Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# A Meeting of the Minds: Informal Agreements and Social Norms

Erin L. Krupka, Stephen Leider, Ming Jiang

To cite this article:
Erin L. Krupka, Stephen Leider, Ming Jiang (2017) A Meeting of the Minds: Informal Agreements and Social Norms. Management Science 63(6):1708-1729. https://doi.org/10.1287/mnsc.2016.2429

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 6, June 2017, pp. 1708–1729
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

# A Meeting of the Minds: Informal Agreements and Social Norms

**Erin L. Krupka,[a,b] Stephen Leider,[c] Ming Jiang[d]**

[a] School of Information, University of Michigan, Ann Arbor, Michigan 48109; [b] Institute for the Study of Labor (IZA), D-53113 Bonn, Germany; [c] Ross School of Business, University of Michigan, Ann Arbor, Michigan 48109; [d] Antai College of Economics and Management, Shanghai Jiao Tong University, 200240 Shanghai, China

**Contact:** ekrupka@umich.edu (ELK); leider@umich.edu (SL); mjiang@sjtu.edu.cn (MJ)

**Received:** January 20, 2014
**Revised:** March 17, 2015; September 26, 2015
**Accepted:** November 2, 2015
**Published Online in Articles in Advance:** May 31, 2016

**https://doi.org/10.1287/mnsc.2016.2429**



**Abstract.** Using coordination games, we elicit social norms directly for two different games where either an agreement to take the first best action has been reached or where no such agreement exists. We combine the norms data with separately measured choice data to predict changes in behavior. We demonstrate that including social norms as a utility component significantly improves predictive performance. Then we compare social norms to guilt aversion and lying aversion. We estimate that honoring an agreement in the double dictator game is worth giving up approximately 10% of total earnings and more than 120% in the Bertrand game. We show that informal agreements affect behavior *through* their direct effect on social norms as well as through an indirect effect on beliefs.

**History:** Accepted by Uri Gneezy, behavioral economics.
**Funding:** Funding for this research was provided by the University of Michigan School of Information and the Ross School of Business as part of E. L. Krupka and S. Leider's research support.
**Supplemental Material:** The online appendices are available at https://doi.org/10.1287/mnsc.2016.2429.

**Keywords:** economics • behavior and behavioral decision making • social norms • informal agreements

## 1. Introduction

Many transactions are supported by verbal promises or other informal agreements rather than formal contracts.[1] For example, the motto of the London Stock Exchange is "My word is my bond." Several papers demonstrate that such promises have a substantial impact on an individual's behavior,[2] even when fulfilling the promise entails a personal cost (Ellingsen and Johannesson 2004, Charness and Dufwenberg 2006, Vanberg 2008, Kessler and Leider 2012, Dufwenberg et al. 2011).[3] These papers in economics build on an earlier literature in social psychology on the role of communication and promises in social dilemmas (Loomis 1959, Dawes et al. 1988, Orbell et al. 1991). In particular, making promises leads to substantially better outcomes than standard theory would predict.

The most prominent explanations for why informal agreements have the power to affect behavior are a desire to conform with social norms (Kessler and Leider 2012), guilt aversion (Charness and Dufwenberg 2006), and lying aversion (Ellingsen and Johannesson 2004, Gneezy 2005). Each of these explanations has been tested separately but never been tested head-to-head within a single experimental framework. Furthermore, the empirical work either offers only indirect evidence (see Kessler and Leider 2012) or provides only a partial explanation for behavior (see Charness and Dufwenberg 2006, Ellingsen and Johannesson 2004). In this paper, we combine choice data with collected data on norms. Our goal is to understand how informal agreements work and to demonstrate that making an informal agreement changes the social norm governing a decision. Furthermore, we compare social norms to guilt aversion and lying aversion and identify which mechanism (or combination of mechanisms) can best explain behavior in both a setting where informal agreements are present and where they are absent.

A social norm is inherently a social construction in which there is joint recognition that a particular behavioral rule exists, that the rule characterizes what one ought to do, and that it is applicable to the relevant situation (Bicchieri 2006, Krupka and Weber 2013). Individuals experience utility from complying with actions that are collectively judged to be appropriate and experience disutility when they take actions that are collectively deemed inappropriate. In the context of informal agreements, the social norm reflects a collectively shared belief—or a meeting of the minds—that informal agreements ought to be honored (López-Pérez 2008, Kessler and Leider 2012).[4] Under the *social norms mechanism*, changes in behavior stem from changes in an action's appropriateness when an informal agreement is present.

On the other hand, guilt aversion posits that actors experience guilt if they disappoint others (Charness and Dufwenberg 2006, Battigalli and Dufwenberg 2007). The guilt aversion model states that an individual has a (second-order) belief about the outcomes that the other party expects (first-order belief) and

experiences disutility when generating outcomes for the other party that are worse than expected.[5] Thus, under the *guilt aversion mechanism*, changes in behavior stem from the power of an informal agreement to affect the interacting parties' beliefs in a specific way: namely, the actor believes that the other party expects him to comply with that agreement.

However, while guilt aversion can show that *given* a change in beliefs those beliefs will be fulfilled in equilibrium, it has little to say (on its own) about why interacting parties have expectations that informal agreements will be honored and "does not suggest which forms of communication move beliefs" (Charness and Dufwenberg 2006, p. 1595). To explain why parties' expectations are affected by informal agreements, Charness and Dufwenberg (2006) argue that norms shape expectations, and deviation from those expectations generates guilt.[6]

Echoing this intuition, more recent work relies on social norms to determine ex ante expectations about what actions ought to be taken or what actions will likely be taken by transacting parties (Sliwka 2007, Hart and Moore 2008, Fehr et al. 2009). In essence, this body of work suggests that norms may have an indirect effect on behavior through beliefs and subsequent expectations (which can give rise to guilt). However, it is also possible that to best explain behavior one may need to account for the direct and independent impact of norms on behavior, not merely the indirect effect through beliefs—a pathway we explore here.

Lying aversion posits that deviating from what the actor said he was going to do generates disutility. This aversion is a personal preference and does not work through beliefs. An aversion to lying may stem from a social norm prohibiting such behavior,[7] or conversely, an innate aversion to lying may be why the social norm exists. As such, lying aversion and a desire to comply with social norms may be difficult to distinguish from one another in the *presence* of informal agreements. However, lying aversion has little to say about behavior when *no* agreement is reached; in such cases, lying aversion is not an appropriate model, whereas a social norms model may still be able to explain behavior both with and without an agreement. Thus, lying aversion may be too limited to describe the full range of behavior across games and agreement conditions even if it is the direct result of a social norm or the reason the norm exists in the first place.

Thus, a key feature that distinguishes a social norms mechanism from guilt aversion is that a guilt aversion model will predict behavior using the individual beliefs held by the interacting parties about each other. In this model, each individual in a pair can hold different beliefs and each pair can hold different beliefs from another pair. The social norms mechanism, on the other hand, will rely on common beliefs (held by all individuals regardless of whom they interact with) to predict behavior. A key feature that will distinguish the social norms mechanism from lying aversion is that norms can also affect behavior when no agreement is present.[8]

To identify the relationship between promises and social norms, and to disentangle norms from the alternative mechanisms, we need data on choice behavior as well as the relevant social norms and beliefs. We collect behavioral data in our *choice experiments* using two games: a double dictator game (DDG) and a Bertrand game (BG; see Kessler and Leider 2012, Dufwenberg and Gneezy 2000, Dufwenberg et al. 2007). In the former, partnered subjects make a simultaneous transfer decision that results in a division of their endowments.[9] In the later, subjects simultaneously select any whole number between 0 and 100, and whoever chooses a smaller number has a payoff equal to his number while the other player gets a payoff of 0. If both players choose the same number, than profits are divided equally. Both games have a rich enough action space to distinguish between the three mechanisms.[10] To implement the games in our *choice experiment*, we replicate the Kessler and Leider (2012) experimental design but add an elicitation of second-order beliefs to the protocol so that we can directly test the guilt aversion model. Using two games for our experiment is attractive because it allows us to test how well the three different mechanisms explain behavior across settings—one game has strategic independence while the other has strategic complements—as well as across agreement and no agreement conditions.[11] Finally, studying a context in which both parties come to a mutual agreement approximates a key aspect of informal agreements that we wish to have as our focus. This focus contrasts with decision contexts used to test lying aversion and guilt aversion that have typically been restricted to unilateral promises. Consistent with the previous work on promise making, we find that having an informal agreement leads to substantially higher actions than when there is no agreement in place.

To collect data on the social norms, we conduct a separate *norm elicitation experiment* using the Krupka and Weber (2013) protocol to elicit the social norms for each of the *choice experiment's* games and agreement or no agreement conditions. Just using the social norms data, we demonstrate that making a promise to take a highly prosocial action significantly and substantially changes the social norm: fulfilling the promise *increases* in appropriateness while taking even very prosocial actions that fall short of the promise become socially inappropriate. We then take the new norms data and merge it with the separately collected choice data to estimate a choice model describing the behavior of subjects in the two games and agreement conditions. We find that including social norms improves

our explanatory power across games and agreement conditions and also captures qualitative moments of the data.

We then estimate choice models for guilt and lying aversion and compare the explanatory power across the three models. The social norms model performs better than lying aversion, whereas the guilt aversion model does a better job of explaining behavior in one game but does not do well across both games. Furthermore, both lying and guilt aversion miss key qualitative features of behavior across games and conditions that the social norms model captures. Finally, we show that adding social norms to either guilt or lying aversion improves the predictive power of either model. We conclude that the evidence is consistent with a direct effect of social norms on behavior and an indirect effect via changing beliefs.

Our main contribution is to provide an analysis of informal agreements that directly demonstrates the role of social norms and considers together all three proposed mechanisms in the literature. Similarly, although we use largely the same methods as Krupka and Weber (2013), the goals and setting of this paper are different. Krupka and Weber focus on demonstrating that a social norms framework is a viable explanation for (unilateral) dictator behavior; in this paper we use the norms data to distinguish three separate mechanisms in strategic games.[12] Our results provide evidence on a relatively simple mechanism, norms, by which informal agreements operate to produce the observed behavior changes across two different games.

## 2. Promise Mechanisms

Three major mechanisms have been proposed to explain why nonbinding verbal promises may have a substantial effect on behavior. One approach focuses on social norms as a generally important influence on behavior and notes that there is a widely recognized and quite strong social norm against violating one's word. Making a promise to take a specific action therefore increases the psychological cost of choosing another action. Another approach uses the framework of psychological game theory and guilt aversion and argues that individuals dislike disappointing others. Here, a promise serves to change the beliefs of the other party, which then increases the costs of choosing actions different from that belief. Finally, one can directly assume a psychological cost for lying. Therefore, an action may become psychologically costly if it makes a previous statement into a lie. Social norms and guilt aversion differ fundamentally in that social norms are collectively defined beliefs about expectations for behavior, whereas guilt aversion depends on individually held beliefs about expectations. Lying aversion depends only on a personal cost for lying; however, it can also easily be seen as a special case of the general social norms framework.

