisors Robert Sugden, Anders Poulsen, and Odile Poulsen for helpful guidance and financial support. I would also like to thank David Hugh-Jones for serving as a discussant at the design stage of the experiment and participants of the annual CCC (CBESS-CEDEX-CREED) meeting held in 2018 for their feedback.

[†] School of Economics and Centre of Behavioural and Experimental Social Science, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK. Email: k.grubiak@uea.ac.uk.

1 Introduction

Trust plays an important role in many economic interactions. It is a prerequisite for interactions where legal contracts are not enforceable or simply too expensive to implement. Moreover, trust can provide substantial efficiency gains, for instance, by speeding up the process of decision making. Despite its potential benefits, however, trust carries the risk of betrayal.

Yet, abundant evidence documents that people are far more trustworthy than the standard economic model resting on the assumption of pure self-interest would assert. Prominent explanations relate to intrinsic preferences for concepts like fairness, equality, or reciprocity. But also situational factors like the ability to talk and exchange promises have widely been observed to increase trust and trustworthiness. The inclination to keep a promise can theoretically and empirically be accounted for by the *commitment-based* (Vanberg, 2008) as well as the *expectations-based* (Charness and Dufwenberg, 2006) explanations. According to the former, people keep their promises because they have an intrinsic preference for keeping their word. According to the latter, promises are kept because they induce a shift in promisee expectations and, thus, higher experienced guilt by the promise maker. Although these theories are not mutually exclusive, follow-up research has used ever more sophisticated experimental protocols in an attempt to cleanly distinguish between these two motivations of promise keeping (e.g., Vanberg, 2008; Schwartz, Spires and Young, 2018; Bhattacharya and Sengupta, 2016; Ederer and Stremitzer, 2017; Ismayilov and Potters, 2016; Mischkowski, Stone and Stremitzer, 2016; Di Bartolomeo et al., 2019). Although guilt aversion appears to play a significant role, promises are frequently kept even when guilt is ruled out as an explanation. On balance, these studies provide remarkable support in favor of an intrinsic preference for promise keeping.

Alternative explanations of promise keeping which have yet received little attention in the literature are *image concerns*. People may keep promises because they don't want to be thought of badly by others. Although the conventional workhorse in the literature on promise keeping is a hidden action trust game, choices in most variants of this game remain transparent to the experimenter. Consequently, subjects may be concerned about how they are being perceived by the experimenter. Even if experimenter observability were to play a negligible role, these studies are silent on the interesting question of how the threat of being exposed as a promise breaker affects behavior. The first contribution of our paper is to address this question.

Distinct from *social*-image concerns as outlined before are *self*-image concerns. People like to think of themselves as fair and honorable human-beings and where these perceptions are at stake, as in the case of opportunistic temptation, so is their *self-concept*. Whereas psychologists have long recognized the importance of self-image concerns for behavior (e.g., Baumeister et al., 1998; Bem, 1972; Fiske, 2018), economists have only recently started to incorporate this construct into what could be coined "third-generation" theories of moral behavior, based on models of identity management (Bénabou and Tirole, 2011). Theories of self-concept maintenance (Mazar, Amir and Ariely, 2008), self-signalling (Bodner and Prelec, 2003; Bénabou and Tirole, 2004, 2006; Grossman and Van der Weele, 2017) and identity (Akerlof and Kranton, 2000, 2010) have moreover proven useful in organizing experimental findings unexplained by standard theories of social preference. Identity management, in particular *self-deception*, is able to explain why people avoid costless information (Dana, Weber and Kuang, 2007), sort-out of morally demanding environments (Dana, Cain and Dawes, 2006; Lazear, Malmendier and Weber, 2012), trade off good deeds with bad deeds (Mazar and Zhong, 2010; Merritt, Effron and Monin, 2010) or delegate the execution of opportunistic decisions to third-parties (Hamman, Loewenstein and Weber, 2010).

The second contribution of our paper is to investigate whether self-deception can also mitigate the effectiveness of promises. For the purpose of illustration, consider Bob who promised Ann to proofread her paper under the condition that he finishes his exam revision on time. Bob would rather want to avoid the additional workload but he also does not want to be considered by Ann as a promise breaker. One obvious remedy for Bob is to lie to Ann by claiming that he had not finished his revision on time. This strategy, however, may not be feasible for Bob if he also does not want to think badly of himself. Yet, there are ways out. Bob could engage in self-deception. He could (subconsciously) reduce his work pace or delay completion e.g. by prolonging breaks or by prioritizing other duties. He could eventually convince himself that he would have kept his promise were it not due to (seemingly) external circumstances that prevented him from doing so. Restricting Bob's strategy space by excluding the latter options as done in previous research may lead to the false conclusion that Bob's observed trustworthiness is the result of his intrinsic desire to keep his word.

Analogously to this example and our previous discussion on social image, in this paper we seek to investigate the effectiveness of promises when circumstances allow people to deceive others and themselves about the underlying cause of a broken promise. Our study adds to an evolving literature on social image concerns and, to the best of our knowledge, is the first to test for self-image concerns in promise keeping.

Our methodological vehicle is a laboratory experiment. The remainder of this paper is structured as follows. Section 2 reviews the related literature in more detail. Section 3 elaborates on the experimental design, hypotheses and procedures. Section 4 presents the results. Section 5 contains a discussion. Section 6 concludes the analysis.

2 Related Literature

Our study connects two strands of the literature which, by and large, have only been considered in isolation from each other: the literatures on *promise keeping* and on *image concerns*. In this section, we review each respective literature and comment on how a joint perspective could improve our understanding of the effectiveness of non-binding verbal commitments.

2.1 Promise Keeping

Although standard economic theory discards any influence that pre-play communication can have on subsequent behavior, numerous studies have documented that communication, in particular the use of promises, can substantially increase cooperation. Unaccounted for by the standard approach, people may be averse to lying or dislike letting others down on what they promised them they would do, which may eventually render cheap talk *credible*.

In a seminal paper, Charness and Dufwenberg (2006) introduce a hidden-action trust game framework with pre-play communication and find that promises significantly increase cooperation. The cooperative (*In, Roll*) profile occurred 20% of the time without communication and 50% of the time with communication. They argue that their results square well with a model of guilt aversion by which promises feed expectations which the promisor dislikes to violate (*expectations-based* explanation).

Yet, a popular alternative explanation of their results is that people may hold an intrinsic preference for keeping their word (*commitment-based* explanation). A series of papers have been dedicated to disentangling these two explanations of promise keeping. The first of which, Vanberg (2008), uses a variant of the hidden-action trust game where subjects are informed that there is a 50% chance that they will be re-matched to a different subject than the one they previously communicated with. Only the promisor is informed about whether his partner was switched and he is allowed to inspect the message that his new partner has received earlier, before the switch occurred (hence, he knows whether or not a promise was received). From the perspective of the promisee who is unaware whether or not a switch occurred, first-order beliefs about the promisor's trustworthiness should not differ across conditions. Anticipating this, the promisor's second-order belief and hence the guilt potentially experienced should not differ either. Holding second-order beliefs constant, Vanberg finds that a dictator's own promise affects behavior whereas a promise that was made by others does not.[1] He argues that this result appears to be incompatible with the expectations-based explanation of promise keeping and lends support to the commitment-based explanation.

[1] Dictators who promised chose the cooperative outcome 73% of the time whereas those who didn't promise chose it 52% of the time.

Ederer and Stremitzer (2017) claim that the lack of evidence of guilt aversion in promise keeping observed by Vanberg (2008) may result from the possibility that guilt is only experienced if one is directly responsible for inducing an increase in the expectations of a promisee. Recall that in Vanberg's study, the increase in expectations in the control condition is induced by *another* dictator's promise, whereas expectations are affected by the dictator's *own* promise in the main condition. The authors use an "unreliable random device" to generate exogenous variation in second-order beliefs and provide evidence of guilt aversion in promise keeping. However, since their design does not include an analog to Vanberg's partner-switching mechanism, Ederer and Stremitzer cannot assess the strength of the expectations-based explanation relative to the strength of the commitment-based explanation (and their study is also not intended to do so).

In a unified framework, Di Bartolomeo et al. (2019) study an environment that allows for exogenous variation of *both* promises and expectations allowing them to test which channel is quantitatively more important. They essentially combine the earlier designs by Vanberg (2008) and Ederer and Stremitzer (2017). More precisely, they make the partner-switching probability in Vanberg's design a separate treatment variable that randomly takes the value *low* (25%) or *high* (75%) to generate exogenous variation in expectations. Their results suggest that promise keeping is *independent* of beliefs. Promise keeping rates are as high when beliefs are low (as induced by a high switch probability) as when beliefs are high (as induced by a low switch probability). Nonetheless, they observe an independent effect of higher expectations on cooperation as predicted by guilt aversion.[2]

The overall picture documents that (i) the use of promises is a powerful way of increasing cooperation and efficiency and (ii) that its effect is mediated by *both* an intrinsic preference for promise keeping and guilt aversion. Yet, another motivation for promise keeping not accounted for in these studies is *image* motivation.

2.2 Social-Image Concerns

Social-image concerns relate to people's concerns over how their actions are perceived by others. A vast body of research has documented that choices depend on observability (Andreoni and Petrie, 2004; Andreoni and Bernheim, 2009; Ariely, Bracha and Meier, 2009; Bohnet and Frey, 1999; Bursztyn and Jensen, 2017; Chaudhuri, 2011; Dana, Cain and Dawes, 2006; Ekström, 2012; Fehr and Gächter, 2000; Rege, 2004; Rege and Telle, 2004; Soetevent, 2005; Tadelis, 2011). Altruistic behav-

[2] Mischkowski, Stone and Stremitzer (2016) use a vignette study to manipulate expectations about the behavior of others in a more controlled fashion by simply asking subjects to assume a counterpart would hold specific beliefs. The authors provide empirical support for an expectations per se effect, a promising per se effect and an interaction effect by which a subject becomes more sensitive to guilt towards a promisee.

ior in the well-known dictator game, for instance, has been shown to be sensitive to the possibility that the experimenter could infer choices (Hoffman et al., 1994; Hoffman, McCabe and Smith, 1996) and many studies have documented that what looks like *giving* can oftentimes be better classified as *giving-in* to social pressure (Cain, Dana and Newman, 2014).

In situations where people are directly confronted with pro-social requests, many follow reluctantly in an attempt to avoid the experience of shame. A reluctance to enter sharing environments has been documented in several field and laboratory studies. In a door-to-door fundraising study, DellaVigna, List and Malmendier (2012) observe that informing household about an upcoming donation request significantly reduces the share of households opening the door. Dana, Cain and Dawes (2006) as well as Lazear, Malmendier and Weber (2012) document the same pattern in laboratory experiments where subjects are willing to (silently) sort-out of a dictator game at a cost.

Rather recently, scholars have started to investigate the robustness of several concepts which have previously been thought of as resulting from intrinsic preferences. Malmendier, te Velde and Weber (2014) document that a preference for reciprocating others' kindness is less strong than previously thought when accounting for social pressure. Another example is Kriss, Weber and Xiao (2016) who observe that third-parties punish norm violators reluctantly, i.e., although they indicate a preference for punishment, they ultimately avoid the act of punishing if excuses allow them to do so without exposure. Attributing responsibility to nature allows subjects to maintain a positive image in the eyes of other subjects and the experimenter.

One of the aims of our paper is to assess the role that social image concerns play in promise keeping. We are only aware of a few studies approaching this or similar territories. Deck, Servátka and Tucker (2013) argue that the effectiveness of promises observed in previous studies could be driven by experimenter observability. The authors, however, are unable to document image concerns in their study due to the fact that they could not reproduce an effect of promises under *both* a single-blind *and* a double-blind experimental protocol. Schütte and Thoma (2014) vary the ex-post observability of a promising party's action to test for social-image concerns. They are unable to document a robust effect. One possibility is that the very high rate of promise keeping observed in their baseline treatment (81%) limited the scope of image-concerns to be detectable. Greenberg, Smeets and Zhurakhovska (2015) investigate ex-post disclosure of dishonest messages in a sender-receiver game and find that it almost doubles the incidence of truth-telling. Although they can rule out guilt aversion as an explanation, it is unclear whether the effect of disclosure results from being perceived as *dishonest* or, more generally, *selfish*.[3]

[3] Although the receiver is uninformed about the payoff consequences to the sender from lying, it is likely that he will associate lies with outcomes that favor the sender. Consequently, the sender

Our paper adds to this literature by providing an experimental environment which allows to test the relevance of social image concerns in promise keeping, accounting for the limitations of previous studies as outlined before. We are moreover extending the analysis to the consideration of self-image concerns.

2.3 Self-Image Concerns

The role of *self-image concerns* has recently attracted a lot of interest in the economics literature. Bodner and Prelec (2003) provide a theoretical model in which utility can be decomposed into *outcome utility* and *diagnostic utility.* The basic underlying thought of their model is that people may draw inferences from their actions about their dispositions in situations where the latter are not directly accessible. In their setup, very much like in the one of Bénabou and Tirole (2006, 2011), an agent can be thought of as being comprised of multiple selves, one of which being uncertain about the agent's true dispositions. Through his actions, the agent can send an informative signal to the "observer-self" in an attempt to obtain positive diagnostic utility. Conversely, this framework allows agents to engage in self-deception as means to avoid negative diagnostic utility associated with opportunistic behavior.

Consistent with these theories, several studies have documented that behavior can be biased in self-serving ways that allow people to *save face* in front of others and even themselves. Haisley and Weber (2010) show that subjects use ambiguity in an experimental labor market as an excuse for letting-off workers. Exley (2015) documents that subjects use risk as an excuse not to give to Charity. Di Tella et al. (2015) find that subjects manipulate their beliefs about others' altruism to justify selfish behavior. Spiekermann and Weiss (2016) show that subjects selectively acquire information which, in expected terms, lowers normative obligations.

The importance of self-image concerns is also emphasized in a seminal study by Dana, Weber and Kuang (2007). In a series of experiments, they provide robust evidence documenting that subjects use moral excuses (or, *moral wiggle room*) to not only deceive others but also themselves about the true cause of an unfair allocation. They demonstrate that dictators justify selfish choices by remaining willfully ignorant about the welfare consequences of their actions to the receiver. Dictators also refrain from implementing socially optimal outcomes if others are able to do so on their behalf. Their final "plausible deniability" treatment is specifically designed to discern social- and self-image concerns. Here, a dictator can choose between an allocation favoring himself over the recipient, or an equal and efficient allocation. The twist in this treatment is that the dictator can lose agency if he delays his decision in which case the computer intervenes to implement either outcome with equal probability. The recipient can never tell whether a selfish outcome resulted from a

may not only worry about being perceived as a liar, but also about being perceived as selfish.

willful decision or an unlucky dictator. Interestingly, since the computer implements a lottery between the two outcomes which is lower in expected value than choosing either outcome directly, delegation is inconsistent with purely outcome-based theories of behavior. Self-image concerns, instead, become a natural candidate for explaining subjects' willingness to delegate the decision to the computer. With 50% probability, the computer would choose the fair outcome the dictator would have felt compelled to choose anyway, but otherwise the selfish outcome would obtain and the dictator could maintain the illusion of not being responsible for its implementation. Indeed, a substantial proportion of dictators in their study (24%) allowed themselves to be cut off, thereby avoiding to make a decision.[4] The deniability mechanism has also been applied to the analysis of reciprocal preferences e.g. by van der Weele et al. (2014) and Regner (2018).

In our paper, we implement a variant of the cut-off mechanism to investigate the relevance of self-image concerns in promise keeping. To the best of our knowledge, all studies on promise keeping make it perfectly transparent to the decision maker that he himself is responsible for a broken promise, or, put differently, promise breaking is an act of commission. Yet, the responsibility for a broken promise can also be shifted to external circumstances, thereby granting a decision maker a moral excuse for selfish behavior without compromising his self-image.[5]

3 The Experiment

3.1 Design

We combine a binary dictator game with a *matrix solving* task and systematically vary between subjects (i) the degree to which a "plausible deniability" mechanism allows subjects to obfuscate responsibility for outcomes and (ii) whether or not the experiment features a communication stage. Table 1 summarises our 2x2 factorial treatment design. The sequence of stages in the experiment is depicted in Figure 1.

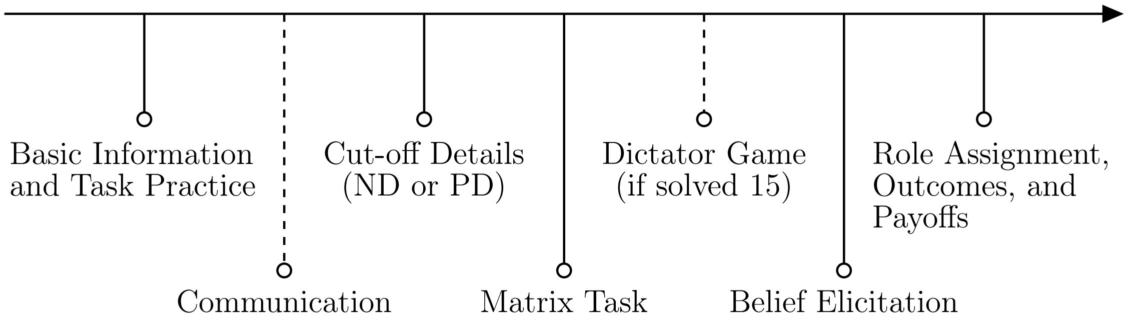
Table 1: Factorial Treatment Design

	No Deniability	Plausible Deniability
No Communication	NC_ND	NC_PD
Communication	C_ND	C_PD

[4] Note that the cut-off timer was calibrated in a way such that subjects who really wanted to make a decision themselves had enough time to do so.

[5] Note how self-deception may alleviate both channels found to affect promise keeping. One might not feel to have broken a promise *and/or* the perceived lack of responsibility may decrease or even fully erase any form of experienced guilt.

Figure 1: Sequence of Stages in the Experiment


Basic Information and Task Practice
Communication
Cut-off Details (ND or PD)
Matrix Task
Dictator Game (if solved 15)
Belief Elicitation
Role Assignment, Outcomes, and Payoffs

Subjects are randomly paired in groups of two. Role assignment takes place *at the end* of the experiment, i.e., all subjects simultaneously play as A players (potential dictators) knowing that outcomes in this role would only count for half of them whereas the other half would eventually serve the role of player B (recipient).[6]

All treatments have in common that the dictator game stage is only reached if a preceding matrix task is solved successfully. In case of *success*, the subject enters the dictator game stage and decides how to allocate money between herself and her counterpart by choosing one of two possible allocations: A=(£10,£0) or B=(£6,£6). Conversely, in case of *no success*, the subject skips the dictator game stage and is forced to let the computer randomly implement either of the two allocations with equal probability on her behalf.

The matrix task, borrowed from Abeler et al. (2011), consists of subjects counting *ones* (1s) in a series of 5x5 matrices comprised of randomly ordered zeros and ones.[7] Importantly, we modified the task to feature a cut-off mechanism which (in some of our treatments) can serve as a plausible excuse for the implementation of the selfish allocation A (£10, £0).[8] Successful completion requires a subject to solve a target amount of 15 matrices *on time*, i.e. before being cut off by the computer.

We employ different variants of the cut-off mechanism in our experiment. In our *No Deniability* (ND) treatments (Table 1, first column), subjects are given 300 seconds (5 minutes) to work on the task until a cut-off occurs. The time allotted in these treatments is extremely generous based on the results of an informal and un-incentivized pretest where subjects needed on average 104s to solve 15 matrices and no subject took longer than 138s. Our aim was to erase the opportunity of using

[6] In the instructions, we refer to "you" and "your counterpart" instead of "dictator" and "recipient". Instructions can be found in Appendix B.

[7] Appendix C provides screenshots of the experimental interface.

[8] Recall that in Dana, Weber and Kuang (2007), 24% of the subjects allowed themselves to be cut-off by the computer, thereby preferring a mixture of two outcomes over each one separately. This observation is “inconsistent with a theory of rational choice with utilities defined only over outcomes” (p. 74). For subjects who are feeling compelled to choose the other-regarding option in order not to threaten their self-image, however, being cut off can be desirable. In half of the cases, the outcome would obtain which the dictator would have felled compelled to choose anyway. In another half of the cases, the opportunistic outcome would obtain allowing the subject to uphold the illusion of not being responsible for its implementation.

the cut-off mechanism as a plausible excuse for selfish allocations whilst keeping the experimental protocol as close as possible to the treatments we describe next.

In our *Plausible Deniability* (PD) treatments (Table 1, second column), instead of telling subjects that the cut-off would occur after 300 seconds sharp, we tell them that the cut-off can occur at any randomly determined second within the 300 seconds interval.[9] The PD treatments offer room for two distinct dimensions of deniability:

- *Deniability towards the counterpart.* A subject can exploit the fact that her counterpart cannot ascertain whether an outcome came about by the subject's own choice or by the computer. Our plausible deniability treatments therefore alleviate the social-image cost that is usually associated with selfish behavior under full transparency.

- *Deniability towards the self.* A subject who feels compelled to choose the other-regarding outcome because she does not want to think badly of herself may prefer to be cut off by the computer. A cut-off results in a fair chance (50%) of obtaining the opportunistic outcome whilst allowing to maintain the illusion of not being responsible for its implementation.

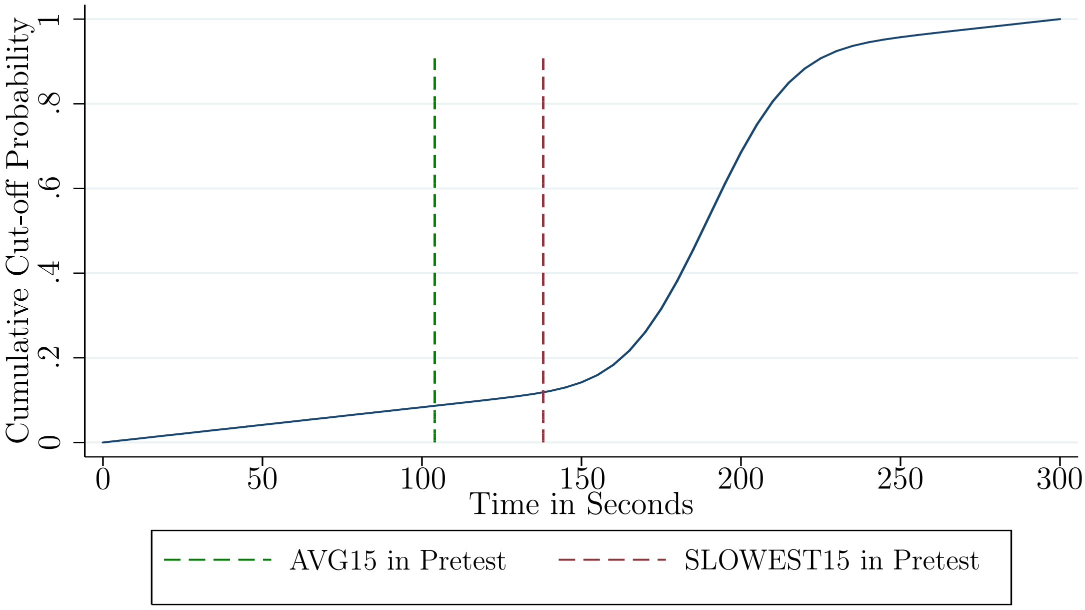
We assumed that self-deceivers in our experiment would work on the task half-heartedly, waste time, or commit more errors all of which delaying the completion of the task.[10] To identify whether subjects in our PD treatments indeed procrastinated, an additional control treatment was conducted. This treatment was designed as closely as possible to the NC_PD treatment. The only difference was the absence of a counterpart. In this treatment, successful completion of the matrix task allowed the dictator to choose her own payoff only (£10 or £6). Since any incentives for procrastination in the matrix task were removed in this treatment, we aimed to obtain an unbiased distribution of performances in the matrix task against which to compare performances in our main treatments. Instructions for the control treatment can be found in Appendix B.2.

No information was disclosed to subjects regarding the underlying distribution that generated the cut-offs in our PD treatments (and the control). Whilst it is technically true that a cut-off could occur anywhere within the specified time interval, we used a distribution which favored later cut-offs. To be precise, we combined a

[9] If a cut-off occurred, a subject was asked to work on a follow-up task for the remainder of the 300 seconds. The task was not incentivized and consisted of adding up numbers on screen. The purpose of this task was to maintain a constant sound of mouse clicks in the background, thereby ruling out that subjects could infer from the lack of this sound information about the timing of cut-offs of their peers.

[10] Previous studies which utilized a cut-off mechanism required self-deceivers to be passive and to wait for the computer to intervene. We decided to embed our cut-off mechanism into a real effort task instead of the dictator game itself to reduce potential demand effects and to mimic a richer (and in our opinion, more realistic) environment that would allow subjects to hide their intentions in an non-obvious way, by disguising their true ability in an active task.

Figure 2: Calibrated Cut-off Distribution



discretized normal distribution with a uniform distribution such that cut-offs would be drawn from the function: $f(x) = \mathcal{N}(190, 20) + \mathcal{U}\{1, 300\}$.[11] Figure 2 depicts the associated cumulative distribution function which illustrates the probability of being cut off in the matrix task as a function of time. Dotted lines mark the times that the average as well as the slowest subject took to successfully complete the matrix task in the informal pretest. These times were used as benchmarks for our calibration. We calibrated the cut-off distribution with the following two objectives in mind:

- *Minimizing data loss.*

  Early cut-offs are associated with data loss because neither is the time data of a particular subject rich enough to identify procrastination nor do we obtain choice data in the subsequent dictator game. To minimize data loss, our cut-off distribution is shifted to the right. Recall that in the pretest, subjects needed on average 104s to succeed in the matrix task. But even up to the 150 seconds mark, the cumulative probability of being cut off in our experiment was merely 12% (after which it increased more rapidly).

- *Minimizing selection effects.*

  Some of our hypotheses derived in Section 3.2 are tested by comparing aggregate choice behavior in the dictator game stage between our ND and PD treatments. For these tests to be reliable, we have to rule out the possibility

[11] We refrained from shifting the distribution to the utmost right and added a uniformly distributed element to it to preclude subjects from working out the underlying distribution ex-post e.g. through communication with fellow participants.

that our cut-off mechanism changed the composition of our PD compared to our ND samples. This would be the case e.g. if one assumed cut off subjects to be overly selfish or other-regarding. The shift of our cut-off distribution was specifically motivated to handle this potential concern. Since, for most of the cases, a cut-off would not occur until very late, we made it very difficult for subjects to successfully self-deceive. A cut-off could only be enforced through excessive procrastination which we assumed to be incompatible with maintaining the perception of irresponsibility. Consequently, we expected most subjects in our experiment to finish the task (with only few being cut off). In Section 4.2 we confirm that this was indeed the case in our experiment.

On the second dimension of our factorial treatment design, we varied whether subjects could communicate with their counterpart before entering the matrix solving stage. In the communication stage, we allowed subjects to exchange pre-formulated messages. Within a group, one subject was randomly chosen to send the first message by choosing one of the following alternatives:

Message 1: "I promise to do my best to implement Option B, if you promise to do the same."

Message 2: "I don't want to commit myself to anything."

The second subject could then reply by choosing between:

Message 1: "I promise to do my best to implement Option B."

Message 2: "I don't want to commit myself to anything."

Payoffs were calibrated providing an equality as well as total earnings maximizing argument in favor of option B(£6, £6) over the opportunistic option A(£10, £0). We presumed that subjects would use the communication stage to exchange promises as a means to achieve cooperation on the former allocation.

The experiment was designed such that the aforementioned deniability manipulations took place only after the communication stage had concluded. This means that, at the time when subjects exchanged messages, they did not know whether they would be assigned to the *No Deniability* or *Plausible Deniability* condition. It was only after messages had been exchanged and the communication stage had concluded that they learned which condition applied to them.[12] By this means, we were able to vary by treatment whether deniability was possible or not without systematically influencing the content of exchanged messages.

[12] In the instructions, we only provide minimal information about the cut-off mechanism. Subjects are told that additional details would follow in the later course of the experiment. After the conclusion of the communication stage, treatment-specific details regarding the cut-off mechanism were read out aloud by the experimenter. Scripts can be found in Appendix B.3.

By comparing the marginal effect of *adding* communication (and thereby promise exchange) to our existing ND and PD conditions, our experiment allows to shed light on the relevance of image concerns particular to promise keeping. We also collected data on subjects' beliefs about the behavior and expectations of their counterpart to investigate whether any observed effects of our treatment variables are correctly anticipated by subjects to affect behavior more generally. Subjects' second-order beliefs which serve as the conventional measure of guilt in the literature are moreover informative in assessing the extent to which our results could potentially be accounted for by an aversion to guilt as opposed a "pure" image concern.