### 2.1. Defining and Identifying Social Norms

We define (injunctive) social norms as *jointly recognized beliefs, among members of a population, regarding the appropriateness of different behaviors*. Following Elster (1989), we note two important features of social norms. First, social norms generally prescribe or proscribe behaviors or actions rather than outcomes. Allowing norms to govern actions, rather than outcomes, suggests that two actions that produce the same outcome but differ in other respects may be governed by different social norms (see Krupka and Weber 2013). Second, the "social" element of norms requires that they be jointly recognized, or collectively perceived, by members of a population.[13] These two features—that social norms typically apply to actions rather than outcomes and that they must be jointly recognized—are present in most researchers' definitions (Bettenhausen and Murnighan 1991, Fehr and Gächter 2000, Ostrom 2000, Bicchieri 2006).[14]

Furthermore, we distinguish norms regarding what one "ought" to do, or injunctive norms, from customs or actions that people regularly take, or descriptive norms (Deutsch and Gerard 1955, Bicchieri 2006). Both kinds of norms influence behavior (Cialdini et al. 1990, Krupka and Weber 2009, Bicchieri and Xiao 2009). However, our focus here is on injunctive social norms, i.e., those described by Elster (1989) as prescribing what one "should do" or "should not do."[15] From here on, when we talk about injunctive social norms, we will refer to them as *norms*. When we wish to distinguish injunctive norms from actions taken by most others, we will refer to the latter as *descriptive norms*.

To measure the extent to which actions are jointly recognized to be socially appropriate or inappropriate, we follow Krupka and Weber (2013) and present respondents with a description of a choice environment, including all the possible available actions. We ask respondents to judge the social appropriateness of *each* action on a six-point scale that ranges over "very socially inappropriate," "socially inappropriate," "somewhat socially inappropriate," "somewhat socially appropriate," "socially appropriate," and "very socially appropriate."[16] We provide respondents with incentives to *match* their ratings to the responses of other subjects in the session, rather than to provide their personal opinions. Thus, respondents play a coordination game in which the incentive is to anticipate the extent to which others will rate an action as socially appropriate or inappropriate and to respond accordingly.[17] From a game-theoretic point of view, matching games have a number of equilibria, and nothing intrinsic to the game makes one equilibrium favored (or focal) over the other, although common culture and shared experiences can create focal points (Schelling 1960, Mehta et al. 1994, Sugden 1995).

In our experiment, we assume that collectively recognized social norms create focal points in the matching game. (In Online Appendix I, in which we provide additional information regarding the norms experiment, Section 3 describes several tests of this assumption.)[18] That is, if there is a social norm that some actions are more or less socially appropriate, respondents attempting to match others' appropriateness ratings are likely to rely on this shared perception to help them do so. Thus, the incentive in the coordination game elicits collective perceptions of appropriateness, which we will call our empirical measure of the social norm.

More formally, we let $A = \{a_1, \ldots, a_K\}$ represent a set of $K$ actions available to a decision maker. The social norm function $N(a_k)$ is an empirically observed collective judgment that assigns to each action a degree of appropriateness or inappropriateness that reflects the norm of the relevant group. Thus, if, for an action $a_k$ there is collective recognition among group members that the action constitutes "norm-consistent" behavior, then $N(a_k) > 0$.[19] If there is joint recognition that an action constitutes "norm-inconsistent" behavior, then $N(a_k) < 0$. This formalization makes apparent that the social norm applies to the *entire set of possible actions*; as such, the elicited social norm function can be interpreted as a characterization of the *profile* of appropriateness ratings over all the actions available to a decision maker that stems from the social norm the researcher is trying to measure.[20,21]

We can now embed this definition of social norms into a simple utility framework that will motivate our subsequent estimation of the concern that individuals have for norm compliance relative to other payoff relevant preferences. We motivate our empirical work by assuming that the individual cares about both the monetary payoff $x_i(a_k, a^{-i})$ produced by the selected action, $a_k$ (given the actions of other individuals, denoted by $a^{-i}$), and the degree to which the action is collectively perceived as socially appropriate such that

$$u_i(x, a_k) = V_i(x_i(a_{i,k}, a^{-i})) + \gamma_i N(a_k). \tag{1}$$

For an individual $i$, the function $V(\,)$ represents the value the individual places on the monetary payoffs from a particular action, $a_k$, and is concave and increasing in $x_i(a_k, a^{-i})$. One important feature of this model is that *actions* are arguments in the utility function; in this sense, the social norms model is different from standard social preference models (see Fehr and Schmidt 1999). The moral weight of an action therefore depends only on the action itself, not on the actions that others take (nor on the outcomes that follow). The parameter $\gamma_i \geq 0$ represents the degree to which the individual cares about adhering to a particular norm.[22] An individual entirely unconcerned with social norms ($\gamma_i = 0$) will always select the payoff-maximizing action. On the other hand, as $\gamma_i$ increases, an individual will derive greater utility from selecting actions that are socially appropriate relative to the utility from those that are not. Note that it is generally not possible to separately identify both $\gamma_i$ and $N(\cdot)$ from behavioral data alone. This is why it is necessary to have some independent means of identifying either $\gamma_i$ or $N(\cdot)$. Our approach is to use a separate group of subjects to empirically identify $N(\cdot)$ using the norm elicitation experiment. We then combine the empirical measure of $N(\cdot)$ with the behavioral data from our choice experiment and are able to estimate $\gamma$.

With this framework we can see how the social norms mechanism might result in different behavior across choice environments even when they are payoff equivalent. The framework also provides a testable relationship between the degree of social appropriateness of actions and individuals' willingness to take those actions, provided one has a reasonable method for independently capturing the "social appropriateness" of the different available actions. We now provide hypotheses about what features the social norm might reasonably have in a the games we study.

In previous research, Krupka and Weber (2013) found that subjects judge prosocial behavior as generally socially appropriate, while more selfish behavior is generally considered less socially appropriate (though the relationship is not clearly monotonic). In our choice experiments, "higher" actions are prosocial in the sense that they (weakly) increase the total surplus. In the context of our norm elicitation experiment, this leads to the following straightforward hypothesis regarding the appropriateness ratings.

**Hypothesis 1.** *Actions that are more prosocial will be seen as socially appropriate, and actions that are more selfish will be considered less socially appropriate.*

Numerous experiments have demonstrated that preplay communication of various forms can increase the prosociality of individual behavior (for an early experiment, see, for example, Dawes et al. 1977; for an early survey, see Sally 1995). Promises to take a particular action have been shown to be particularly powerful in changing behavior (Charness and Dufwenberg 2006, Vanberg 2008, Kessler and Leider 2012). These results can be interpreted to suggest that there is a norm of promise keeping that will be active in our agreement condition. This would suggest that the only socially appropriate actions are those that fulfill the promise. Furthermore, the appropriateness of an action may change when an informal agreement is made. For example, while sending 80% of the endowment in the double dictator game may be seen as relatively prosocial when there is no agreement, if subjects have an informal agreement to send the entire endowment,

then sending only 80% is a violation of that promise. Thus, sending 80% of the endowment in the former case may be judged "socially appropriate," whereas sending 80% of the endowment in the latter case may be judged a "socially inappropriate" action because it violates the informal agreement. This yields Hypotheses 2A and 2B—that making an agreement to take a particular action will significantly impact the social norm in the following way.

**Hypothesis 2A.** *The agreed-upon action will be substantially more appropriate than other actions.*

**Hypothesis 2B.** *Compared with the no agreement case, an agreement will increase the appropriateness of the agreed-upon action and (weakly) decrease the appropriateness of all other actions.*

### 2.2. Guilt Aversion and Promises

Battigalli and Dufwenberg (2007) develop a model of guilt aversion in which individuals care about what others expect of them and feel disutility (guilt) when their actions fall short of those expectations. Charness and Dufwenberg (2006) apply guilt aversion to explain behavior in a stochastic trust game where individuals can communicate, and therefore make promises, before choosing an action. They argue that making a promise is likely to increase the expectations of the other party and therefore likely to increase the promise maker's second-order beliefs. If beliefs change in response to making a promise, then promise making can be self-enforcing, as promise makers will have increased disutility for choosing lower actions (compared with a nonpromise maker). Let $x'_j$ denote player $i$'s beliefs about player $j$'s expectations of his ($j$'s) payoffs. Then $\max\{x_j(a_{i,k}, a^{-i}) - x'_j, 0\}$ indicates how much player $i$ believes that his action will disappoint player $j$ (we will refer to this "guilt aversion" term as GA). We can then describe an individual's intrinsic propensity to feel guilt by a parameter $g_i$ and represent player $i$'s utility as

$$u_i(x, a_k) = V_i(x_i(a_{i,k}, a^{-i})) - g_i \max\{x_i(a_{i,k}, a^{-i}) - x'_j, 0\}. \quad (2)$$

In addition to testing for the effect of guilt on choices, we can also directly test the assumption that promises change first- and second-order beliefs, since this is a necessary precondition for the guilt aversion mechanism. We expect to find similar results as Charness and Dufwenberg (2006) that both first- and second-order beliefs will on average be higher when subjects have made a promise.

**Hypothesis 3A.** *Average first-order beliefs will be higher in the agreement case than in the no agreement case.*

**Hypothesis 3B.** *Average second-order beliefs will be higher in the agreement case than in the no agreement case.*

### 2.3. Lying Aversion and Promises

Several papers have examined how communication can affect behavior by assuming that individuals directly experience disutility from lying (Ellingsen and Johannesson 2004, Chen et al. 2008, Özer et al. 2011). In the situation where individuals can make promises before playing a game, such promises create psychological incentives to take actions that fulfill the promise (so that the promise will not be a lie). To model lying aversion, we specialize Chen et al. (2008) by assuming that the disutility of lying increases linearly[23] in the difference between one's action and the promised action $a^*_k$ (we will refer to this "lying aversion" term as LA)[24] such that

$$U_i^{\mathrm{LA}}(x, k) = V_i(x_i(a_{i,k}, a^{-i})) - k|a_{i,k} - a^*_k|, \quad \text{where } i \neq j. \quad (3)$$

In the final section of the paper, we use data collected in our *choice experiment* and our *norm elicitation experiment* to test how well the elicited social norms, guilt aversion, and lying aversion explain the actual choices made by subjects in the choice experiment. Using data obtained from the norm elicitation experiment, we elicit measures of social appropriateness ($N(a_k)$) over possible action choices in the two games (Bertrand and double dictator games) and agreement conditions (when there exists an agreement to take the first best action and when no such agreement exists). Using behavior data and the first- and second-order beliefs collected in the choice experiment, we can estimate a model of guilt aversion and lying aversion.