Belief elicitation took place after the conclusion of the dictator game stage, but before roles and payoffs were assigned. Table 2 reproduces what subjects saw on their screen. Subjects were first asked how likely they thought it was that their counterpart (i) succeeded in the matrix task, and (ii) chose the generous allocation (conditional on having succeeded). Subsequently and on a separate screen, we elicited subjects' second-order beliefs by asking them to second-guess their counterpart's responses to the aforementioned questions. Subjects were paid a flat payment of £1 for providing their initial responses. We decided not to incentivize the accuracy of these responses because the conventional approach would have required us to reveal information on a counterpart's true behavior (which our PD conditions were specifically designed to avoid). This constraint did not apply to the elicitation of second-order beliefs which were formed upon a counterpart's beliefs rather than behavior. Consequently, we incentivized the accuracy of subjects' second-order beliefs by awarding a bonus of £1 for every response that was correctly matched.

Table 2: Belief Elicitation

<table>
<tr><td colspan="6">How likely do you think it is that your counterpart correctly solved 15 matrices on time?</td></tr>
<tr><td></td><td>Very Likely</td><td>Somewhat Likely</td><td>50-50</td><td>Somewhat Unlikely</td><td>Very Unlikely</td></tr>
<tr><td>Your Guess</td><td>○</td><td>○</td><td>○</td><td>○</td><td>○</td></tr>
</table>

<table>
<tr><td colspan="6">Now, assume your counterpart correctly solved 15 matrices on time and made a choice between Options A and B. How likely do you think it is that your counterpart chose Option B (£6, £6)?</td></tr>
<tr><td></td><td>Very Likely</td><td>Somewhat Likely</td><td>50-50</td><td>Somewhat Unlikely</td><td>Very Unlikely</td></tr>
<tr><td>Your Guess</td><td>○</td><td>○</td><td>○</td><td>○</td><td>○</td></tr>
</table>

We opted for a one-shot version of the game because we presumed that learning associated with repeated play would eventually reduce or even erase the scope for self-deception to be operative. To assist subjects in their understanding of the rules and processes of the experiment, we initiated a practice phase in which they were guided through the stages of the experiment, supplemented with detailed on-screen explanations. In the course of this practice phase, subjects were also able to work on scaled-down versions of the matrix task with computer simulated counterparts. A late cut-off round (60s) familiarized them with how the matrix task worked, followed by an early cut-off round (12s) which was meant to familiarize subjects with the cut-off mechanism and its consequences.[13] The practice phase concluded with a quiz to ensure that subjects understood the instructions and processes of the experiment.

3.2 Hypotheses

We start this section by stating a set of more general hypotheses about the contents and effects of exchanged messages before turning our attention to *image motivation* in particular.

Hyp. 1: *Subjects will use the communication stage to exchange promises.*

Since the focus of our paper is on promise keeping, it was our intention to induce high rates of promise exchange in our experiment. Although some subjects may want to avoid commitment[14], we expected promise induced cooperation on the other-regarding allocation to be appealing to many subjects due to its equal and total-earnings maximizing payoff properties. Moreover, our restrictive communication protocol with pre-formulated messages made promise exchange suggestive and erased any ambiguities surrounding the classification of messages oftentimes observed under protocols of free form communication.

Hyp. 2: *Generosity is higher in treatments featuring communication.*

It is a well-documented finding in the literature that promises are oftentimes kept, even in one-shot encounters and in the absence of punishment threats. According to the *commitment-based explanation* of promise keeping, people keep their promises because they have an intrinsic preference for keeping their word. Consequently, we would expect some promise keeping to occur (and thereby increase

[13] To make it more apparent to subjects that a cut-off could be desirable, we programmed the computer to pick the opportunistic outcome in the early cut-off round. Thus, every subject experienced at least once that a cut-off could result in the implementation of the opportunistic outcome on the subject's behalf.

[14] Think of subjects who prefer keeping promises but expect their counterpart to make opportunistic promises which are bound to be broken. It is then rational for a subject not to engage in mutual promise exchange.

generosity) under *both* our No Deniability *and* Plausible Deniability conditions.

Hyp. 3: *Beliefs about generosity are higher in treatments featuring communication.*

Hypothesis 3 naturally follows from hypothesis 2 under the assumption that subjects believe the underlying theory. It is the process of promises feeding expectations which also underlies the *expectations-based* explanation of promise keeping based on guilt and according to which people dislike letting others down on their promise-induced expectations.

We next turn our attention to understudied explanations of promise keeping which rest on the relevance of social- and self-image concerns. We contribute to the literature by assessing the empirical relevance of these explanations in our experiment.

3.2.1 Social-Image Concerns

From the stream of research discussed in subsection 2.2, we know that subjects care about how they and their actions are being *perceived by others*. The assumption is that being perceived in a negative light by others imposes a psychological cost on the subject. Recall that the cut-off mechanism in our Plausible Deniability conditions could serve as an excuse for selfish outcomes. Since a subject's counterpart cannot ascertain how an outcome came about, we would expect social-image concerns to be mitigated in these treatments. Conversely, subjects in the No Deniability conditions cannot use early cut-offs as excuses for selfish outcomes. Therefore, we would expect social-image concerns to be amplified in these treatments.

The image concern that we are interested in arises over promise keeping. To rule out an alternative image concern, namely that of being perceived as *selfish* (or, greedy, unfair), we also conducted treatments where communication opportunities were removed. Our identification strategy is to compare the relative effectiveness of *adding* communication within our No Deniability as compared to our Plausible Deniability conditions.[15] Under the assumption that there exist subjects who suffer an image cost of being perceived as a promise breaker by others, we would expect communication to be more effective under No Deniability compared to Plausible Deniability.

Hyp. 4: *Communication increases generosity more strongly under ND than PD.*

Again, given that subjects believe the underlying theory behind hypothesis 4, they will anticipate social image concerns to be amplified in others under ND compared to PD. We can state the following hypothesis:

[15] A similar strategy was applied by Schütte and Thoma (2014) in the context of a trust game.

Hyp. 5: *Communication increases beliefs about generosity more strongly under ND than PD.*

3.2.2 Self-Image Concerns

Our last set of hypotheses derive from the literature on self-image concerns which we discussed in subsection 2.3. The message of this stream of research is that people desire to *perceive the self* in a favorable light. Psychological discomfort can be experienced when behavior threatens a person's self-concept. One way of maintaining a desired self-concept in light of opportunistic temptation is to engage in self-deception.

Our idea is that self-image concerns may be relevant for promise-keeping. As a consequence, the strength of promises may be diluted in environments which allow people to self-deceive about the existence of a broken promise. In our experiment, a subject who feels compelled to live up to her promise in order not to threaten her self-image may want to procrastinate in the matrix task in the hope of being cut off by the computer. A cut-off results in a fair chance of obtaining the opportunistic outcome whilst allowing to maintain the perception of not having acted against one's promise. Recall that we conducted a control treatment where no counterpart was involved and successful completion of the matrix task allowed the dictator to choose her own payoff only. The assumption behind this treatment was that image related incentives for procrastination would be removed, thereby allowing us to obtain an unbiased approximation of subjects' ability in the matrix task against which to compare performances in our Plausible Deniability treatments (where we assumed such incentives to be present).

As argued before, image concerns can relate to *outcomes* (perceiving the self as selfish) and/or the *process* by which outcomes are reached (perceiving the self as a promise breaker). Considering our No Communication conditions first where only the former concern was at stake, we would expect self-deceivers in treatment NC_PD to have worked significantly more slowly and/or to have committed more errors compared to subjects in our control treatment.

Hyp. 6: *Matrix task performance is worse under NC_PD than CONTROL.*

In treatment C_PD, we assume that the additional self-image concern stemming from promise making induces higher generosity. This provides yet more subjects with an incentive to self-deceive and to procrastinate in the matrix task. Consequently, we predict matrix task performance in treatment C_PD to be worse compared to treatments NC_PD (and CONTROL).

Hyp. 7: *Matrix task performance is worse under C_PD than NC_PD.*

Recall that beliefs about generosity are expected to be higher in conditions featuring a communication stage. If anything, guilt aversion would therefore predict *more* instead of less effort in the matrix task which would bias our results *against* hypothesis 7.

3.3 Procedures

The experiment was programmed in z-Tree (Fischbacher, 2007) and conducted in the *Laboratory for Economic and Decision Research* (LEDR) at the University of East Anglia. A total of 254 participants recruited from the local student population took part in the study. We ran 16 sessions in March 2018, each of which lasting between 35-45 minutes, depending on the treatment. We ran more PD sessions to compensate for the small data loss expected to occur by early cut-offs. The number of sessions per treatment were: 3 x NC_ND, 3 x C_ND, 4 x NC_PD, 4 x C_PD, 2 x CONTROL. 16 subjects took part in each session, except for one NC_PD session where only 14 subjects turned up. Average earnings were £10, with a minimum earning of £4 and a maximum earning of £16 (including a £3 participation fee).

Upon arrival, participants were randomly assigned to computer terminals by drawing their desk number. Each computer was located in a separate cubicle which inhibited visual interaction or communication. Anonymity amongst participants was secured because at no point during or after the experiment did any participant receive identifying information about his or her peers. We also took great care in the instructions emphasizing that the experimenter would not be able to link the generated data to any participant as a person. Participants received a hard copy of the instructions and were asked to follow along as the experimenter read the instructions out aloud. Clarifications were provided on an individual basis. Participants were asked to answer a set of five control questions after the completion of the practice phase to ensure that they understood the instructions and processes of the experiment. Two further control questions were displayed after details regarding the cut-off mechanism were publicly announced by the experimenter. The experiment concluded with a brief questionnaire asking for socio-demographic characteristics and an assessment of the difficulty of the experimental tasks. Privacy was guaranteed during the payment phase by asking participants to individually collect their final earnings from an experimental assistant at the end of the experiment.

4 Results

Section 4.1 looks at the communication contents of our experiment. Section 4.2 analyses the effects of communication, focusing on social-image effects in Section 4.2.1 and self-image effects in Section 4.2.2.

4.1 Communication Contents

Table 3 summarizes the observed message profiles (pairs of messages) broken down by treatment condition. Recall that by design, our deniability manipulations took place only after the communication stage concluded. Up to that point, the experimental protocol including the instructions was identical. We would therefore expect no significant differences in the contents of exchanged messages across treatments. This is confirmed by our data which is why we henceforth refer to the pooled data provided in the last column of Table 3.

By looking at the first two rows of Table 3, we can see that 46 out of 56 first-movers (82.1%) sent the cooperative message 1 stating a promise intent. Amongst the 46 second-movers who received a promise intent, 42 (91.3%) reciprocated with a promise thereby establishing mutual promise exchange. Unsurprisingly, amongst the few cases (10 out of 56) where first-movers refrained from proposing a mutual exchange of promises by stating that they do not want to commit themselves, the majority of second-movers (8 out of 10) decided not to commit either. Two subjects decided to commit despite not having received an intention to commit by their counterpart. In line with hypothesis 1, we can state the following result:

Result 1. *Most pairs of subjects (75%) used communication to exchange promises.*

Table 3: Overview of Message Profiles by Treatment

	By Treatment			Pooled
$\text{Message}_{\text{F-Mover}}/\text{Messages}_{\text{S-Mover}}$	C_ND	C_PD	Z-stat.[a] (p-value)	C_ND + C_PD
Promise Intent/Promise	17/24 (70.8%)	25/32 (78.1%)	-0.624 (0.533)	42/56 (75%)
Promise Intent/No Commitment	3/24 (12.5%)	1/32 (3.1%)	1.348 (0.178)	4/56 (7.1%)
No Commitment/Promise	1/24 (4.2%)	1/32 (3.1%)	0.208 (0.835)	2/56 (3.6%)
No Commitment/No Commitment	3/24 (12.5%)	5/32 (15.6%)	-0.331 (0.741)	8/56 (14.3%)

[a] The Z-statistic reflects two-tailed tests of differences in proportions.

4.2 Communication Effects

Having established that subjects used the communication stage to exchange promises, we can investigate *whether* and *by what means* promise exchange increased generosity in our communication treatments.

Our analysis is based upon subjects who successfully completed the matrix task and for which choice data in the dictator game is available. Discarding subjects who were cut off before the completion of the task may raise self-selection concerns. As discussed before, we designed our experiment to minimize these concerns. As expected, the proportions of subjects who were cut off in our Plausible Deniability conditions were small: 6/64 (9.4%) in treatment C_PD, 9/62 (14.5%) in treatment NC_PD, and 4/32 (12.5%) in treatment CONTROL. Moreover, if selection issues were present in the sense that procrastinators successfully managed to enforce a cut-off, we would expect the proportion of cut-offs to be higher in treatments C_PD and NC_PD (where incentives for procrastination were present) compared to treatment CONTROL (where incentives for procrastination were removed). This however was not the case according to pairwise Fisher's exact tests (p = 0.441 and p = 0.529 respectively, one-tailed). Appendix C.1 provides details on cut-off times and matrix task progress of subjects who were cut off from the task before completion. It is noteworthy that a considerable proportion of these subjects (11/21 or 52.4%) did not manage to solve a single matrix in the practice stage, suggesting that our cut-off mechanism sorted out subjects who lacked a sufficient understanding of the task.

4.2.1 Social-Image Effects

All data referred to in this section is also subsumed in Table 4 which provides detailed summary statistics on the frequency of cut-offs, on choices in the dictator game stage, and on reported beliefs, all broken down by treatment and, if applicable, by communication history. Unless otherwise stated, reported Z statistics reflect tests of proportions (see Glasnapp and Poggio, 1985) when comparing choice data in the dictator game and Wilcoxon rank sum tests (see Siegel and Castellan, 1988) when comparing reported belief data.