## 3. The Experimental Design

We would like our experimental data to accomplish two goals: (1) directly identify the social norm in the double dictator and Bertrand games for when there is an agreement or when there is no agreement and (2) allow us to predict behavior in those games and agreement conditions. To that end, our experimental design consists of two separate experiments: a norm elicitation experiment and a choice experiment.

In the choice experiment, we use a within-subject design with 20 rounds of play. In the first 10 rounds, subjects make choices in the double dictator game, and in the second 10 rounds, they make choices in the Bertrand game. In the double dictator game, subjects are randomly placed into A–B pairs. Each subject in the A–B pair starts with 20 units worth of tokens and must simultaneously choose whether to send between 0 and 20 tokens to the other person. Payoffs are calculated in the following way: A's earnings are 20 – (2 × what A sends) + (6 × what B sends); B's earnings are 20 – (2 × what B sends) + (6 × what A sends). In the Bertrand game, subjects are randomly placed into A–B pairs. Each A–B pair must simultaneously select any whole

number between 0 and 100. Whoever chooses a smaller number has a payoff equal to his number, whereas the other player gets a payoff of zero. If A and B choose the same number, then their payoff is equal to 1/2 of that number.

For each round, subjects are first paired with a (different) subject in the room. They are told which game they are playing and are given an opportunity to say whether they would like to have (or not like to have) an unenforceable agreement with the other subject to take the first best action for that game.[25] The computer then randomly determines whether the round is an "Agreement" round or a "No Agreement" round by flipping a virtual coin. If it is an Agreement round, then the computer checks to see if both A and B indicated that they wanted an informal agreement. If both said they wanted an informal agreement, then the computer informs A and B that they "have an agreement." If one or neither of the pair wished for an informal agreement, then the computer informs them of this. If the computer determines that it is a No Agreement round, then subjects are informed that no agreements can be made in this round (note that this is not a failure to reach an agreement, but a lack of opportunity to have one in place).

In both the No Agreement and Agreement situations, subjects are then prompted to select an action to take and then they are asked for their (incentivized) first- and second-order beliefs. For correctly guessing first or second-order beliefs, they receive \$0.25. Finally, subjects are informed about their pair's choice; what their payoff would be if the round were selected for payment; and whether they received \$0, \$0.25, or \$0.50 bonus for their guesses. This concludes the round. Our design enables us to collect a subject's choices for each game (Bertrand and double dictator) and for each agreement condition. Payment is determined by randomly selecting one of the 20 rounds for payment from each game.

Our norm elicitation experiment uses coordination games to elicit subjects' beliefs about normative evaluations and, in aggregate, identifies the norm for that decision context. We elicit the norms for the double dictator game with and without agreement and for the Bertrand game with and without agreement in Module 1 of this experiment, though the entire experiment consisted of five modules.[26] Thus, in our norm elicitation experiment, our subjects read a vignette that describes the choices an "individual A" would be faced with in the double dictator game or the Bertrand game. Subjects who read about the double dictator game with no agreement read the following vignette[27]:

> Individual A and individual B are randomly paired with each other. A and B each start with tokens worth 20 units. A must choose an action. B will also be choosing an action at the same time. The action that A and B choose will determine their earnings. A and B are told that their payoffs will be calculated in the following way: A's earnings are $20-(2\times$what A sends$)+(6\times$what B sends). B's earnings are $20-(2\times$what B sends$)+(6\times$what A sends). Beyond these basic instructions, [*in the case of No Agreement*] A and B were not given the opportunity to make any kind of agreement about what action they were each going to take.

Subjects reading about the Bertrand game with no agreement read the following vignette:

> Individual A and individual B are randomly paired with each other. A must choose an action. B will also be choosing an action at the same time. A's action, and B's action, consists of selecting any whole number between 0 and 100. Whoever chooses a smaller number has a payoff equal to his number while the other player gets a payoff of zero. If A and B choose the same number, then their payoff will be equal to 1/2 of that number. [*In the case of No Agreement*,] A and B were not given the opportunity to make any kind of agreement about what action they were each going to take.

In the Agreement treatments, subjects were instead told that "A and B were given the opportunity to make an agreement about what action they were each going to take. They agreed to each take action 10 [100]."

After reading about the situation and completing a comprehension check for the norms rating task,[28] subjects were asked to evaluate the "social appropriateness" of a number of the actions available to A[29] and to rate how sure they were that each of their ratings matched with each of the ratings of another subject. Subjects only rated one game (either the double dictator game or the Bertrand game) for only one agreement environment (either Agreement or No Agreement).[30]

We told subjects that by "socially inappropriate" we meant "consistent with what most people expect individual A ought to do."[31] We also told them that we would pay them not to reveal their own personal opinions but instead to try to match the appropriateness ratings of others. To incent subjects to think about what others think is appropriate, we introduced a proper scoring rule (Lambert and Shoham 2009) to the Krupka and Weber (2013) norm elicitation protocol. This scoring rule elicits subjects' median belief[32] about the distribution of others' ratings by matching a subject with another subject and then paying them according to the payoff function

$$\pi_i = \$15 - \$4|x_i - x_{-i}|, \quad \text{for each subject } i, \qquad (4)$$

where $\pi_i$ is the payoff of subject $i$, and $x_i$ and $x_{-i}$ are the appropriateness ratings for subject $i$ and the matched other subject, respectively. We chose to elicit an estimate of the median because (unlike a quadratic scoring rule to elicit the mean) this yields fewer extreme ratings when the distribution of the other's ratings is

particularly skewed (as might be the case for actions that are, as an example, extremely self-regarding or other-regarding). Furthermore, while there may be no changes in the modal rating an action receives, the median rating can change between treatments. As an example, even if the modal rating for taking the most prosocial action is unchanged when there is an agreement or not, the degree to which appropriateness ratings vary for actions that deviate from the most prosocial action may vary when an agreement is in place. This, in turn, will change the median rating.

To test our hypotheses, we converted subjects' norm ratings into numerical scores. A rating of "very socially inappropriate" received a score of 1, "socially inappropriate" a score of 2, "somewhat socially inappropriate" a score of 3, "somewhat socially appropriate" a score of 4, "socially appropriate" a score of 5, and "very socially appropriate" a score of 6.[33] Subjects' earnings for the norm elicitation experiment were calculated using the coordination payoffs described above. Subjects also received a \$5 show-up fee. Subjects were paid privately at the end of the experiment.

## 4. Results

Students from the University of Michigan were recruited to take part in either the *norm elicitation* or *choice* experiment. In the norm elicitation experiment, there were a total of 358 participants recruited in 36 sessions. Sessions were conducted using an even number of participants, ranging from 6 to 22 per session, and the average length of each session was 1 hour and 15 minutes. The average payoff for each subject was \$29.72. In the choice experiment, there were a total of 62 subjects in four sessions. The average payment, including the \$5 show-up fee, was \$16.63, and the average length of a session was about an hour. Table S2 in Online Appendix I details participation rates and average payoffs by treatment and experiment.

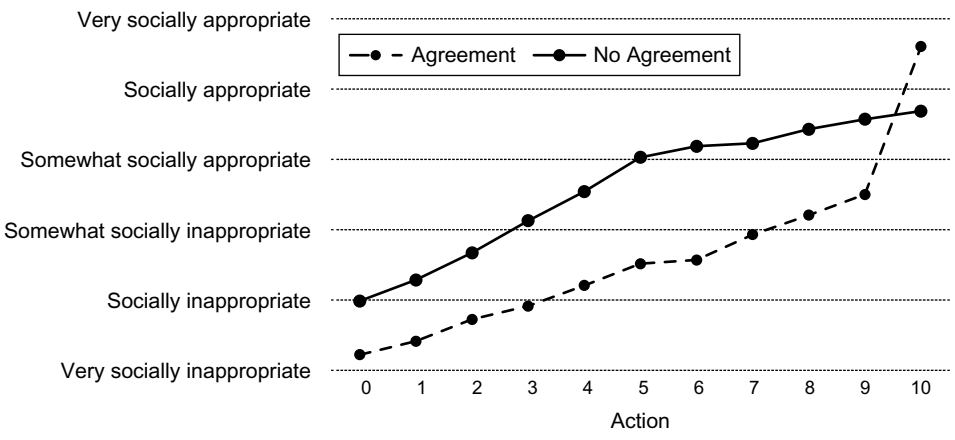
We begin our discussion of results by analyzing the data generated from our norm elicitation experiment and testing whether norms differ when there is an informal agreement. We then present the results from our choice experiment and test for the effect of informal agreements on behavior. We then combine the norms data with the choice data to predict behavior. We conclude our analysis of results by comparing the explanatory power of guilt and lying aversion to a social norms model.

### 4.1. Norm Elicitation: Norm Ratings With and Without Agreement Across Both Games

Recall that in module 1, subjects read a vignette about an individual A in either the Agreement or No Agreement condition for either the double dictator game or the Bertrand game and then provided social appropriateness ratings for all actions available to A (if rating the double dictator game) or a selection of actions (if rating the Bertrand game) in that situation. These responses yield our primary outcome measure—the "between-subjects" elicited ratings of social appropriateness, $N(a_k)$, for these two games and these two agreement conditions. In Section 3 of Online Appendix I, we conduct a number of robustness checks of our norm elicitation mechanism, which we do not discuss further here.

Figure 1 displays the average appropriateness ratings for the double dictator game with and without agreement. We see that subjects are using the full range of appropriateness ratings in both games, and we find that sending a small amount is seen as fairly socially inappropriate while sending a large amount is seen as more appropriate in the double dictator game (consistent with Hypothesis 1). Additionally, a rank-sum test shows that in the Agreement condition, sending 10 tokens is significantly more appropriate than sending 9 tokens in the double dictator game ($p < 0.01$, supporting Hypothesis 2A for the double dictator game).

**Figure 1.** Average Appropriateness Ratings for the Double Dictator Game With and Without Agreement (Data from Module 1)



There are also notable differences between the environments where an agreement exists and where none exists. First, every action other than sending the full amount is seen as less appropriate in the Agreement condition than in the No Agreement condition (consistent with Hypothesis 2B). A rank-sum test finds that appropriateness ratings are significantly higher in the No Agreement condition than in the Agreement condition for actions 0–9 ($p < 0.01$ for all). Second, sending the entire endowment of 10 tokens is seen as more appropriate in the Agreement condition than in the No Agreement condition (rank-sum test, $p < 0.01$). Additionally, the greatest increase in appropriateness in the No Agreement condition is for relatively low actions, and then ratings change little and remain fairly flat for higher transfer decisions; in particular, the average rating for sending all 10 tokens is not significantly higher than sending all nine tokens (signed-rank test, $p = 0.52$). This contrasts with the large difference in the Agreement condition between action 9 and action 10.

In Figure 2, we plot the average appropriateness ratings for the Bertrand game with and without agreement. Hypothesis 1 is not fully supported. Taking a higher action is considered more appropriate up until action 50, and actions greater than 50 are viewed as less appropriate than taking action 50.[34] Hypothesis 2A is fully supported for the Bertrand game: taking action 100 is significantly more appropriate than any other action in the Agreement condition (signed-rank test, $p < 0.01$ for all). A rank-sum test supports Hypothesis 2B: for any action less than action 100, average ratings in the No Agreement condition are greater than average ratings in the Agreement condition ($p = 0.03$ for action 0, $p < 0.01$ for all other comparisons). Choosing action 100 is considered more appropriate in the Agreement condition than in the No Agreement condition ($p < 0.01$). In fact, in the No Agreement condition, the average appropriateness rating increases from 0 to 50, peaks at the middle (action 50), and declines from 50 to 99 (signed-rank test of appropriateness ratings in the No Agreement condition for action 50 > action 40 is $p < 0.01$; action 50 > action 60 is $p = 0.04$). Moreover, in the No Agreement condition there is no significant difference in appropriateness rating between action 50 and action 100 (signed-rank test, $p = 0.57$), though there is in the Agreement condition.

The regression analysis reported in Table 1 supports these results.[35] Columns (1) and (3) report the results of regressing subjects' appropriateness rating for each action on the action number, a dummy for the agreement condition, and an interaction between agreement and action number. This captures the simplest forms of Hypotheses 1, 2A, and 2B, that more prosocial (higher) actions are deemed more appropriate, and that high actions should be particularly appropriate in the Agreement condition. This specification does a reasonable job of capturing the patterns we see in Figure 1 in the double dictator game—there is a positive coefficient on the variable *Action* ($b = 0.275$, $p < 0.01$), and the increase in appropriateness for higher actions becomes steeper in the Agreement condition ($b = 0.078$, $p < 0.01$).

However, this specification is not flexible enough to capture the nonmonotonicity and the sharp discontinuities in the Bertrand game very well (see Figure 2). In the specifications reported in columns (2) and (4) of Table 1, we add an additional dummy variable denoting the "highest action" that captures the "jump" in ratings at the highest action in the Agreement condition and aligns with our a priori prediction (Hypotheses 2A and 2B) that agreements should change the perception of the promised action. We also interact the dummy for highest action with the agreement dummy. In both specifications, it is clear that there is a substantial increase ($b = 2.298$, $p < 0.01$ in the double dictator game; $b = 3.423$, $p < 0.01$ in the Bertrand game) between the highest and next highest actions

**Figure 2.** Average Appropriateness Ratings for the Bertrand Game With and Without Agreement (Data from Module 1)

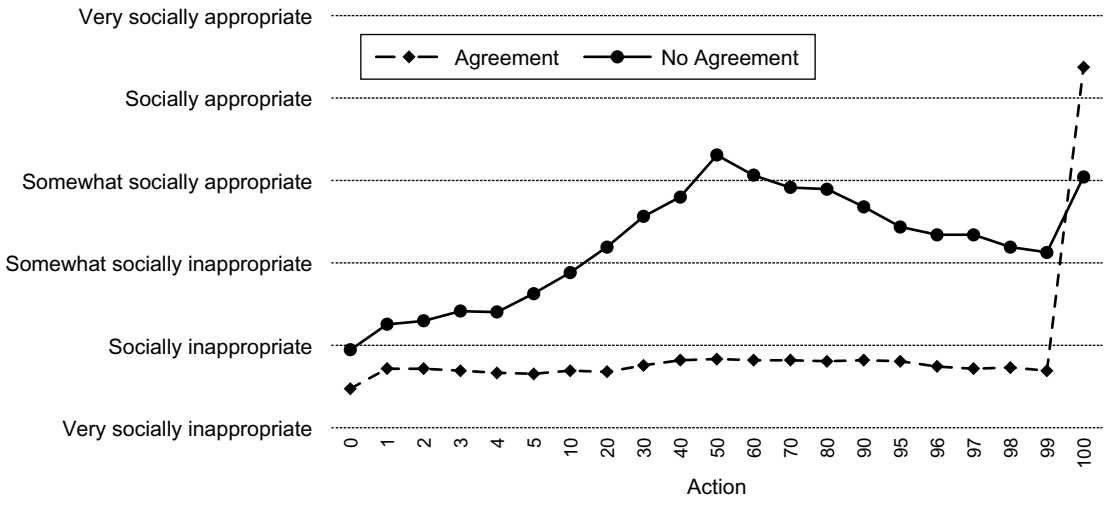


**Table 1.** Ordinary Least Squares Regressions on Appropriateness Ratings for the Double Dictator Game and the Bertrand Game (Ratings Elicited in Module 1 of the Experiment)

| | DDG | | BG | |
|---|---|---|---|---|
| Variable | (1) | (2) | (3) | (4) |
| *Action* | 0.275*** | 0.302*** | 0.0758*** | 0.0749*** |
| | (0.0147) | (0.0155) | (0.00540) | (0.00558) |
| *Agreement* | −1.380*** | −0.962*** | −1.106*** | −0.780*** |
| | (0.143) | (0.151) | (0.116) | (0.126) |
| *Agreement* × *Action* | 0.0781*** | −0.0263 | −0.0178** | −0.0623*** |
| | (0.0187) | (0.0190) | (0.00781) | (0.00660) |
| *Highest Action* | | −0.590*** | | 0.0726 |
| | | (0.207) | | (0.201) |
| *Agreement* × *Highest Action* | | 2.298*** | | 3.423*** |
| | | (0.254) | | (0.274) |
| *Constant* | 2.017*** | 1.910*** | 2.391*** | 2.398*** |
| | (0.122) | (0.131) | (0.0988) | (0.108) |
| Observations | 1,914 | 1,914 | 3,864 | 3,864 |
| No. of subjects | 174 | 174 | 184 | 184 |

*Notes.* The dependent variable is the norm rating for each action in the double dictator game (columns (1) and (2)) and the Bertrand game (columns (3) and (4)). Standard errors clustered at the subject level are reported in parentheses.
***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

in the Agreement condition—an increase not matched in the No Agreement condition. Furthermore, in the Bertrand game, the net effect of the estimated coefficients in the Agreement condition is that the appropriateness ratings should be flat for all actions less than 100, with a sharp increase at action 100. In short, the regressions restate what the graphs show: the effect of an informal agreement is to increase the appropriateness of the agreed-upon action and decrease the appropriateness of all other actions.

In summary, the graphs and supporting regressions show that social norms are significantly different when an agreement has been reached even when all other aspects of the choice environment remain the same. Furthermore, the promise does not just shift the norms ratings up or down but also changes the shape of the profile (see the results in the Bertrand game). Before we turn to running our "horse races," we use our choice experiment data to test for differences in behavior when agreements have been reached.

### 4.2. Choice Experiment: The Effect of Agreement on Behavior

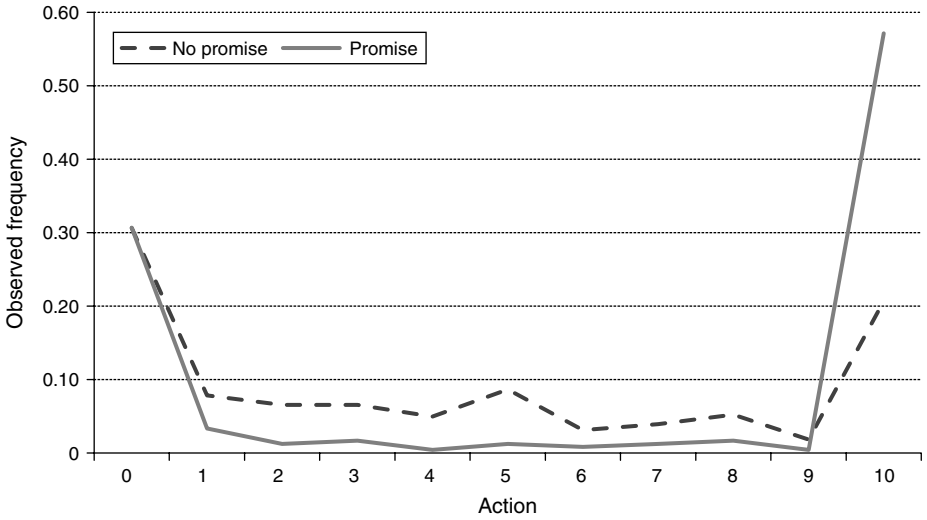
Recall that our choice experiment consisted of a within-subject design. There were 20 rounds. In the first 10 rounds, subjects made decisions in the double dictator game, and in the second 10 rounds they made decisions in the Bertrand game. Within each round, subjects indicated their desire to have an informal agreement to take the first best action. The fraction of subjects who requested an informal agreement across all periods was 89% in the double dictator game and 88% in the Bertrand game. Figures 3 and 4 display the frequency of subjects' action choices with and without an agreement. We find results that are in line with Kessler and Leider's (2012) findings and with the literature on informal agreements: having an agreement increases actions by 50% in the double dictator game and 61% in the Bertrand game.[36]

We confirm these results by regressing subjects' actions on a dummy for whether or not they have an agreement. The estimates are presented in columns (1) and (4) of Table 2 for the double dictator and Bertrand games, respectively.[37] We see that having an agreement in place significantly increases the chosen action in both games ($\beta = 2.686$, $p < 0.01$ in the DDG; $\beta = 30.620$, $p < 0.01$ in the BG). Columns (2) and (5) show that having an agreement significantly increases subjects' first-order beliefs ($\beta = 3.301$, $p < 0.01$ in the DDG; $\beta = 33.382$, $p < 0.01$ in the BG) and second-order beliefs ($\beta = 1.651$, $p < 0.01$ in the DDG; $\beta = 12.511$, $p < 0.01$ in the BG). The significant increase in both actions and first-order beliefs is consistent with Kessler and Leider's (2012) results. The increase in both first-order beliefs and second-order beliefs is consistent with Hypotheses 3A and 3B, and therefore with guilt aversion being a potential mechanism behind promises.

### 4.3. Predicting Choice Behavior Using Social Norms

Thus far, we have used a separate set of subjects to provide us with an independent measure of the social norms for these games and treatments. We have shown evidence that, for each game, promises work to change the social norm governing a decision. We have also shown that when subjects actually play the game, the

**Figure 3.** Average Taken in the Double Dictator Game With and Without Agreement




informal agreement significantly affects their chosen action as well as their first- and second-order beliefs. Because we separately identify the social norms from behavior data, we can now examine whether our measured norms can explain behavior in these games and whether subjects' choices are guided by a desire to comply with the social norm. To do so, we fit individual utility functions to the choice data. Recall that if norms are an important motivation for behavior, then a model that incorporates concerns for norms ought to outperform models that do not.