Figure 3 summarizes our main findings by depicting the proportions of subjects choosing the generous allocation for each treatment separately. Our communication protocol is found to be effective in increasing generous allocations both under conditions of No Deniability (20.8% vs. 58.7%; Z = -3.756, $p < 0.01$, one-tailed) as well as under conditions of Plausible Deniability (18.9% vs. 37.9%; Z = -2.215, p = 0.013, one-tailed). A strong effect of communication is in line with hypothesis 2 and research discussed in subsection 2.1. We state the following result:

Result 2. *Generosity is higher in treatments featuring communication.*

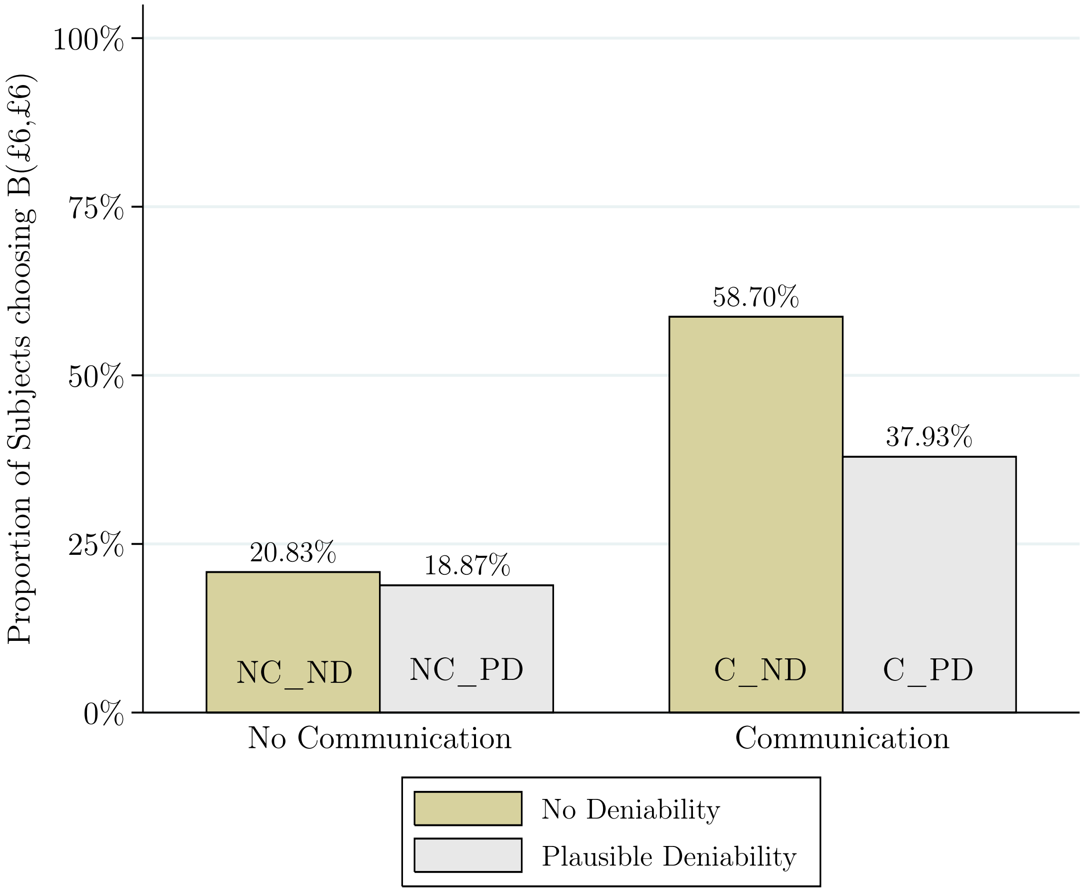
It is evident from our data however that communication has a stronger effect on generosity under ND compared to PD which squares well with hypothesis 4 and the idea that subjects dislike being perceived as a promise breaker by others.

Result 3. *Communication increases generosity more strongly under ND than PD.*

Table 4: Summary Statistics

	n	Cut off n(%)	Generous n(%)	Selfish n(%)	Question 1 FO _Belief	Question 1 SO_Belief	Question 2 FO_Belief	Question 2 SO_Belief
Communication	**112**	**8(7.1%)**	**49(47.1%)**	**55(52.9%)**	**4.44**	**4.46**	**2.88**	**2.88**
C_ND	48	2(4.2%)	27(58.7%)	19(41.3%)	4.80	4.87	2.98	2.93
C_ND_PromiseEx.	34	1(2.9%)	25(75.8%)	8(24.2%)	4.79	4.85	3.24	3.21
C_ND_NoPromiseEx.	14	1(7.1%)	2(15.4%)	11(84.6%)	4.85	4.92	2.31	2.23
C_PD	64	6(9.4%)	22(37.9%)	36(62.1%)	4.16	4.14	2.81	2.84
C_PD_PromiseEx.	50	5(10.0%)	22(48.9%)	23(51.1%)	4.18	4.18	3.22	3.20
C_PD_NoPromiseEx.	14	1(7.1%)	0(0.0%)	13(100%)	4.08	4.00	1.38	1.62
No Communication	**110**	**9(8.2%)**	**20(19.8%)**	**81(80.2%)**	**4.50**	**4.47**	**2.26**	**2.23**
NC_ND	48	0(0.0%)	10(20.8%)	38(79.2%)	4.88	4.83	2.33	2.31
NC_PD	62	9(14.5%)	10(18.9%)	43(81.1%)	4.17	4.13	2.19	2.15
CONTROL	**32**	**4(12.5%)**	**1(3.6%)**	**27(96.4%)**	**n/a**	**n/a**	**n/a**	**n/a**

Figure 3: Proportions of Generous Choices between Treatments



To see this, it is sufficient to realize that our deniability manipulation affected generosity within our Communication conditions only, not however within conditions where no communication was possible. Looking at our Communication conditions first, we observe that PD significantly decreased the proportion of subjects choosing the generous allocation from 58.7% to 37.9% (Z = 2.107, p = 0.018, one-tailed). Considering promise keeping proportions in particular as presented in Table 4, we observe a significant decline from 75.8% in treatment C_ND to 48.9% in treatment C_PD (Z = 2.396, $p < 0.01$, one-tailed). Inspecting our No Communication conditions reveals that PD decreased the proportion of generous allocations by merely two percentage points. Although the effect goes in the anticipated direction, the difference is insignificant (Z = 0.248, p = 0.402, one-tailed). It appears that subjects in the No Communication treatments were not particularly bothered about the transparency of their decisions. Or, put differently, purely outcome based image concerns (such as being perceived as selfish, egoistic, or unfair) seem not to have played a major role in our experiment. On the contrary, our results lend support to the existence of an image concern particular to promise keeping per se.

Do subjects predict the effects of our treatment variables on their counterpart's behavior? We collected data on subjects' beliefs about their counterpart to answer this question. Responses were submitted on a 5 point Likert scale ranging from 1 (="very unlikely") to 5 (="very likely"). As illustrated earlier in Table 2, we first asked subjects how likely they thought it was that their counterpart succeeded in the matrix task. The purpose of asking this first question was to check whether our deniability manipulations were successful in diffusing a counterpart's perceived responsibility for outcomes. As is evident from the data provided in Table 4, this was indeed the case. Plausible deniability decreased average first-order beliefs relating to question 1 within both our Communication (4.80 vs. 4.16; Z = 4.623, $p < 0.01$, one-tailed) and No Communication (4.88 vs. 4.17; Z = 4.808, $p < 0.01$, one-tailed) conditions. The same pattern holds for second-order beliefs. Allowing subjects to communicate, on the other hand, had no impact on a subject's belief about their counterpart's success in the matrix task.

Looking at first-order responses to question 2, we can see that communication and the exchange of promises raised subjects' *own* beliefs about a counterpart's generosity (2.26 vs. 2.88; Z = -3.488, $p < 0.01$, one-tailed). On top of that, communication was correctly predicted by subjects to also move their *counterparts'* beliefs about the subjects' own generosity as evidenced by subjects' second-order beliefs (2.23 vs. 2.88; Z = -3.592, $p < 0.01$, one-tailed). In line with hypothesis 3, we state the following result:

Result 4. *Beliefs about generosity are higher in our communication treatments. This suggests that subjects anticipated an effect of promise exchange on generosity.*

Comparing subjects' first-order responses to question 2 between our deniability conditions allows us to investigate whether subjects also anticipated their counterpart to exploit the diffusion of responsibility inherent in our PD conditions. Relatedly, subjects' second-order responses are informative as to whether subjects anticipated their counterpart to anticipate such an effect to be present. In light of the fact that we did find an effect of deniability on behavior as stated in result 3, it is surprising that subjects appear not to have anticipated deniability to matter to others. In the case of subjects' first-order beliefs (and equivalently so for second-order beliefs), we observe no statistical differences between our deniability conditions. This result holds both within our No Communication conditions (2.33 vs. 2.19; Z = 0.762, p = 0.446, two-tailed) and within our Communication conditions (2.98 vs. 2.81; Z = 0.607, p = 0.544, two-tailed).

Result 5. *The effect of communication on beliefs does not differ under ND and PD. This suggests that subjects failed to anticipate promise keeping to be sensitive to our deniability manipulations.*

It is interesting to notice that guilt aversion – whilst providing a possible explanation (through result 4) for some of the generosity we observe – can be ruled out as an explanation for the differences that we observe between our deniability manipulations. We observe higher generosity in treatment C_ND than C_PD despite there being no significant differences in subjects' reported beliefs about generous behavior between these treatments. Our findings appear to be consistent with a "pure" image concern according to which subjects dislike being perceived as a promise breaker by others. Appendix A.2 reproduces our results by providing regression-equivalents to the non-parametric tests reported in this section.

4.3 Self-Image Effects

Recall that despite being able to exploit deniability in treatment C_PD, a considerable proportion of subjects (22/45 or 48.9%) honored their promise. Conventional theories of promise keeping would argue that this effect is due to either an intrinsic preference for promise keeping (Vanberg, 2008), or an aversion to letting promisees down on their payoff expectations (Charness and Dufwenberg, 2006). Even social image concerns could still be prevalent under the assumption that our PD treatments mitigated instead of fully erased perceived responsibility. An alternative explanation which has yet received little attention in the literature on promise keeping is that subjects honor their word to maintain their *self-image* as an honest person.

If self-image concerns contribute to the effectiveness of promises, we would expect its effect to be mitigated in environments which allow subjects to self-deceive about the cause of a broken promise. We hypothesized that self-deception in our experiment would take the form of subjects procrastinating in the matrix task to delegate their choice to the computer.

To obtain a benchmark for subjects' abilities in the matrix task against which to compare performances in our plausible deniability treatments, we conducted our control treatment which erased incentives for procrastination. The following analysis is based on a comparison of performances in the matrix task observed between treatments C_PD, NC_PD, and CONTROL.

Table 5: Success Times and Accuracy in the Matrix Task across Treatments

Treatment	n	Cut off n(%)	Time15 mean/median	Incorrect15 mean/median
NC_PD	62	9(14.5%)	102s/102s	1.49/1
C_PD	64	6(9.4%)	103s/100s	1.22/1
CONTROL	32	4(12.5%)	111s/104s	1.29/1

Figure 4: Cumulative Distribution Functions of Success Times

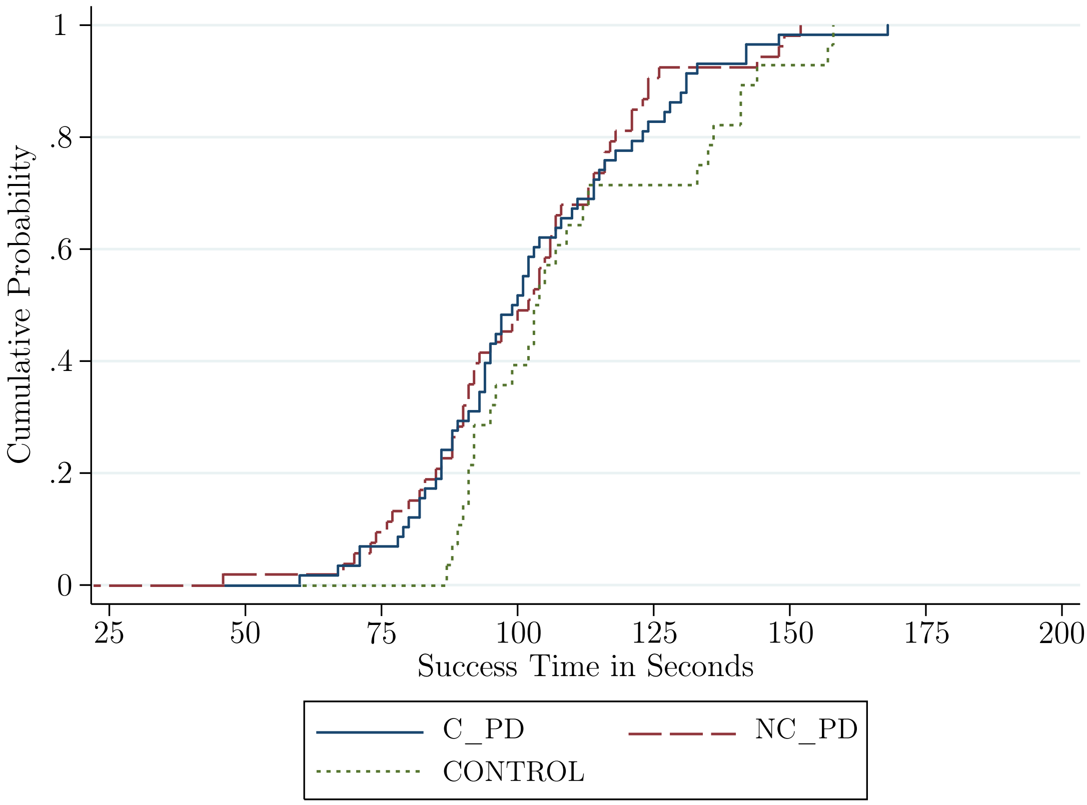
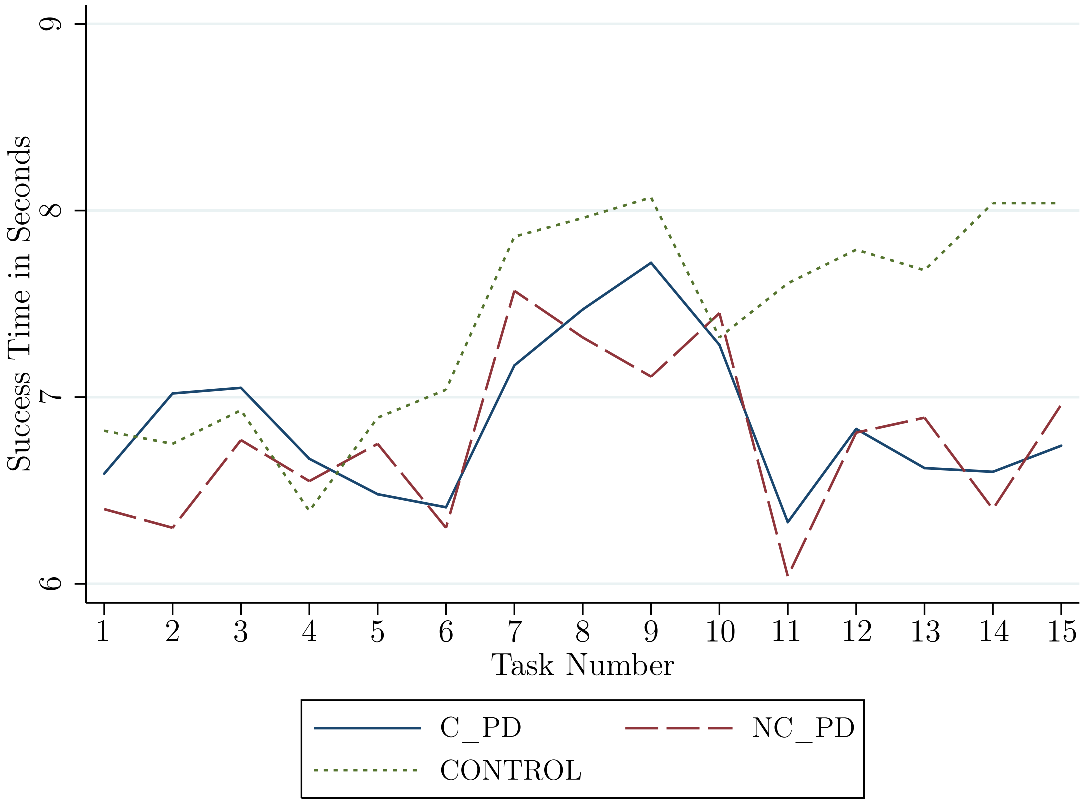


Table 5 reports summary statistics on the speed and accuracy with which subjects solved the target amount of 15 matrices.[16] Figure 4 provides the associated cumulative distribution functions of success times across treatments. If subjects procrastinated in our main treatments, we would expect the respective CDF's to lie further to the right compared to our control treatment where incentives for procrastination were removed. On the contrary, we observe the opposite. It appears that subjects in our main treatments performed even better than subjects in our control treatment which is particularly pronounced at the segment of high performing subjects. However, according to pairwise Kolmogorov-Smirnoff tests, the distributions of treatments C_PD and NC_PD do not differ significantly from CONTROL ($p = 0.157$ and 0.227, respectively).

We also looked at within-subject variation of performances in the matrix task. It is possible that procrastination would take the form of subjects slowing down on the task the closer they approach the target amount of 15 matrices. Figure 5 depicts for every treatment separately the average time spent on each of the 15

[16] We continue to condition our analysis on the sample of subjects who have not been cut off. Recall our previous discussion on p. 18 and Appendix A.1 for a justification of this approach. An advantage of doing so is that our cut-off mechanism simultaneously sorted out subjects who lacked sufficient understanding of the task. To keep these subjects in our sample would have made it complicated to discern motivated procrastination from delay due to misunderstanding.

Figure 5: Average Times Taken per Task



tasks. Again, eye-balling the results suggests that subjects in our main treatments performed better than subjects in the control treatment.

We ran a random effects panel model estimation to quantify what is observed in Figure 5. Results are presented in Table 6. Our dependent variable is the natural logarithm of the time (in seconds) taken by a subject to solve a given task. *TREAT* is a dummy distinguishing our treatment conditions with CONTROL being the reference category. *TASK_N* is the task number allowing us to measure changes in performance over time. We also include an interaction term between *TREAT* and *TASK_N* to allow performance changes to be treatment specific. The coefficient for *TASK_N* is positive and significant suggesting that subjects in our control treatment exhibit performance reductions as they move through the tasks. Such an effect could be due to boredom, or fatigue. On the contrary, no time trend is observed in treatments NC_PD and C_PD. This is evident from the negative coefficients of our interaction terms which are significant and fully compensate the negative time trend observed in our control treatment. Overall, performance in the matrix task appears to be worse in our control treatment with there being no difference between treatments C_PD and NC_PD. This result contradicts hypotheses 6 and 7 and lets us conclude with:

Result 6. *We find no evidence of procrastination in treatments NC_PD and C_PD.*

Table 6: Random Effects Panel Model Estimations

Dep. Variable *LN_ TIME*	Coef.	Robust[a] Std. Error	Z	p-value
TREAT				
NC_ PD	-0.015	0.050	-0.30	0.762
C_ PD	0.023	0.050	0.47	0.641
TASK_ N	0.012	0.004	3.15	0.002
TREAT × *TASK_ N*				
NC_ PD	-0.010	0.005	-2.23	0.026
C_ PD	-0.012	0.004	-2.82	0.005
_ CONS	1.847	0.041	45.31	0.000
		Prob > chi2		0.013
		R-Squared		0.015
		Number of Groups		139
		Number of Observations		2085

[a] Standard errors are clustered on the subject level.

5 Discussion

Similar to previous studies looking at the effectiveness of non-binding verbal commitments (or, “cheap talk”), we observe strong effects of communication on cooperative behavior. It is noteworthy that the effects that we observe originate from a rather reserved protocol of pre-formulated message exchange which is commonly perceived to be less powerful than free-form communication (see e.g. Charness and Dufwenberg, 2010). We ascribe this result to the nature of our experimental protocol . Since we generate promise exchange in a dictator instead of a trust game framework, our environment is less susceptible to reciprocity effects which usually generate significant rates of trustworthiness in baseline conditions and thereby limit the scope for treatment effects to be detectable. The idea for this design feature goes back to Vanberg (2008)’s random dictatorship game. Our results suggest that this protocol may be of interest to researchers who prefer to resort to pre-formulated message exchange without making compromises on the effectiveness of promises, or those who are concerned about “ceiling effects” in trust game studies.

A separate examination of the effectiveness of communication under No Deniability as compared to Plausible Deniability revealed that promise keeping was sensitive to whether a promisee could undoubtedly blame the promisor for outcomes. Note that the observed effect cannot be attributed to an intrinsic preference which underlies the commitment-based explanation of promise keeping. This theory predicts promise keeping to be independent of image concerns. Our analysis was moreover able to rule out alternative explanations such as an image concern of being perceived as selfish, or an aversion to guilt. We also judge it unlikely that our results were driven by experimenter observability or demand because (i) the presence of the experimenter was not altered between treatments, and (ii) our treatment manipulations required only subtle changes to the experimental protocol. Instead, our results square well with the hypothesis that promise keeping is partly driven by subjects' aversion to being perceived as a promise breaker by their counterpart.

An interesting finding is the observation that subjects appear not to have correctly anticipated their counterpart to be sensitive to our deniability manipulations. This is surprising, given that subjects themselves did respond to the increased transparency embedded in our ND conditions by keeping their promises more often. It is possible that the emotion of shame, whilst being an important factor of a subject's own decision making process, is underestimated to play as important a role in others' behavior. Under this premise, efforts in de-biasing subjects could be promising e.g. when it comes to the initiation of relationships based on trust.

Albeit to a lesser extent, promises remained to be effective even within our Plausible Deniability conditions. Both the commitment-based and the expectations-based explanations of promise keeping provide potential candidates for explaining this finding and our experiment was not designed to discern the empirical relevance of these theories from one another. Instead, we focused on a plausible alternative explanation of promise keeping which stems from the idea that subjects keep their promises in order not to threaten their self-image. This theory gave rise to the hypothesis that subjects would engage in self-deception – which would take the form of procrastination in the matrix task – to hide a reluctance to keep promises. We tested this hypothesis and report a null result.

One way of interpreting our null result is to take it as corroborating evidence of the strength of promises: subjects did not self-deceive because they truly desired to live up to their promise. At the same time, our result may call into question the generalizability of evidence supporting self-deception in dictator game studies to morally richer environments, as similarly pointed out by van der Weele et al. (2014, p. 262). The authors implement a cut-off mechanism to investigate the robustness of reciprocal behavior and likewise report a null result. There are caveats in order here, however.

Firstly, Regner (2018) reports a positive result observing that subjects do use the cut-off mechanism to avoid reciprocating others's kindness under different payoff calibrations of the trust game. He points out that the lack of a treatment effect in van der Weele et al. (2014) could be attributed to a ceiling effect stemming from the high proportion of selfish decisions (62.5%) observed in their baseline. Whilst a ceiling effect could have also been at work in our No Communication treatments, it is less likely that the same applied to our Communication treatments where the proportion of selfish allocations in our ND baseline was merely 41.3%.

Lastly, it is important to point out differences in the way we designed our experiment as compared to the aforementioned studies and in particular compared to the seminal paper by Dana, Weber and Kuang (2007). Whereas in their study, self-deception required subjects to deliberately wait for the computer to intervene, in our study subjects could delegate their decision in a more subtle and inconspicuous way by means of procrastination in an *active* task. One could argue that our design is less susceptible to demand effects and therefore provides a more natural testing ground for self-deception. At the same time, our experiment is more complex. It is possible that the additional complexity of our experiment made it more difficult for subjects to fully process the "exploitability" of our cut-off mechanism. However, as discussed in the design section of our experiment, we initiated a practice phase to assist subjects' general understanding of our game. In the course of this practice phase, we also exposed subjects to outcomes which hinted the possible desirability of being cut off in our experiment.

6 Conclusion

Trust is oftentimes referred to as the glue to social capital formation. Although its efficiency enhancing nature is desirable, trust carries the risk of betrayal. Communication and the exchange of promises are amongst the most prominent mechanisms to promote trust.

Our experiment was specifically set out to assess the relevance of two under-studied explanations of promise keeping, namely social- and self-image concerns. We observe strong evidence of social-image concerns in treatments which feature ex-ante opportunities for promise exchange. Ruling out alternative explanations, our results are consistent with subject exhibiting an aversion to being perceived as a promise breaker by others. Surprisingly, subjects seem not to anticipate social-image concerns to be present in others.

Our test of self-image concerns yielded a null result: there is no evidence suggesting that subjects in our experiment engaged in self-deception to evade their promise-induced commitments. This resilience can be interpreted as corroborating evidence of the strength of promises.

Our study contributes to the literature on promise keeping by documenting the significance of social-image concerns and, to the best of our knowledge, by being the first to have tested for self-image concerns.

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. “Reference Points and Effort Provision.” *American Economic Review*, 101(2): 470–492.

Akerlof, George A., and Rachel E. Kranton. 2000. “Economics and Identity.” *The Quarterly Journal of Economics*, 115(3): 715–753.

Akerlof, George A., and Rachel E. Kranton. 2010. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being.* Princeton University Press.

Andreoni, James, and B. Douglas Bernheim. 2009. “Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects.” *Econometrica*, 77(5): 1607–1636.

Andreoni, James, and Ragan Petrie. 2004. “Public Goods Experiments without Confidentiality: A Glimpse into Fund-Raising.” *Journal of Public Economics*, 88(7): 1605–1623.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. “Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially.” *American Economic Review*, 99(1): 544–555.

Baumeister, Roy F., Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. 1998. “Ego Depletion: Is the Active Self a Limited Resource?” *Journal of Personality and Social Psychology*, 74(5): 1252–1265.

Bem, Daryl J. 1972. “Self-Perception Theory.” In *Advances in Experimental Social Psychology.* Vol. 6, 1–62. Elsevier.

Bénabou, Roland, and Jean Tirole. 2004. “Willpower and Personal Rules.” *Journal of Political Economy*, 112(4): 848–886.

Bénabou, Roland, and Jean Tirole. 2006. “Incentives and Prosocial Behavior.” *American Economic Review*, 96(5): 1652–1678.

Bénabou, Roland, and Jean Tirole. 2011. “Identity, Morals, and Taboos: Beliefs as Assets.” *The Quarterly Journal of Economics*, 126(2): 805–855.

Bhattacharya, Puja, and Arjun Sengupta. 2016. “Promises and Guilt.” Available at SSRN: https://ssrn.com/abstract=2904957.

Bodner, Ronit, and Drazen Prelec. 2003. “Self-Signaling and Diagnostic Utility in Everyday Decision Making.” *The Psychology of Economic Decisions*, 1: 105–26.

Bohnet, Iris, and Bruno S Frey. 1999. “The Sound of Silence in Prisoner’s Dilemma and Dictator Games.” *Journal of Economic Behavior & Organization*, 38(1): 43–57.

Bursztyn, Leonardo, and Robert Jensen. 2017. "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure." *Annual Review of Economics*, 9: 131–153.

Cain, Daylian M, Jason Dana, and George E Newman. 2014. "Giving versus Giving In." *Academy of Management Annals*, 8(1): 505–533.

Charness, Gary, and Martin Dufwenberg. 2006. "Promises and Partnership." *Econometrica*, 74(6): 1579–1601.

Charness, Gary, and Martin Dufwenberg. 2010. "Bare Promises: An experiment." *Economics Letters*, 107(2): 281–283.

Chaudhuri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics*, 14(1): 47–83.

Dana, Jason, Daylian M Cain, and Robyn M Dawes. 2006. "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games." *Organizational Behavior and Human Decision Processes*, 100(2): 193–201.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory*, 33(1): 67–80.

Deck, Cary, Maroš Servátka, and Steven Tucker. 2013. "An examination of the Effect of Messages on Cooperation under Double-blind and Single-blind Payoff Procedures." *Experimental Economics*, 16(4): 597–607.

DellaVigna, Stefano, John A List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *The Quarterly Journal of Economics*, 127(1): 1–56.

Di Bartolomeo, Giovanni, Martin Dufwenberg, Stefano Papa, and Francesco Passarelli. 2019. "Promises, Expectations & Causation." *Games and Economic Behavior*, 113: 137–146.

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. 2015. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about others' Altruism." *American Economic Review*, 105(11): 3416–3442.

Ederer, Florian, and Alexander Stremitzer. 2017. "Promises and Expectations." *Games and Economic Behavior*, 106: 161–178.

Ekström, Mathias. 2012. "Do Watching Eyes Affect Charitable Giving? Evidence from a Field Experiment." *Experimental Economics*, 15(3): 530–546.

Exley, Christine L. 2015. "Excusing Selfishness in Charitable Giving: The Role of Risk." *The Review of Economic Studies*, 83(2): 587–628.

Fehr, Ernst, and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3): 159–181.

Fischbacher, Urs. 2007. “z-Tree: Zurich Toolbox for Ready-made Economic Experiments.” *Experimental Economics*, 10(2): 171–178.

Fiske, Susan T. 2018. *Social Beings: Core Motives in Social Psychology.* 4th edition, Hoboken, NJ : John Wiley & Sons, Inc.