In the choice experiment, a subject made choices for both agreement conditions and for both games. We assume that individuals have a logistic choice rule, where the likelihood of choosing any action, $a$, depends on the relative utility of that action compared to the other action. We write this as

$$P(a = a_i) = \frac{\exp(U_i)}{\sum_j \exp(U_j)}. \tag{5}$$

Our first specification assumes that utility only depends on one's monetary payoff. (One way to think of this is that we set $\gamma_i = 0$ in Equation (1).) To estimate the weight placed on monetary payoffs, we impose a linear restriction on $V(\,)$ such that for any final payoff, $x$, $V(x) = \beta x$. Additionally, we use subject's first-order

**Figure 4.** Average Taken in the Bertrand Game With and Without Agreement

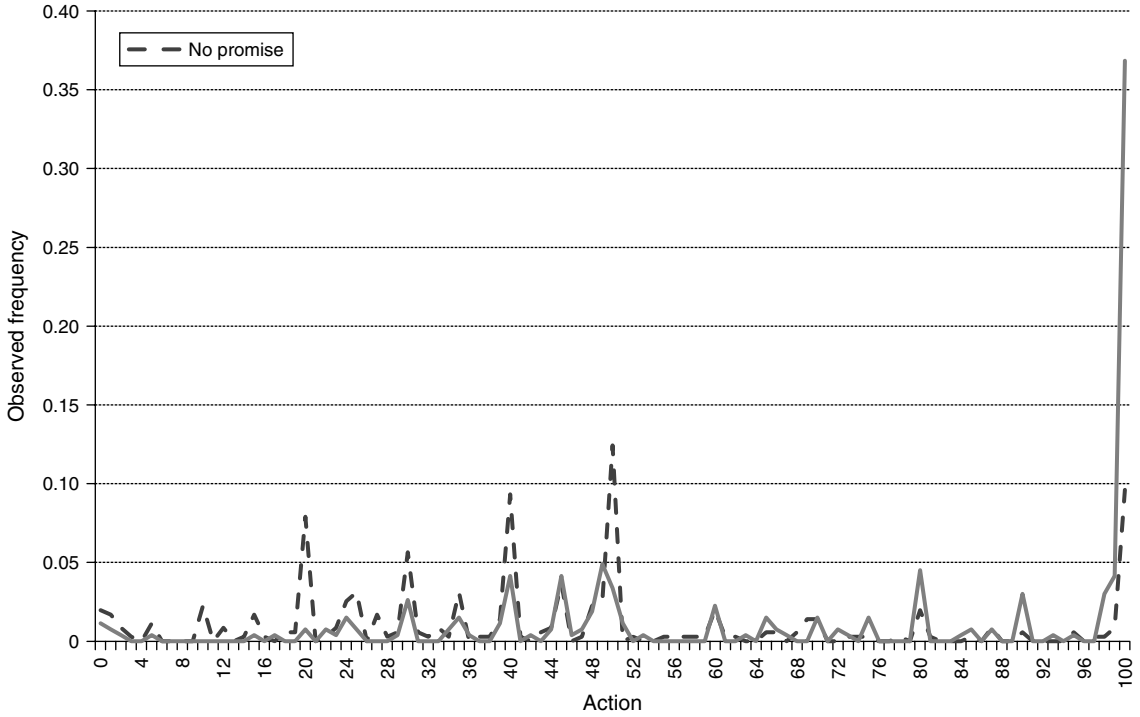


**Table 2.** Statistical Tests of the Effect of Having an Agreement on Behavior and First- and Second-Order Beliefs in the Double Dictator and Bertrand Games

| | DDG | | | BG | | |
|---|---|---|---|---|---|---|
| DV: | *Chosen action* (1) | *FOB* (2) | *SOB* (3) | *Chosen action* (4) | *FOB* (5) | *SOB* (6) |
| *Have agreement* | 2.686*** | 3.307*** | 1.652*** | 30.620*** | 33.382*** | 12.512*** |
| | (0.468) | (0.421) | (0.628) | (4.892) | (4.708) | (3.307) |
| *FOB* | | | 0.614*** | | | 0.630*** |
| | | | (0.082) | | | (0.064) |
| *Constant* | 3.913*** | 5.063*** | 2.158*** | 43.726*** | 54.677*** | 24.413*** |
| | (0.632) | (0.563) | (0.333) | (7.027) | (6.669) | (4.263) |
| Model | | | OLS, RE | | | |
| Observations | 620 | 620 | 620 | 620 | 620 | 620 |
| No. of subjects | 62 | 62 | 62 | 62 | 62 | 62 |
| $R^2$ | 0.054 | 0.146 | 0.553 | 0.1878 | 0.238 | 0.604 |

*Notes.* The dependent variable (DV) is the chosen action (columns (1) and (3)), first-order belief (columns (2) and (4)), and second-order belief (columns (3) and (6)) in the double dictator game (columns (1)–(3)) or in the Bertrand game (columns (4)–(6)). Standard errors are clustered by session and reported in parentheses. FOB, first-order belief; SOB, second-order belief; OLS, ordinary least squares; RE, random effects.

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

beliefs about the other player's action $a^{-i}$. Thus, we estimate the weight, $\beta$, that individuals place on the money they receive from a particular choice as

$$u_i(x, a_k) = \beta_1 x(a_{i,k}, a^{-i}). \qquad \text{(Selfish model)}$$

To investigate whether concern with norm compliance guides behavior, we can estimate Equation (1) using the average appropriateness ratings from module 1 and the behavioral data from the choice experiment.[38] We use a conditional logit regression (McFadden 1974),[39] in which the dependent variable is which action was selected, and the independent variables are the characteristics of the possible action choices (specifically, each action's social appropriateness and its expected monetary payoff). For each alternative, we include the average social appropriateness rating ($N(a_k)$), which varies within game by whether there was an agreement or not. The coefficient for appropriateness ratings[40] provides an estimate of the weight on social appropriates, $\gamma$, in Equation (1) such that

$$u_i(x, a_k) = \beta_1 x(a_{i,k}, a^{-i}) + \gamma N(a_k). \qquad \text{(Norms model)}$$

Table 3 reports the estimation results for the double dictator game and the Bertrand game. We cluster at the subject level, and, because the average norm ratings are a measured quantity that may have sampling error, we use bootstrapped standard errors for the models containing the norm ratings.[41]

In each regression, the reported coefficient reflects the relative weight that each component has in the utility function. For the double dictator game, the coefficient on monetary payoffs, though positive, is small and not different from zero in the Selfish model in column (1) ($\beta = 0.003$; $p > 0.05$). Because a transfer of zero is a dominant strategy in the double dictator game, the purely Selfish model does a poor job of explaining the substantial number of nonzero transfers. However, the coefficient on action payoff is positive and significant when we add social norms as an explanatory variable to the regression in column (2) ($\beta = 0.254$, $p < 0.01$).

For the Bertrand game, the payoff characteristic is positive and significant (columns (3) and (4)) in both specifications—indicating that subjects are more likely to choose actions with higher payoffs. In the Norms model (columns (2) and (4)), we see that for both games the coefficient for the appropriateness rating is positive and statistically significant, signifying that actions that are deemed more appropriate are chosen more often. Additionally, augmenting the Selfish model with the norms ratings increases the model's predictive fit (measured both by the likelihood ratio and the Bayesian information criterion (BIC), which penalizes models for the number of parameters).[42]

Moreover, the influence of social appropriateness on behavior is not just statistically significant but also large in magnitude. The ratio $0.15\gamma/\beta_1$ identifies how much money an individual is willing to sacrifice to gain one category of social appropriateness.[43] To make comparisons between the Bertrand game and the double dictator game, we can estimate the average dollar value (with bootstrapped standard errors in parentheses) subjects would place on an increase in appropriateness for taking a promised action rather than the median action that was actually taken by subjects. We estimate that in the double dictator game, subjects are willing to give up \$2.42 to take the agreed-upon action

**Table 3.** Conditional Logit Estimation of Choice Determinants for the Double Dictator Game and Bertrand Game Using Mean Appropriateness Ratings from Module 1

| | DDG | | BG | |
|---|---|---|---|---|
| Variable | Selfish (1) | Norms (2) | Selfish (3) | Norms (4) |
| *Action Payoff* ($\beta$) | 0.003 | 0.254*** | 0.025*** | 0.032*** |
| | (0.006) | [0.019] | (0.002) | [0.002] |
| *Norm Rating* ($\gamma$) | | 1.449** | | 1.202*** |
| | | [0.263] | | [0.0752] |
| *Monetary Value* ($0.15\gamma/\beta$) | | 0.855*** | | 5.634*** |
| | | [0.059] | | [0.569] |
| Observations | 620 | 620 | 620 | 620 |
| Log likelihood | −1,486.59 | −1,370.86 | −2,770.3 | −2,486.81 |
| Bayesian IC | 2,982.0 | 2,759.37 | 5,551.66 | 4,995.71 |

*Notes.* The dependent variable is the chosen action in the double dictator game (columns (1) and (2)) or in the Bertrand game (columns (3) and (4)). Standard errors are clustered at the subject level and reported in parentheses, with bootstrapped standard errors in brackets for specifications with norm ratings. Each observation represents a subject's choice in a particular period. For the conditional logit estimate, each observation corresponds with 11 possible alternatives for the DDG and 101 possible alternatives for the BG. The variable *Norm Rating* converts subject responses in module 1 to numerical scores: "very socially inappropriate" = 1, "socially inappropriate" = 2, "somewhat socially inappropriate" = 3, "somewhat socially appropriate" = 4, "socially appropriate" = 5, and "very socially appropriate" = 6. The ratio $0.15\gamma/\beta$ identifies how much money an individual is willing to sacrifice to gain one category of social appropriateness. We multiple by 0.15 because each token in the Kessler and Leider (2012) experiments was worth \$0.15. IC, information criterion.

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

("action 10") rather than the average action. When no such agreement exists, they are willing to only give up \$0.83. Thus, honoring the informal agreement is worth giving up an additional \$1.58 in the double dictator game (approximately 10% of the average earnings for the whole session).

By similar calculations, we estimate that in the Bertrand game subjects are willing to give up \$19.92 to take the agreed-upon action ("action 100") rather than the average action; however, they must be paid \$0.10 to take action 100 when no agreement exists. Thus, in the Bertrand game, honoring the informal agreement is worth giving up an additional \$20.02 (approximately 120% of the average earnings for the whole session). The greater willingness to follow promises in the Bertrand game is in line with Miettinen (2013), which predicts a greater effect of promises in games with strategic complements.

To get a sense of how well the social norms model can qualitatively account for the data from the choice experiment, we calculated the predicted frequencies of choices in the two games for the two treatments (Agreement and No Agreement). Figures 5–8 predict the behavior data from the coefficients on the Selfish and Norms models in Table 3 (where we cluster at the subject level and use bootstrapped standard errors for

**Figure 5.** Distributions and Predicted Distributions of Actions Taken in the Double Dictator Game (Predictions Based on Selfish Model Coefficients in Table 3, Model 1)

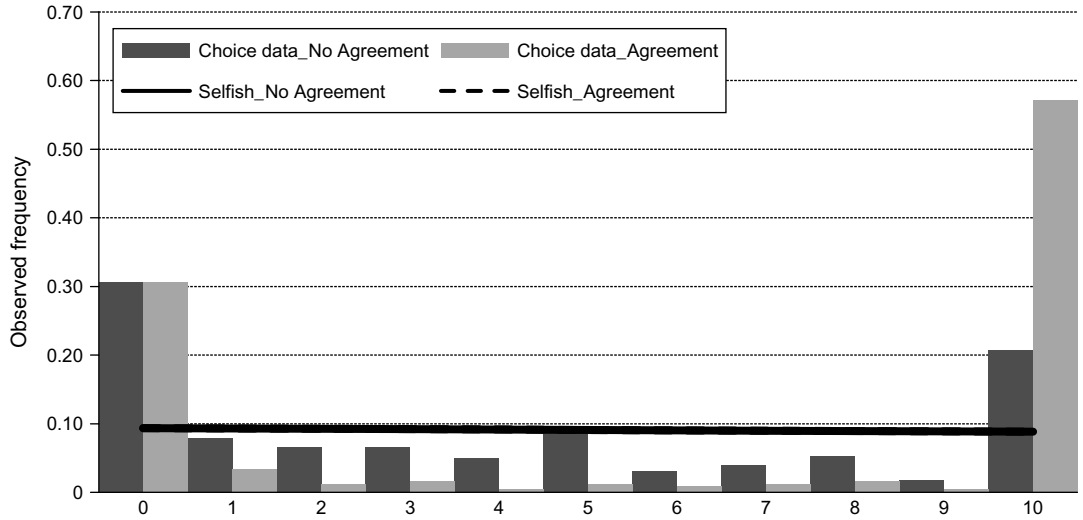


**Figure 6.** Distributions and Predicted Distributions of Actions Taken in the Double Dictator Game (Predictions Based on Norms Model Coefficients in Table 3, Model 2)

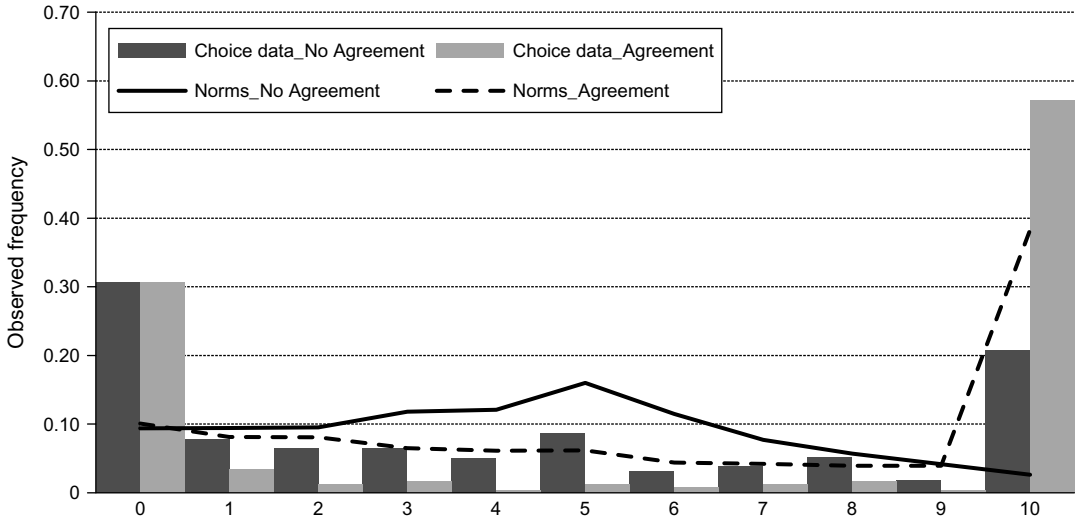



**Figure 7.** Distributions and Predicted Distributions of Actions Taken in the Bertrand Game (Predictions Based on Selfish Model Coefficients in Table 3, Model 3)

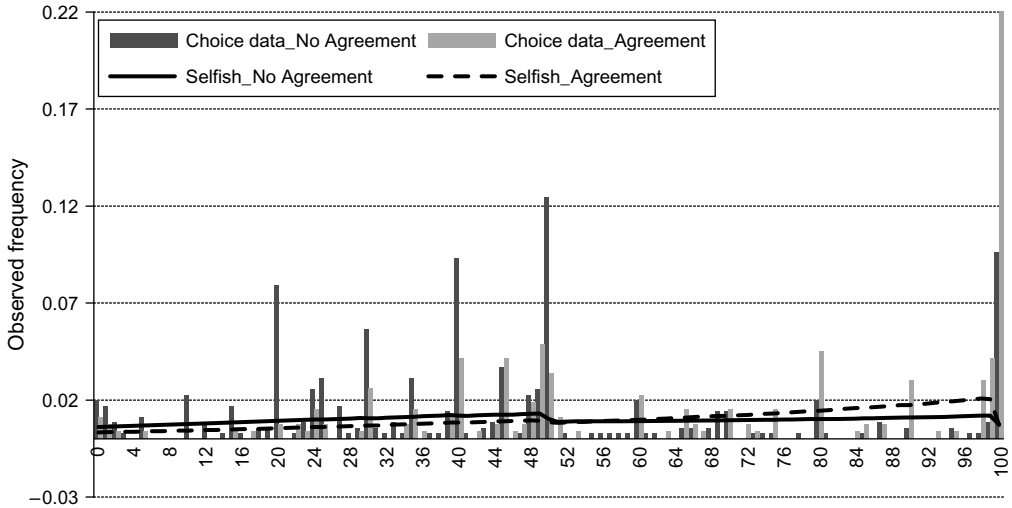
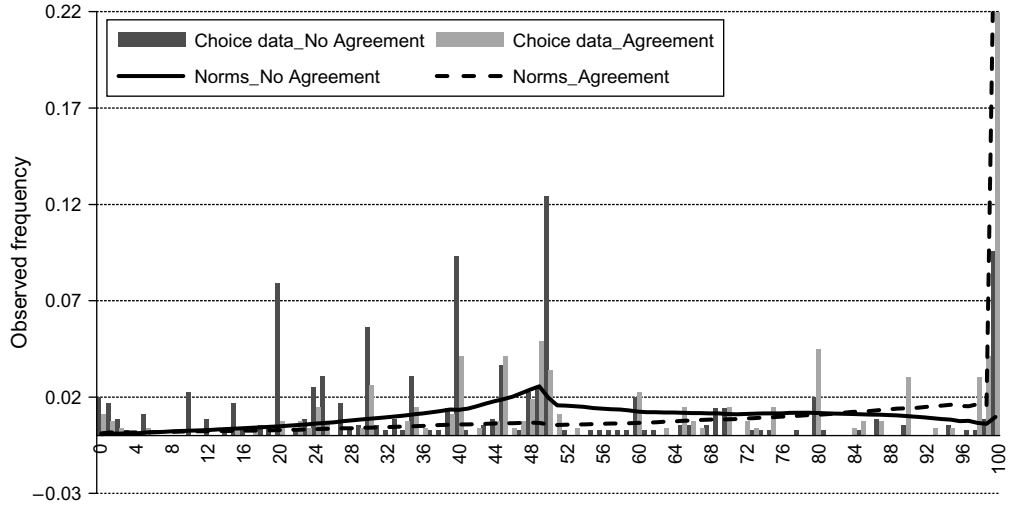



**Figure 8.** Distributions and Predicted Distributions of Actions Taken in the Bertrand Game (Predictions Based on Norms Model Coefficients in Table 3, Model 4)



models containing norms ratings). In the double dictator game, the Selfish model predicts the same distribution of actions for both the Agreement and No Agreement cases (since choosing 0 is the dominant choice even a difference in beliefs cannot lead to different predictions in the Selfish model). However, the Norms model is able to accurately capture the larger share of subjects choosing lower actions (7 and below) in the No Agreement treatment, as well as the large mass of subjects in the Agreement treatment that choose action 10 (although it does not pick up the smaller mass choosing 10 in the No Agreement treatment). In the Bertrand game, the Selfish model (through the change in beliefs) barely captures the upward shift in actions between the No Agreement and Agreement conditions, and it actually predicts a sharp *drop* in the frequency of subjects playing 100 versus 99. The Norms model, however, captures the large number of subjects choosing action 100. It also captures the slight uptick in subjects taking action 50 in the No Agreement condition. Hence, the social norms mechanism not just appears to provide a good statistical fit but also does a good job of capturing the unique importance of fulfilling a promise.

Thus, to summarize, we find that behavior changes across the Agreement and No Agreement treatments in *both* the double dictator game and the Bertrand game can be accounted for by changes in the social appropriateness of seeming identical (in terms of payoffs) actions.

### 4.4. Predicting Choice Behavior with Alternative Models

In addition to demonstrating that the social norms mechanism does a good job of describing the choice data (both quantitatively and qualitatively), we want to consider whether other common mechanisms for the efficacy of agreements can also describe the choice patterns. In particular, we look at *guilt aversion* and *lying aversion*. The lying aversion model is easy to estimate once we pick a functional form, whereas guilt aversion requires information about the second-order beliefs (beliefs about beliefs).

We estimate logistic choice models for guilt aversion and lying aversion using a similar procedure as in the previous section. For guilt aversion, we assume that utility depends on one's own payoff and a measure of guilt aversion based on the difference between the chosen action and one's belief about the other party's expectation (denoted by GA; see Equation (2)). For each game and agreement treatment, we use the second-order beliefs elicited at the end of each round in our choice experiment to form the guilt aversion term for each individual. Hence, we can estimate the relative weight subjects place on this utility component,

$$u_i(x, a_k) = \beta_1 x(a_{i,k}, a^{-i}) - \beta_2 \mathrm{GA}. \qquad \text{(GA model)}$$

To test a model of lying aversion, we assume the cost of lying to increase linearly in the difference between one's action and the promised action (denoted by LA; see Equation (3)). We estimate the lying aversion model as

$$u_i(x, a_k) = \beta_1 x(a_{i,k}, a^{-i}) - \beta_4 \mathrm{LA}. \qquad \text{(LA model)}$$

Finally, in the most general specifications, we assume that utility depends on one's own payoff, guilt aversion, or lying aversion and social norms, respectively, as

$$u_i(x, a_k) = \beta_1 x(a_{i,k}, a^{-i}) - \beta_3 \mathrm{GA} + \gamma N(a_k) \qquad \text{(GA + Norms model)}$$

and

$$u_i(x, a_k) = \beta_1 x(a_{i,k}, a^{-i}) - \beta_4 \mathrm{LA} + \gamma N(a_k). \qquad \text{(LA + Norms model)}$$

In our data, the norms component can be separately identified from both the guilt aversion and lying aversion components. The values of $N(\cdot)$ are set based on the responses of the subjects in the norm elicitation experiment, who do not play the games, and $N(\cdot)$ is assumed to be the same across individuals and rounds when estimating the behavior of subjects in the choice experiment. By contrast, the GA term depends on the measured second-order beliefs from the choice experiment and can therefore vary both across individuals and between rounds. The LA term is different from $N(\cdot)$ by construction: it is defined to be zero in the no promise case, and it is defined to depend linearly on the action chosen in the promise case (although we explore alternatives to this assumption in Online Appendix I, Section 4.4).