Glasnapp, Douglas R, and John P Poggio. 1985. *Essentials of Statistical Analysis for the Behavioral Sciences.* CE Merrill Pub. Co.

Greenberg, Adam Eric, Paul Smeets, and Lilia Zhurakhovska. 2015. “Promoting Truthful Communication through ex-post Disclosure.” Available at SSRN: https://ssrn.com/abstract=2544349.

Grossman, Zachary, and Joël J Van der Weele. 2017. “Self-Image and Willful Ignorance in Social Decisions.” *Journal of the European Economic Association*, 15(1): 173–217.

Haisley, Emily C, and Roberto A Weber. 2010. “Self-Serving Interpretations of Ambiguity in Other-Regarding Behavior.” *Games and Economic Behavior*, 68(2): 614–625.

Hamman, John R, George Loewenstein, and Roberto A Weber. 2010. “Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship.” *American Economic Review*, 100(4): 1826–1846.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L Smith. 1996. “Social Distance and Other-Regarding Behavior in Dictator Games.” *American Economic Review*, 86(3): 653–660.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. 1994. “Preferences, Property Rights, and Anonymity in Bargaining Games.” *Games and Economic Behavior*, 7(3): 346–380.

Ismayilov, Huseyn, and Jan Potters. 2016. “Why Do Promises Affect Trustworthiness, or Do They?” *Experimental Economics*, 19(2): 382–393.

Kriss, Peter H, Roberto A Weber, and Erte Xiao. 2016. “Turning a Blind Eye, but Not the other Cheek: On the Robustness of Costly Punishment.” *Journal of Economic Behavior & Organization*, 128: 159–177.

Lazear, Edward P, Ulrike Malmendier, and Roberto A Weber. 2012. “Sorting in Experiments with Application to Social Preferences.” *American Economic Journal: Applied Economics*, 4(1): 136–163.

Malmendier, Ulrike, Vera L te Velde, and Roberto A Weber. 2014. “Rethinking Reciprocity.” *Annual Review of Economics*, 6(1): 849–874.

Mazar, Nina, and Chen-Bo Zhong. 2010. “Do Green Products Make Us Better People?” *Psychological Science*, 21(4): 494–498.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." *Journal of Marketing Research*, 45(6): 633–644.

Merritt, Anna C, Daniel A Effron, and Benoît Monin. 2010. "Moral Self-Licensing: When Being Good Frees Us to Be Bad." *Social and Personality Psychology Compass*, 4(5): 344–357.

Mischkowski, Dorothee, Rebecca Stone, and Alexander Stremitzer. 2016. "Promises, Expectations, and Social Cooperation." Harvard Law School John M. Olin Center Discussion Paper 887.

Rege, Mari. 2004. "Social Norms and Private Provision of Public Goods." *Journal of Public Economic Theory*, 6(1): 65–77.

Rege, Mari, and Kjetil Telle. 2004. "The Impact of Social Approval and Framing on Cooperation in Public Good Situations." *Journal of Public Economics*, 88(7-8): 1625–1644.

Regner, Tobias. 2018. "Reciprocity under Moral Wiggle Room: Is it a Preference or a Constraint?" *Experimental Economics*, 1–14.

Schütte, Miriam, and Carmen Thoma. 2014. "Promises and Image Concerns." Munich Discussion Paper No, 2014–18.

Schwartz, Steven, Eric Spires, and Rick Young. 2018. "Why Do People Keep their Promises? A Further Investigation." *Experimental Economics*, 1–22.

Siegel, S., and N.J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill.

Soetevent, Adriaan R. 2005. "Anonymity in Giving in a Natural Context — A Field Experiment in 30 Churches." *Journal of Public Economics*, 89(11-12): 2301–2323.

Spiekermann, Kai, and Arne Weiss. 2016. "Objective and Subjective Compliance: A Norm-based Explanation of 'Moral Wiggle Room'." *Games and Economic Behavior*, 96: 170–183.

Tadelis, Steven. 2011. "The Power of Shame and the Rationality of Trust." Haas School of Business Working Paper 2011/3/2.

Vanberg, Christoph. 2008. "Why Do People Keep Their Promises? An Experimental Test of Two Explanations." *Econometrica*, 76(6): 1467–1480.

van der Weele, Joël J, Julija Kulisa, Michael Kosfeld, and Guido Friebel. 2014. "Resisting Moral Wiggle Room: How Robust is Reciprocal Behavior?" *American Economic Journal: Microeconomics*, 6(3): 256–264.

A Supplementary Data

A.1 Cut-offs and Task Progress

Table 7 lists subjects who were cut-off from the task before successfully solving the required number of 15 matrices. Overall, this was the case for 21 out of 254 (8.3%) subjects in our experiment. In the last column, we indicate whether or not a respective subject correctly solved any of the matrices of the practice phase of our experiment. This information may be informative as to whether or not a subject struggled understanding the task. For some subjects, this appears to have indeed been the case as is evident e.g. from subject #249 who made 96 mistakes in the control treatment. Another subject directly expressed to the experimenter confusion about how to solve a given matrix in the practice stage.

It appears that many of the cut-offs observed are consistent with delays due to confusion rather than procrastination. Examining the reported cut-off times and the progress of subjects who demonstrated understanding of the task, it does not appear to be the case that cut off subjects were reluctant to solve the task which further alleviates concerns that our results are affected by selection effects.

Table 7: Cut-offs and Task Performance

	Treatment	ID	Session	Cut-off Time	#Correct	#Incorrect	Solved Practice?
1.	NC_PD	52	9	176	12	4	No.
2.	NC_PD	57	9	23	3	0	Yes.
3.	NC_PD	60	9	193	0	6	No.
4.	NC_PD	61	9	25	5	0	Yes.
5.	NC_PD	72	10	35	3	1	No.
6.	NC_PD	73	10	65	9	0	Yes.
7.	NC_PD	101	12	38	1	4	No.
8.	NC_PD	103	12	113	12	2	Yes.
9.	NC_PD	104	12	22	3	0	Yes.
10.	C_ND	115	5	300	0	14	No.
11.	C_ND	152	7	300	0	7	No.
12.	C_PD	179	2	83	9	1	Yes.
13.	C_PD	182	2	78	10	0	Yes.
14.	C_PD	191	3	86	12	0	Yes.
15.	C_PD	195	3	41	5	0	Yes.
16.	C_PD	206	3	100	12	3	No.
17.	C_PD	220	4	165	6	8	No.
18.	CONTROL	225	8	175	13	3	No.
19.	CONTROL	242	16	66	4	2	No.
20.	CONTROL	246	16	106	14	1	Yes.
21.	CONTROL	249	16	171	2	96	No.

A.2 Regression Results

In Table 8 we report supplementary regression results supporting the conclusions derived from our non-parametric analyses reported in the main text. The dependent variable in models [1]-[2] is a dummy taking value 1 if the generous allocation was chosen, and 0 otherwise. In models [3]-[6], the dependent variable is the respective question 2 belief measured on a 5 point Likert scale. As independent variables we include dummies for our treatment conditions and the interaction thereof. What we find is that communication exerts a strong influence on generosity *and* on reported beliefs which is consistent with the idea that an effect of communication could partly be mediated through guilt aversion. The negative interaction term in models [1]-[2] moreover suggests that communication exerts a stronger effect on generosity within our No Deniability conditions. Interestingly and in line with our previous findings, this asymmetry is not mirrored by beliefs which is evident from the insignificant interaction term reported in models [3]-[6].

The result that communication and promise exchange affect behavior more strongly under No Deniability, coupled with the finding that beliefs were not affected, suggests that the effect of our deniability manipulations is unlikely to be attributed to an aversion to guilt. Instead, our findings are consistent with subjects exhibiting an aversion to being perceived as a promise breaker by others.

Table 8: Regression Results

	[1]	[2]	[3]	[4]	[5]	[6]
Model:	Probit	OLS	Ord. Logit	OLS	Ord. Logit	OLS
Dep. Variable:	Generous	Generous	FO-Belief	FO-Belief	SO-Belief	SO-Belief
Communication	1.032*** (0.173)	0.379*** (0.053)	0.904** (0.408)	0.645** (0.272)	0.870*** (0.203)	0.622*** (0.155)
Pl. Deniability	-0.071 (0.202)	-0.020 (0.057)	-0.224 (0.445)	-0.145 (0.287)	-0.252 (0.355)	-0.162 (0.239)
Communication × Pl. Deniability	-0.456* (0.236)	-0.188** (0.074)	0.006 (0.542)	-0.023 (0.351)	0.121 (0.437)	0.072 (0.294)
Constant	-0.812*** (0.163)	0.208*** (0.047)		2.333*** (0.188)		2.313*** (0.141)
(Pseudo) R-Squared	0.084	0.108	0.022	0.074	0.022	0.075
N	205	205	205	205	205	205

[a] All regressions cluster observations on the session level. Robust standard errors are reported in parentheses. *(**, ***): Coefficient significantly different from zero at the 10% (5%, 1%) level.

B Instructions

B.1 Main Treatments

Information in brackets [...] only applies to treatments featuring a communication stage. Otherwise, instructions are identical across our four main treatments. Subjects received information regarding the cut-off mechanism just before entering the matrix solving stage.

Instructions

Welcome to this experiment and thank you for participating. Please follow along carefully as the experimenter reads the instructions out aloud. The purpose of this experiment is to study how people make decisions in particular situations. You were awarded £3 for showing up on time. Your additional earnings in this experiment depend on the decisions you and other participants make during the experiment and on chance. At the end of the experiment, the entire amount will be paid to you *individually* and *privately* in cash by an assistant.

Please do not speak to other participants during the experiment and keep your phones switched off. If you have any questions at any time over the course of the experiment, please raise your hand and an experimenter will come to assist you.

Note that your behaviour in this experiment is recorded by the computer and stored in a database. The records of this database are anonymous, i.e. not traceable to you as a person. For accounting reasons only, you will be asked to fill in and sign a receipt of your earnings at the end of the experiment. To secure anonymity, these receipts will be kept entirely separate from any data on your behaviour generated in the experiment.

Please remain seated until you are individually asked by the experimenter to collect your final earnings at the end of the experiment.

The Experiment

At the beginning of the experiment, you will be paired with another randomly determined participant in the room who will from now on be called your **counterpart**. No participant will get to know the identity of his/her counterpart during or after the experiment.

All participants in this experiment are provided with the same set of instructions and will encounter the same stages as described below:

Stage 1: Matrix Task.

In stage 1 of the experiment, you will work on a matrix solving task. The task consists of counting *ones* (1s) in a series of matrices comprised of random 0s and 1s. A sample matrix is depicted in Figure 1 below.

Figure 1: Sample Matrix

0	1	1	1	0
0	0	0	0	0
1	0	0	1	1
1	1	1	1	0
1	1	0	1	0

You will be able to work on this task for a *maximum* of 300 seconds (5 minutes). Importantly, you will be timed-out by the computer at some point during this time interval. If this happens, the matrix task will end. You will then be asked to work on a follow-up task for the remainder of the 300 seconds.

All participants will be provided with additional details regarding the time-out mechanism in the later course of the experiment.

Outcomes in the matrix task (not however in the follow-up task) have direct consequences for the decision environment in stage 2 of the experiment:

- If you correctly solve at least 15 matrices before you are timed-out by the computer, you will be able make a decision in stage 2 of the experiment.

- If you do not correctly solve at least 15 matrices before you are timed-out by the computer, you will *not* be able to make a decision in stage 2 of the experiment.

After the conclusion of the matrix and follow-up task (i.e. after 300 seconds), you will move forward to stage 2 of the experiment.

Stage 2: Decision Stage.

In stage 2 of the experiment, you will *potentially* be able to choose between two options. Your choice indicates how you would like to allocate money between you and your counterpart. The possible options are:

- Option A: **£10** to you and **£0** to your counterpart.

- Option B: **£6** to you and **£6** to your counterpart.

If you succeeded in solving at least 15 matrices in stage 1 of the experiment, *you yourself* will choose between Option A and Option B.

If you did not succeed in solving at least 15 matrices in stage 1 of the experiment, *the computer instead* will randomly choose between Option A and Option B with equal probability.

The resulting option (A or B) will be called your **individual stage 2 outcome**. You will know whether the outcome of your stage 2 was determined by your own choice or by the choice of the computer. Your counterpart, however, *will not* know how your stage 2 outcome came about.

Determining the Relevant Player.

After both you and your counterpart have individually completed the stages above, one of you will be randomly determined by the computer to become the **Relevant Player**.

If you become the Relevant Player, your stage 2 outcome will be implemented. If you *do not* become the Relevant Player, your stage 2 outcome *will not* be implemented and will therefore have *no consequences* for payoffs in the experiment. In this case, your payoffs will solely be determined by the stage 2 outcome of your counterpart because he or she was assigned the role of Relevant Player.

Note that it is *equally likely* that you or your counterpart will be assigned the role of Relevant Player.

[Communication Phase.

Before stage 1 of the experiment starts you will be asked to choose one of two predefined messages to be sent to your counterpart.

Note that at this point, you will not know which of you will become the Relevant Player in the experiment. You will receive this information only at the end of the experiment.

Messages will be exchanged sequentially. One participant will be randomly determined to send the first message by choosing one of the following options:

Message 1: *"I promise to do my best to implement Option B, if you promise to do the same."*

Message 2: *"I don't want to commit myself to anything."*

The second participant in a group will then be asked to reply by choosing one of the following options:

Message 1: *"I promise to do my best to implement Option B."*

Message 2: *"I don't want to commit myself to anything."*

Importantly, the sequence in which messages are exchanged is randomly determined and not related to the assignment of roles at the end of the experiment. *Again, this means that at the time when you exchange messages with your counterpart, you will not know which of you will be assigned the role of Relevant Player.*]

Bonus: Guessing.

At certain points during the experiment, you will have the opportunity to earn small amounts of additional money by guessing decisions and outcomes in the experiment. You will learn more about this during the experiment.

Practice.

We will now briefly guide you through the decision stages in order for you to get a better understanding of the interface and processes of this experiment. You will also be able to familiarise yourself with the matrix task. We will conclude the practice phase with a quiz to check your understanding.

Please follow along on screen.