Table 4 reports for each game the results of the Guilt Aversion (columns (1) and (5)) and Lying Aversion (columns (3) and (7)) models, as well as the combined GA + Norms (columns (2) and (6)) and LA + Norms (columns (4) and (8)). For both games and both specifications of the Guilt Aversion model, the coefficient on GA has a negative sign and is significant—indicating that subjects are less likely to choose actions associated with high guilt. Similarly, for both games and both specifications of the Lying Aversion model, the coefficient on LA is negative and significant—as expected, subjects prefer not to break their agreement.

Comparing the Norms, GA, and LA models for each game, we find that, overall, the Norms model does fairly well. In the double dictator game, the GA model has the best fit according to the BIC. However the Norms model has also has a good fit, and a Vuong test does not find a significant difference in the fit of the two models (Norms BIC = 2,759.37, GA BIC = 2,573.64, $p$ = 0.213). The LA model has the worst fit of the three, and

**Table 4.** Conditional Logit Estimation of Choice Determinants for the Double Dictator Game and Bertrand Game Using Alternative Mechanisms

| | DDG | | | | BG | | | |
|---|---|---|---|---|---|---|---|---|
| | GA | GA + Norms | LA | LA + Norms | GA | GA + Norms | LA | LA + Norms |
| Variable | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *Action Payoff* ($\beta$) | 0.309*** | 0.485*** | 0.043*** | 0.251*** | 0.046*** | 0.040*** | 0.018*** | 0.036*** |
| | (0.024) | [0.029] | (0.008) | [0.019] | (0.003) | [0.003] | (0.002) | [0.002] |
| *Norm Rating* ($\gamma$) | | 1.083*** | | 1.379*** | | 0.959*** | | 1.360*** |
| | | [0.217] | | [0.325] | | [0.082] | | [0.105] |
| *Guilt Aversion* | −0.136*** | −0.127*** | | | −0.063** | −0.024*** | | |
| | (0.00) | [0.008] | | | (0.003) | [0.004] | | |
| *Lying Aversion* | | | −0.211*** | −0.035 | | | −0.021*** | 0.013*** |
| | | | (0.027) | [0.030] | | | (0.003) | [0.003] |
| *Monetary Value* ($0.15\gamma/\beta$) | | 0.334*** | | 0.826*** | | 3.594*** | | 5.557*** |
| | | [0.048] | | [0.096] | | [0.473] | | [0.530] |
| Observations | 620 | 620 | 620 | 620 | 620 | 620 | 620 | 620 |
| Log likelihood | −1,277.99 | −1,209.87 | −1,455.03 | −1,370.19 | −2,600.29 | −2,469.10 | −2,744.70 | −2,479.46 |
| Bayesian IC | 2,573.64 | 2,446.24 | 2,927.72 | 2,766.86 | 5,222.6 | 4,971.35 | 5,511.49 | 4,992.06 |

*Notes.* The dependent variable is the chosen action in the double dictator game (columns (1)–(4)) or in the Bertrand game (columns (5)–(8)). Standard errors are clustered at the subject level and reported in parentheses, with bootstrapped standard errors in brackets for specifications with norm ratings. Each observation represents a subject's choice in a particular period. For the conditional logit estimate, each observation corresponds with 11 possible alternatives for the DDG and 101 possible alternatives for the BG. The variable *Norm Rating* converts subject responses in module 1 to numerical scores: "very socially inappropriate" = 1, "socially inappropriate" = 2, "somewhat socially inappropriate" = 3, "somewhat socially appropriate" = 4, "socially appropriate" = 5, and "very socially appropriate" = 6. The ratio $0.15\gamma/\beta$ identifies how much money an individual is willing to sacrifice to gain one category of social appropriateness. We multiple by 0.15 because each token in the Kessler and Leider (2012) experiments was worth \$0.15. IC, independent criterion.

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

both Norms and GA models are significant improvements (LA BIC = 2,927.72; Vuong test, $p < 0.01$ for both comparisons). In the Bertrand game, the Norms model has the best overall fit and is a significant improvement over the GA model (Norms BIC = 4,995.71; GA BIC = 5,222.60; Vuong test, $p < 0.01$). As with the DDG, the Lying Aversion model is the worst overall fit and is significantly worse than both the Norms and GA models (LA BIC = 5,511.49; Vuong test, $p < 0.01$ for both comparisons).

If we consider the models that combine guilt and lying aversion with social norms (columns (2) and (4) for the double dictator game and (6) and (8) for the Bertrand game, respectively), we find that the social norms coefficient is positive and much larger than the coefficient on payoffs or on guilt or lying aversion. For example, in the GA + Norms model for the Bertrand game, the coefficient on the payoff is $\beta = 0.485$, that on GA is $\beta = -0.024$, whereas the coefficient on $\gamma$ is 0.959. Similarly, for every action, the average marginal effect from a change in the norm rating is substantially higher than an equivalent change in the guilt.[44] For the double dictator game, the effect in the Norms model of changing the norm was between 5 and 15 times as large as the effect in the GA model of changing the guilt, and 10 to 14 times as large in the GA + Norms model. In the Bertrand game the effect was 7 to 38 times as large in the separate models, and 39 times as large in the combined model. Second, adding social norms to either the GA or LA model leads to a significant improvement in the BIC and, in the case of the Bertrand game, substantially reduces the size of the coefficients on GA (from −0.063 to −0.024) and LA (from −0.021 to +0.013).

Using the BIC and comparing across all models in Tables 3 and 4, we see that a combined model with GA + Norms is most preferred (it has the lowest BIC, 2,446.24, among all models tested). For both games, the GA + Norms model is a significant improvement over both the Norms model and the GA model (likelihood ratio tests, $p < 0.01$ for all comparisons). These results suggest that the relative impact of concern for complying with the social norm is larger than the impact of guilt or lying aversion in the utility function and, moreover, that the desire for social norm compliance has a direct and separate effect on behavior as well as an indirect effect via guilt or lying aversion.

We can also look at the qualitative fit of the data by graphing predicted behavior against actual behavior. Figures 9–12 report the distributions of predicted actions in each game and treatment for both the Guilt Aversion and Lying Aversion models. In the double dictator game, neither model is able to match the Norms model's ability to capture the key fact of a large mass of subjects in the Agreement condition that choose action 10—both the GA and LA models predict much smaller differences in the frequency of actions between 3 and 10, and they do not capture the small

**Figure 9.** Distributions and Predicted Distributions of Actions Taken in the Double Dictator Game (Predictions Based on the Guilt Aversion Model Coefficients in Table 4, Model 1)

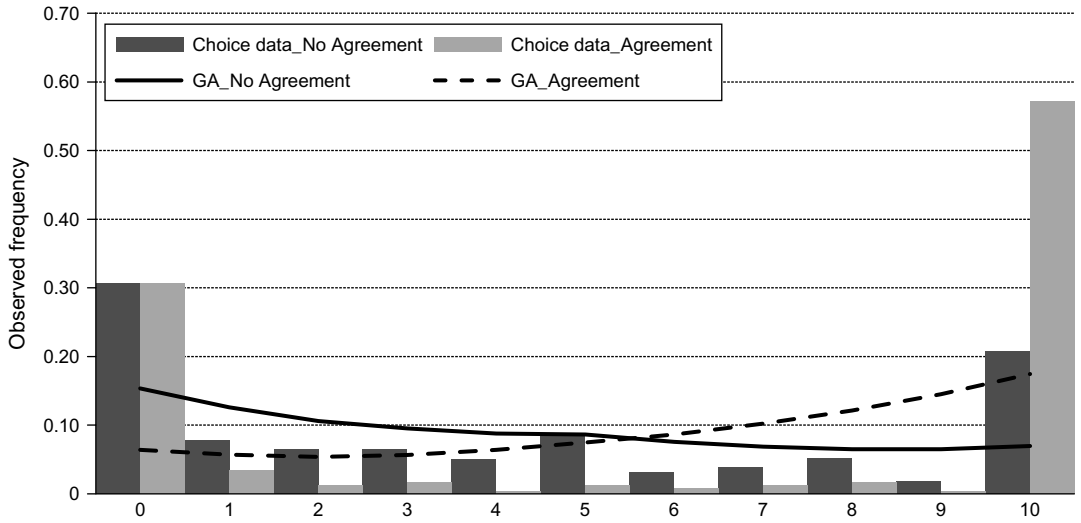



**Figure 10.** Distributions and Predicted Distributions of Actions Taken in the Double Dictator Game (Predictions Based on the Lying Aversion Model Coefficients in Table 4, Model 3)

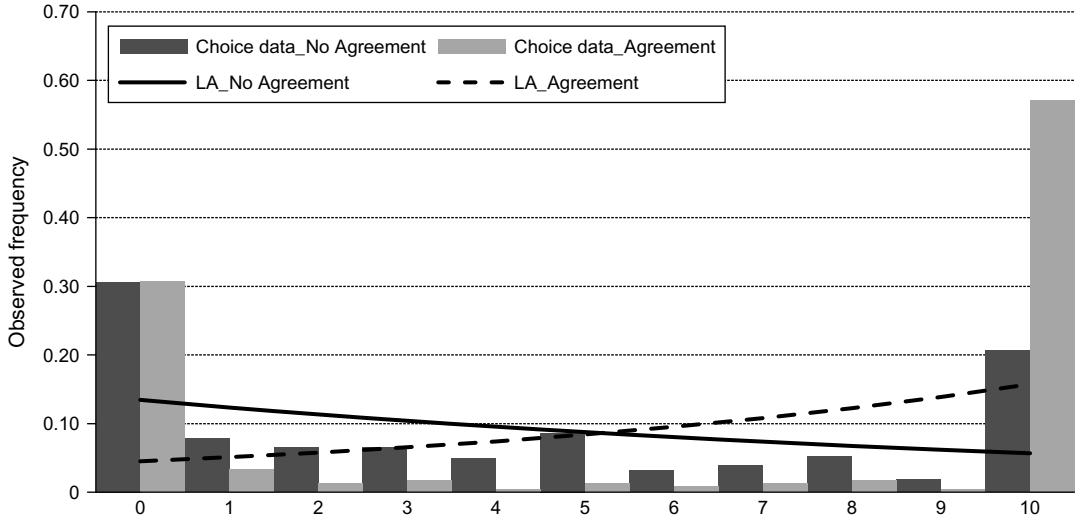



**Figure 11.** Distributions and Predicted Distributions of Actions Taken in the Bertrand Game (Predictions Based on Guilt Aversion Model Coefficients in Table 4, Model 5)