B.2 Control Treatment

In this treatment, we erased the recipient role. All other features of this treatment including the instructions and experimental procedures closely followed treatment NC_PD. We also implemented a counterpart to the role uncertainty feature in the main treatments which meant that outcomes would only count in half of the cases. In the control treatment, we let the computer pick a 'relevant scenario' instead of a 'relevant player'. If a subject's scenario was determined not to count, a compensation of £3 was awarded which lies just in-between the two possible payoff allocations (£6 or £0) which a subject could have expected to be allocated in the main treatments by her counterpart.

Instructions

Welcome to this experiment and thank you for participating. Please follow along carefully as the experimenter reads the instructions out aloud. The purpose of this experiment is to study how people make decisions in particular situations. You were awarded £3 for showing up on time. Your additional earnings in this experiment depend on the decisions you and other participants make during the experiment and on chance. At the end of the experiment, the entire amount will be paid to you *individually* and *privately* in cash by an assistant.

Please do not speak to other participants during the experiment and keep your phones switched off. If you have any questions at any time over the course of the experiment, please raise your hand and an experimenter will come to assist you.

Note that your behaviour in this experiment is recorded by the computer and stored in a database. The records of this database are anonymous, i.e. not traceable to you as a person. For accounting reasons only, you will be asked to fill in and sign a receipt of your earnings at the end of the experiment. To secure anonymity, these receipts will be kept entirely separate from any data on your behaviour generated in the experiment.

Please remain seated until you are individually asked by the experimenter to collect your final earnings at the end of the experiment.

The Experiment

All participants in this experiment are provided with the same set of instructions and will encounter the same stages as described below:

Stage 1: Matrix Task.

In stage 1 of the experiment, you will work on a matrix solving task. The task consists of counting *ones* (1s) in a series of matrices comprised of random 0s and 1s. A sample matrix is depicted in Figure 1 below.

Figure 1: Sample Matrix

0	1	1	1	0
0	0	0	0	0
1	0	0	1	1
1	1	1	1	0
1	1	0	1	0

You will be able to work on this task for a *maximum* of 300 seconds (5 minutes). Importantly, you will be timed-out by the computer at some point during this time interval. If this happens, the matrix task will end. You will then be asked to work on a follow-up task for the remainder of the 300 seconds.

All participants will be provided with additional details regarding the time-out mechanism in the later course of the experiment.

Outcomes in the matrix task (not however in the follow-up task) have direct consequences for the decision environment in stage 2 of the experiment:

- If you correctly solve at least 15 matrices before you are timed-out by the computer, you will be able make a decision in stage 2 of the experiment.
- If you do not correctly solve at least 15 matrices before you are timed-out by the computer, you will *not* be able to make a decision in stage 2 of the experiment.

After the conclusion of the matrix and follow-up task (i.e. after 300 seconds), you will move forward to stage 2 of the experiment.

Stage 2: Decision Stage.

In stage 2 of the experiment, you will *potentially* be able to choose between two options. Your choice indicates how much money you would like to allocate to yourself. The possible options are:

- Option A: **£10** to you.
- Option B: **£6** to you.

If you succeeded in solving at least 15 matrices in stage 1 of the experiment, *you yourself* will choose between Option A and Option B.

If you did not succeed in solving at least 15 matrices in stage 1 of the experiment, *the computer instead* will randomly choose between Option A and Option B with equal probability.

The resulting option (A or B) will be called your individual stage 2 outcome.

Determining the Relevant Scenario.

After you have completed the stages above, the computer will randomly determine whether your stage 2 outcome becomes the **Relevant Scenario**.

If your stage 2 outcome becomes the Relevant Scenario, it will be implemented. If your stage 2 outcome *does not* become the Relevant Scenario, your stage 2 outcome *will not* be implemented and will therefore have *no consequences* for payoffs in the experiment. In this case, you will instead earn a compensation of £3.

Note that it is *equally likely* that your stage 2 outcome *will* or *will not* become the Relevant Scenario.

Practice.

We will now briefly guide you through the decision stages in order for you to get a better understanding of the interface and processes of this experiment. You will also be able to familiarise yourself with the matrix task. We will conclude the practice phase with a quiz to check your understanding.

Please follow along on screen.

B.3 Revelation of Cut-off Details

Just before subjects entered the matrix solving stage, we publicly announced treatment specific details regarding the cut-off mechanism both verbally and on screen.

Script [1] for treatments NC_ND and C_ND:

Details regarding the time-out mechanism:

In this experiment, the time-out in the matrix task will occur at a fixed point in time. You will be timed out when the maximum allotted time of 300 seconds (5 minutes) is reached. Note that you and your counterpart will be timed out at the exact same time.

Script [2] for treatments NC_PD and C_PD:

Details regarding the time-out mechanism:

In this experiment, the time-out in the matrix task will occur at a randomly determined point in time. Any second within the maximum allotted time of 300 seconds (5 minutes) is possible. Note that you will be timed out independently of your counterpart.

Script [3] for treatment CONTROL:

Details regarding the time-out mechanism:

In this experiment, the time-out in the matrix task will occur at a randomly determined point in time. Any second within the maximum allotted time of 300 seconds (5 minutes) is possible.

B.4 Control Questions

Upon completion of the practice stage, subjects were asked to answer a set of 5 control questions. 2 more control questions were assessed after details regarding the cut-off mechanism were announced. Questions in the control treatment differed only marginally which is why they are not explicitly reported here. Correct answers are highlighted. Where correct answers differ, green marks the correct answer in the No Deniability conditions and blue marks the correct answer in the Plausible Deniability conditions.

Control Question 1:

The data generated in this experiment ...

✓ is anonymous, neither the experimenter nor other participants will be able to link my behaviour to me as a person.

links my behaviour in the experiment to me as a person.

links my behaviour in the experiment to me as a person, but only the experimenter will be able to make this connection.

Control Question 2:

Participants in this experiment ...

are provided with different instructions but will encounter the same stages in the experiment.

✓ are all provided with the same instructions and will encounter the same stages in the experiment.

will encounter different stages in the experiment.

Control Question 3:

I will be able to choose between Option A and Option B ...

no matter what.

✓ only if I solve enough matrices on time.

only if I will be timed-out by the computer in the matrix task.

Control Question 4:

My stage 2 outcome will contribute to my earnings in the experiment ...

no matter what.

✓ only if I become assigned the role of Relevant Player at the end of the experiment.

only if I become assigned the role of Relevant Player at the beginning of the experiment.

Control Question 5:

My counterpart ...

will learn whether I succeeded in the matrix task.

will learn whether my stage 2 outcome was chosen by me or by the computer.

✓ will neither learn my performance in the matrix task nor whether my stage 2 outcome was chosen by me or by the computer.

Control Question 6:

In this experiment ...

✓ I will be timed out when the maximum allotted time of 300 seconds is reached.

✓ I can be timed out at any second within the maximum allotted time of 300 seconds.

I will never be timed out.

Control Question 7:

What will your counterpart know after you completed the matrix task?

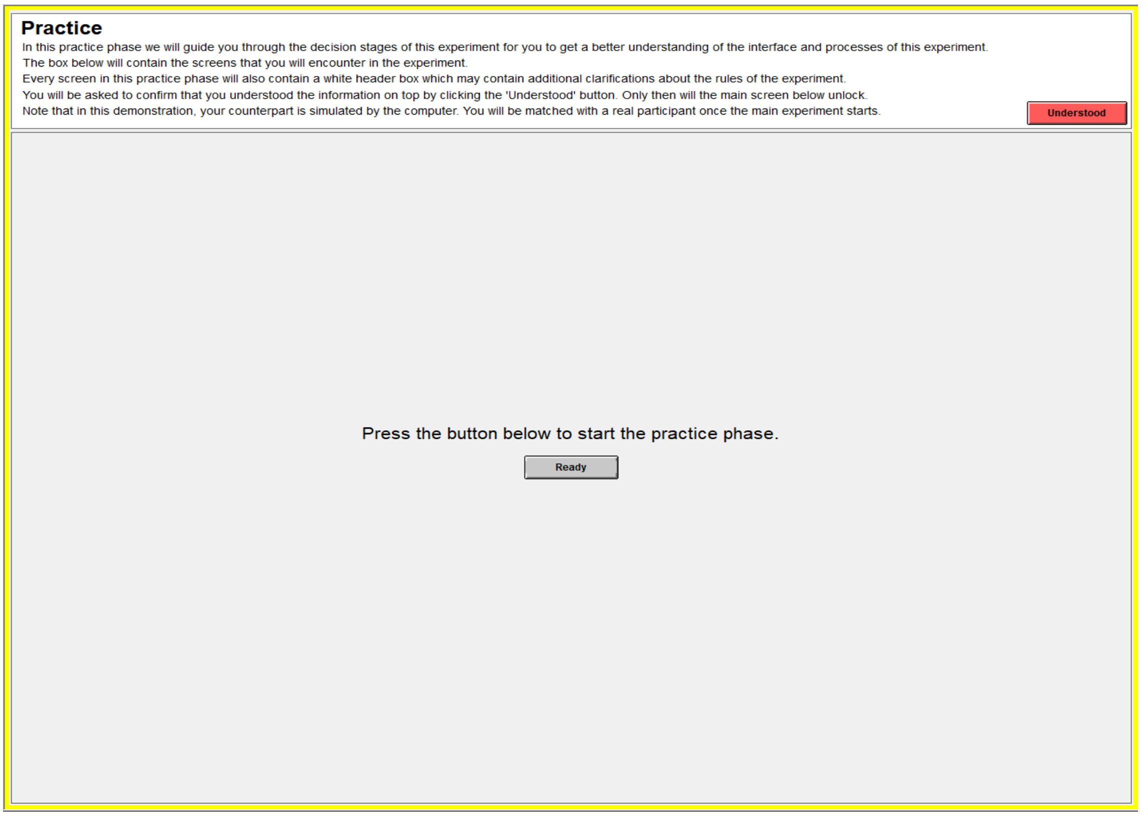
My counterpart will know how many matrices I solved.

✓ My counterpart will know that I was able to work on the matrix task for 300 seconds.

✓ My counterpart will know that I was timed out anywhere within 300 seconds. He will however not know when exactly my time out occurred.

C Sample Screenshots

Figure 6: Practice Stage Screen

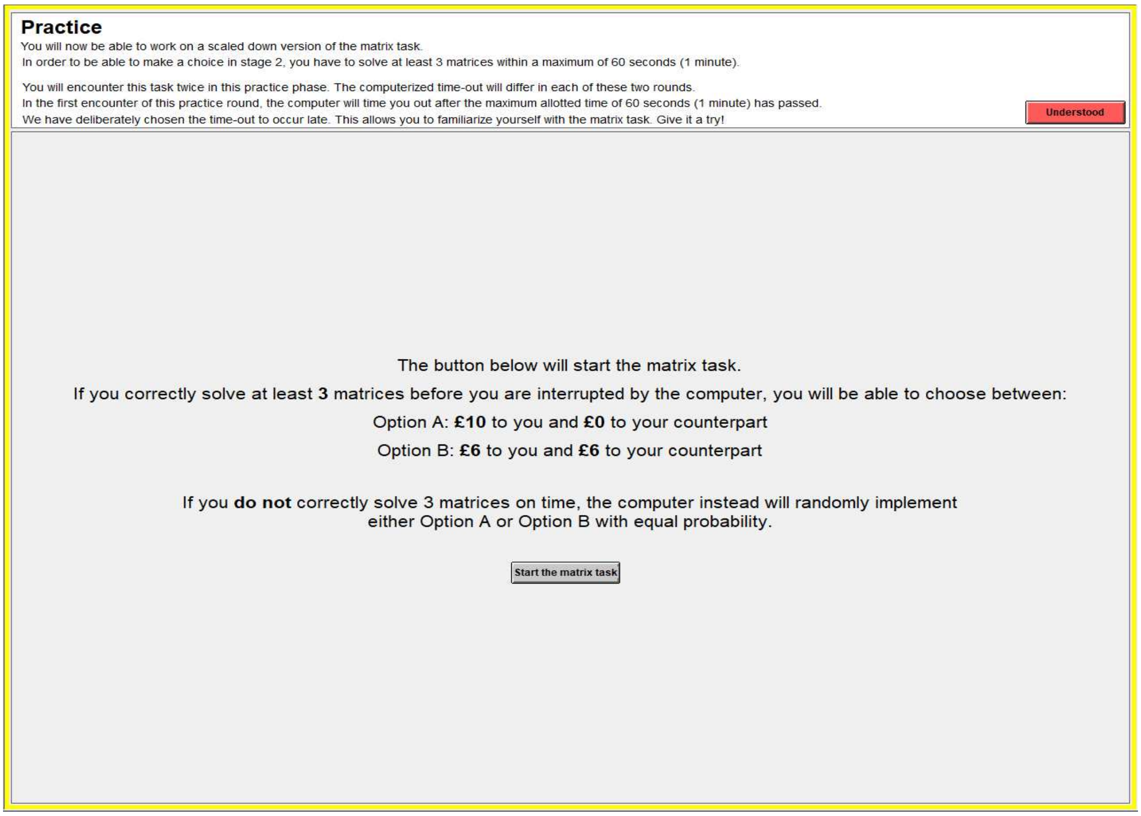


Practice

In this practice phase we will guide you through the decision stages of this experiment for you to get a better understanding of the interface and processes of this experiment.
The box below will contain the screens that you will encounter in the experiment.
Every screen in this practice phase will also contain a white header box which may contain additional clarifications about the rules of the experiment.
You will be asked to confirm that you understood the information on top by clicking the 'Understood' button. Only then will the main screen below unlock.
Note that in this demonstration, your counterpart is simulated by the computer. You will be matched with a real participant once the main experiment starts.

Press the button below to start the practice phase.

Figure 7: Practice Stage Screen (cont.)



Practice

You will now be able to work on a scaled down version of the matrix task.
In order to be able to make a choice in stage 2, you have to solve at least 3 matrices within a maximum of 60 seconds (1 minute).

You will encounter this task twice in this practice phase. The computerized time-out will differ in each of these two rounds.
In the first encounter of this practice round, the computer will time you out after the maximum allotted time of 60 seconds (1 minute) has passed.
We have deliberately chosen the time-out to occur late. This allows you to familiarize yourself with the matrix task. Give it a try!

The button below will start the matrix task.

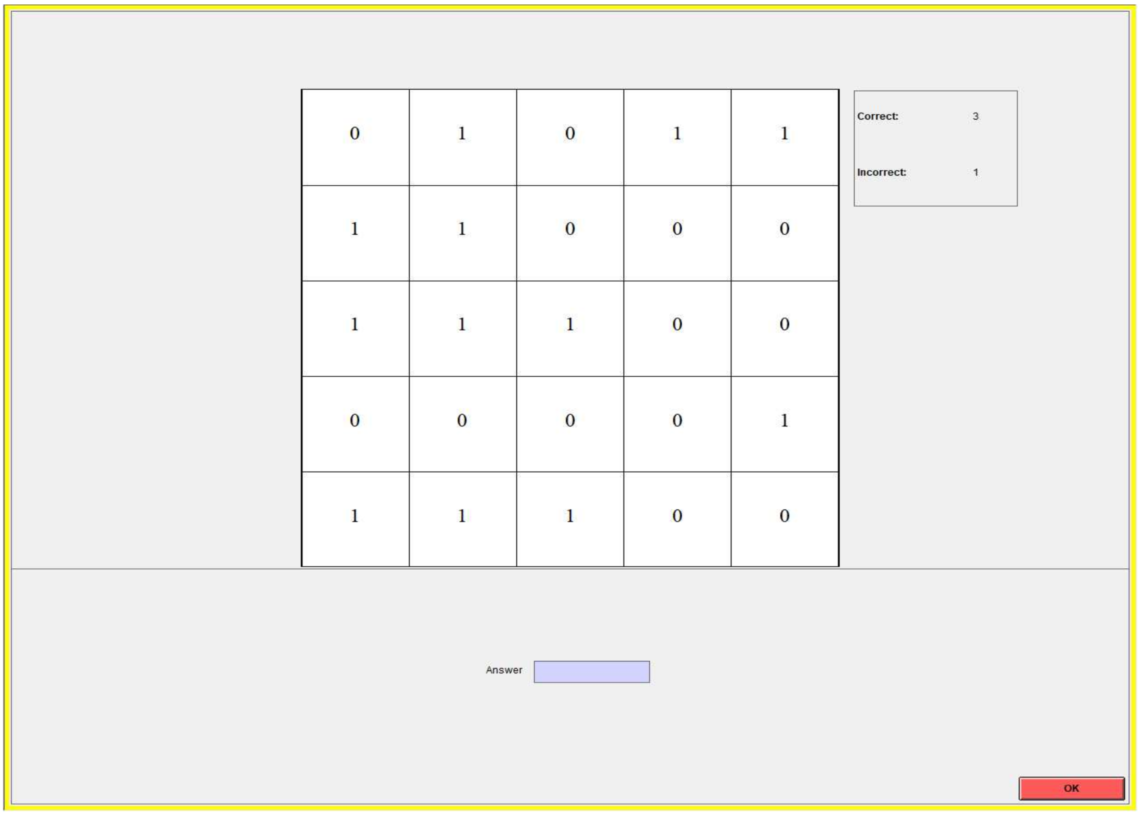
If you correctly solve at least **3** matrices before you are interrupted by the computer, you will be able to choose between:

Option A: **£10** to you and **£0** to your counterpart

Option B: **£6** to you and **£6** to your counterpart

If you **do not** correctly solve 3 matrices on time, the computer instead will randomly implement either Option A or Option B with equal probability.

Figure 8: Matrix Task Screen



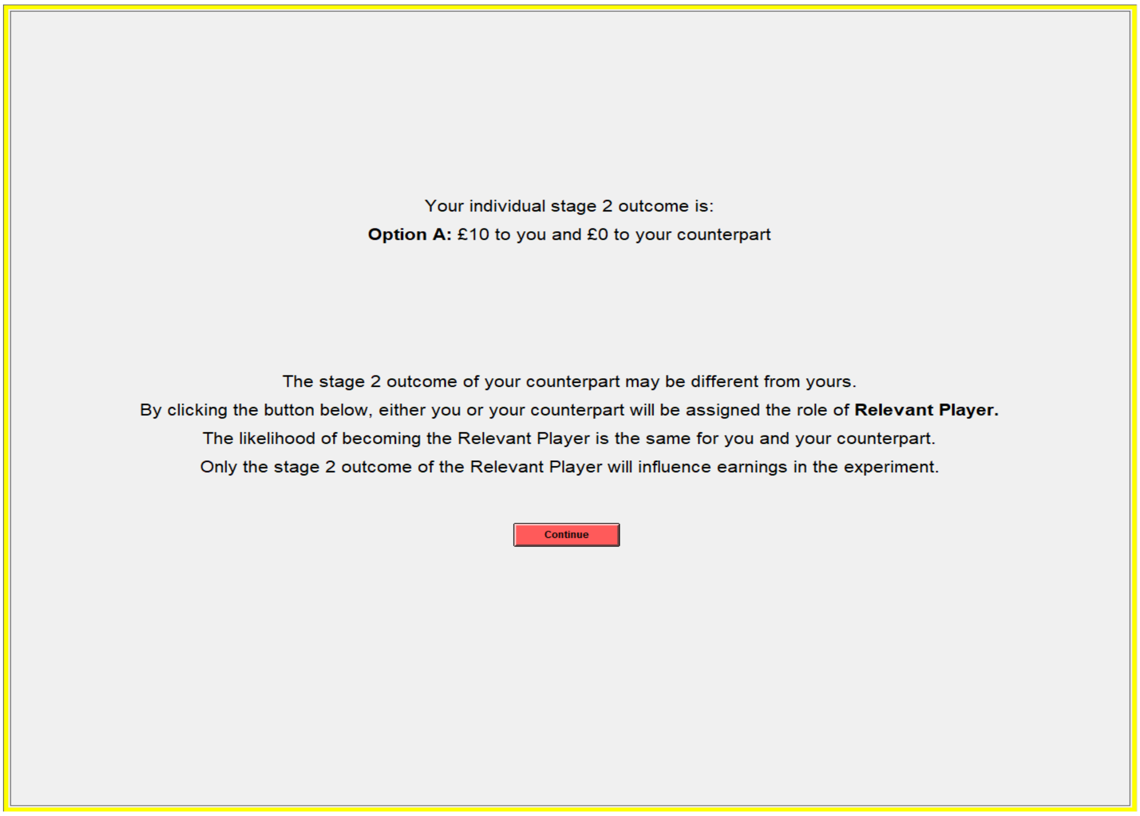
0	1	0	1	1
1	1	0	0	0
1	1	1	0	0
0	0	0	0	1
1	1	1	0	0

Correct: 3

Incorrect: 1

Answer

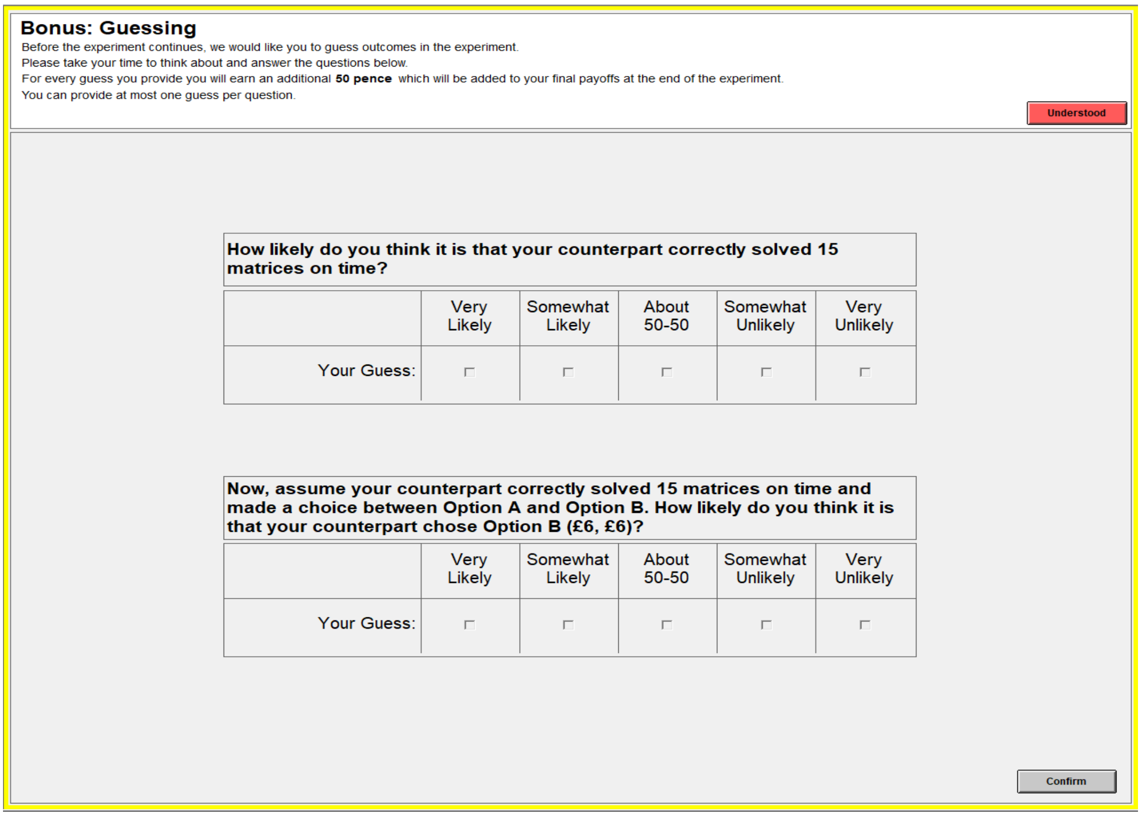
Figure 9: Role Assignment Screen



Your individual stage 2 outcome is:
Option A: £10 to you and £0 to your counterpart

The stage 2 outcome of your counterpart may be different from yours.
By clicking the button below, either you or your counterpart will be assigned the role of **Relevant Player.**
The likelihood of becoming the Relevant Player is the same for you and your counterpart.
Only the stage 2 outcome of the Relevant Player will influence earnings in the experiment.

Figure 10: Belief Elicitation Screen



Bonus: Guessing

Before the experiment continues, we would like you to guess outcomes in the experiment.
Please take your time to think about and answer the questions below.
For every guess you provide you will earn an additional **50 pence** which will be added to your final payoffs at the end of the experiment.
You can provide at most one guess per question.

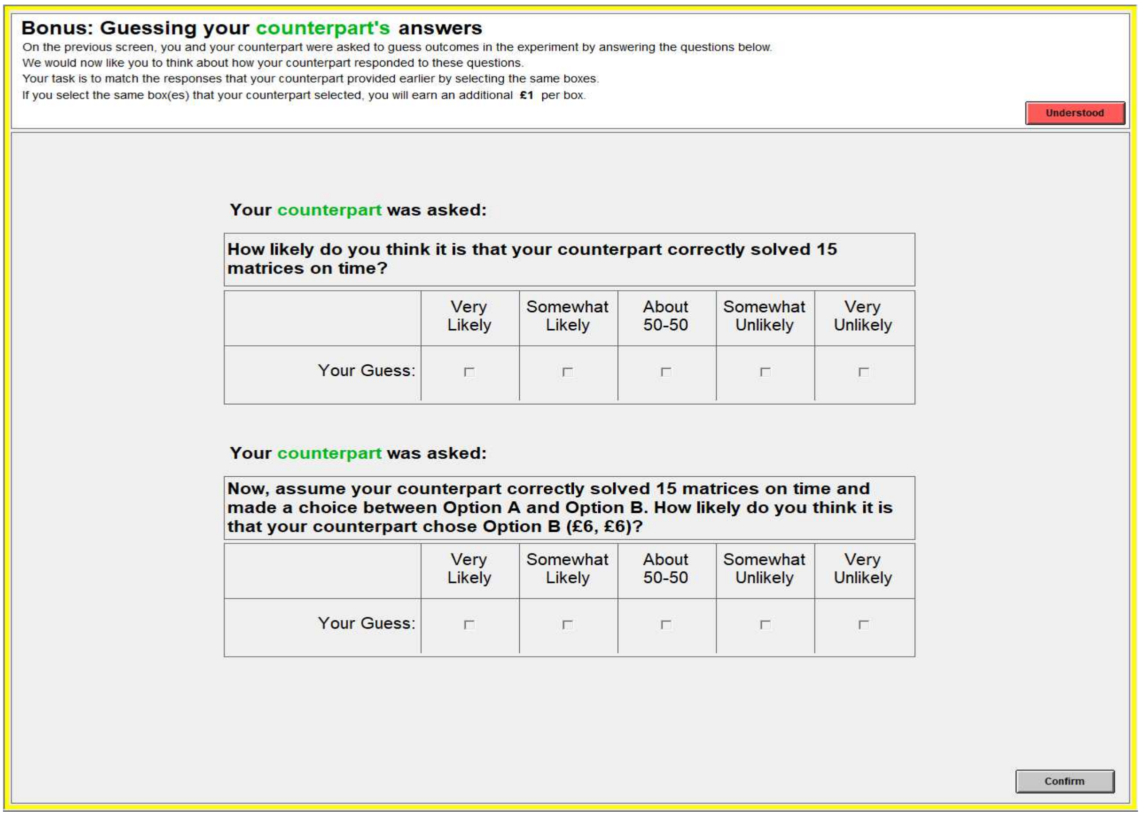
How likely do you think it is that your counterpart correctly solved 15 matrices on time?

	Very Likely	Somewhat Likely	About 50-50	Somewhat Unlikely	Very Unlikely
Your Guess:	□	□	□	□	□

Now, assume your counterpart correctly solved 15 matrices on time and made a choice between Option A and Option B. How likely do you think it is that your counterpart chose Option B (£6, £6)?

	Very Likely	Somewhat Likely	About 50-50	Somewhat Unlikely	Very Unlikely
Your Guess:	□	□	□	□	□

Figure 11: Belief Elicitation Screen (cont.)



Bonus: Guessing your counterpart's answers

On the previous screen, you and your counterpart were asked to guess outcomes in the experiment by answering the questions below.
We would now like you to think about how your counterpart responded to these questions.
Your task is to match the responses that your counterpart provided earlier by selecting the same boxes.
If you select the same box(es) that your counterpart selected, you will earn an additional **£1** per box.

Your counterpart was asked:

How likely do you think it is that your counterpart correctly solved 15 matrices on time?

	Very Likely	Somewhat Likely	About 50-50	Somewhat Unlikely	Very Unlikely
Your Guess:	□	□	□	□	□

Your counterpart was asked:

Now, assume your counterpart correctly solved 15 matrices on time and made a choice between Option A and Option B. How likely do you think it is that your counterpart chose Option B (£6, £6)?

	Very Likely	Somewhat Likely	About 50-50	Somewhat Unlikely	Very Unlikely
Your Guess:	□	□	□	□	□