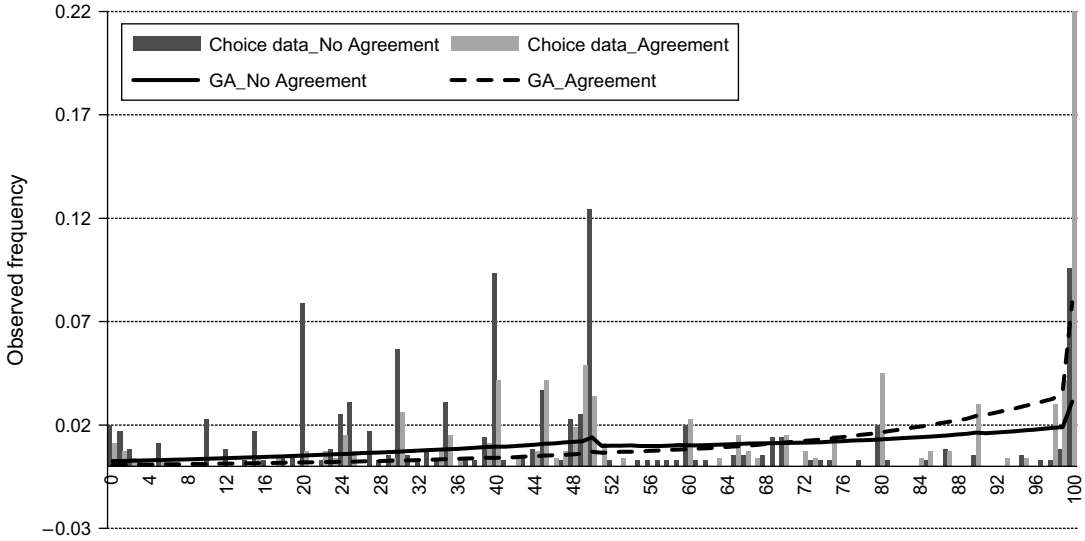
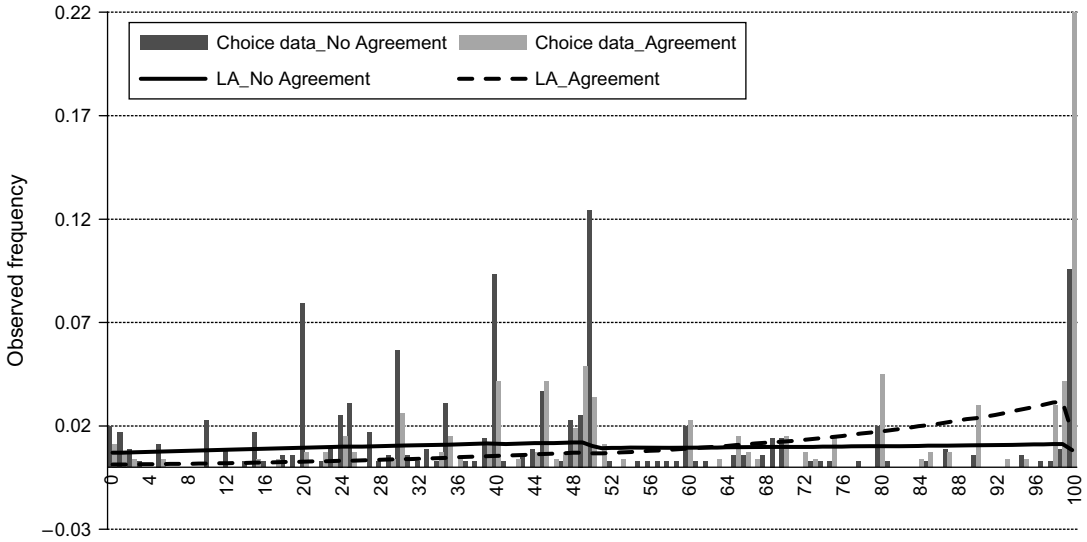


**Figure 12.** Distributions and Predicted Distributions of Actions Taken in the Bertrand Game (Predictions Based on Lying Aversion Model Coefficients in Table 4, Model 7)




uptick at action 5 in the No Agreement treatment, whereas the Norms model does.

We see a similar shortcoming in the Bertrand game. Whereas the Norms model captures the large number of subjects choosing action 100 in the Agreement condition and the small uptick of subjects choosing action 50 in the No Agreement condition, the LA model predicts a *decrease* in the frequency between actions 95–99 and action 100 and almost no uptick at action 50. However, the GA model does predict an uptick in the frequency of choosing action 100, but it does not match the Norms model's ability to capture the magnitude of that uptick, and it predicts a similarly muted uptick at action 50 when there is no agreement. Hence, neither alternative model on its own is as effective as the social norms mechanism on its own for qualitatively capturing the unique importance of fulfilling a promise. Furthermore, the Norms model seems to also do a better job of predicting behavior when no agreement has been reached.

In aggregate, it looks like the GA model on its own does a reasonable job of accounting for behavior; however, the regressions and figures tell a different story. The figures show that once we break out the behavior action by action and look across Agreement *and* No Agreement conditions, the GA model does not capture important moments in these distributions—it gets wrong where the change in behavior will happen (as an example, it does not predict the large spike at 100 in the BG as well as the Norms model does).

The graphical results combined with the regression results suggests that guilt aversion all by itself may be an incomplete model—it can show that for a *given* a change in beliefs, those beliefs will be fulfilled in equilibrium. However, the model has little to say (on its own) about why interacting parties have expectations that informal agreements will be honored and "does not suggest which forms of communication move beliefs" (Charness and Dufwenberg 2006, p. 1595). To explain why parties' expectations are affected by informal agreements, we may point to social norms that shape those expectations.

The LA model can directly explain the effect of the agreement, although not as well as the norms model (evaluated both by BIC and by capturing the importance of fulfilling the promise). Additionally, an aversion to lying may be one particular social norm; however, the social norms framework is more general in that it can also predict behavior in the No Agreement case (where lying aversion is equivalent to selfishness) and in that it is less dependent on functional form assumptions.[45] Hence, models that capture lying aversion reflect the same intuition as our preferred social norms interpretation but provide worse flexibility and explanatory power across conditions.

The norms we elicit can be interpreted to reflect a prohibition against lying about one's intended actions; however, we demonstrate that the specific norms we elicit provide additional information about behavior that a general model of lying cannot fully capture. In particular, elicited norms differ between the double dictator and Bertrand games in how improper it is to make a small deviation from the promised action.

In comparing the models, it is also worth discussing the information upon which each model relies. Of the three mechanisms, lying aversion is the least reliant on measured information (and therefore the most "portable"), making predictions just off the functional form assumptions of the lying costs. Guilt aversion in this setting is in part reliant on measured information (specifically about beliefs) to predict behavior. In a general setting, the guilt aversion model can identify the

sets of beliefs and behaviors that form an equilibrium; however, to predict how communication such as a promise would influence which equilibrium is played, one would either need to add additional assumptions about how beliefs would change (we are not aware of any suggestions for such assumptions beyond those offered by Charness and Dufwenberg 2006) or one would need to directly measure the beliefs (Charness and Dufwenberg 2006 follow the latter approach). This makes the guilt aversion model (applied to informal agreements) less "off the shelf" than the lying aversion model.

The social norms model that we favor similarly relies on measured information (the norm function), and it is therefore arguably of similar "portability" to the guilt aversion model; however, there are at least four reasons to prefer the social norms model. First, the social norms mechanism does a good job at explaining the observed behavior, predicts key moments as well as magnitudes, and is a relatively important determinant of choice even when folded into other models. Thus, the social norms model provides what we feel is a good balance between predictive power and model portability.

Relatedly, we can collect the norm data from third-party subjects who are not playing the game and predict the behavior "out of sample"—suggesting that we are identifying general features of norms rather than just fitting a model ex post to a particular context. Third, the social norms framework is general enough to capture many normative principles—e.g., promise keeping, prosociality, and risk taking. Hence, by collecting norm ratings across a number of different games and decision settings, we can begin to identify features of a decision setting that consistently activate specific normative principles.

Over time, then, we can develop a more general model of what norm functions will be in various settings and construct a portable model that does not rely on measured data on norms. And last, the elicited social norms can guide researchers in making a priori predictions about which forms of communication move beliefs.

## 5. Conclusion

Theory gives social norms a leading role in explaining both the persistence and success of informal agreements. Empirical tests of these theories identify observed behavior consistent with social norms but do not identify the norms directly. In this paper, we elicit social norms separate from behavior and analyze their role in two different games and two different "agreement" conditions. Therefore, we can identify the social norm and then estimate the degree to which actors care to trade off between payoff-related goals and compliance with the social norm.

Our results provide direct evidence of the central role that social norms play in affecting choices in the presence of informal agreements, and they provide evidence that informal agreements affect behavior through their direct effect on the social norm and through an indirect effect by which social norms appear to influence beliefs. Furthermore, we show that the social norms we elicit capture key moments of the choice distribution compared with other mechanisms such as guilt aversion and lying aversion. These results are important because they provide definitive evidence on the most prominent mechanism by which informal agreements are thought to enhance efficiency—social norm compliance.

The evidence also suggests at least two channels by which the act of making an agreement seems to operate on *behavior*: agreement makes a particular norm of obligation salient, and it increases the utility cost of deviating from the obligation. This work also offers compelling new findings regarding how norms vary from environment to environment that can allow for a more general model of norms. In particular, our results in the Bertrand game suggest that strategic complements strongly affect the proscription to comply with an agreement—any action that does not honor that agreement is rated as very socially unacceptable. No such dramatic shift in appropriateness exists when actions are strategically independent and an agreement has been reached.

A strength of our approach is that one need not know the particular social norm (is it a norm of fairness? Of honoring one's obligation? Of not lying?) or the particular manner in which the norm expresses itself ex ante; rather, one can use this technique to characterize the social norm and make and test predictions about behavior that were heretofore not possible. Additionally, by measuring the norms across a variety of decision settings, we can begin to develop a more general model of social norms that can identify what norms are likely to be relevant in a new context based on the features of the decision setting.

### Acknowledgments

The authors thank Tyler Fischer, Caitlin Holman, Jason Johnson, Felicia Kessler, Sally Meyers, and Sarah Pipes for outstanding research assistance. In addition, they thank Abigail Brown, Gary Charness, Rachel Croson, David Danz, Stefano DellaVigna, Martin Dufwenberg, Tore Ellingsen, Ernst Fehr, Simon Gächter, Ulrike Malmendier, Neslihan Uhler, Vera L. te Velde, Roberto Weber, and participants of the Berkeley School of Information, INSEAD, University of Texas at Dallas, and University of Cologne departmental seminars.

### Endnotes

[1] This type of agreement can be thought of as a form of "cheap talk," since the parties engage in "costless," "nonbinding," and "nonverifiable" messages (see Farrell and Rabin 1996).

[2] Promises and informal agreements play a particularly important role in the context of incomplete contracts. Incomplete contracts are extremely common (Tirole 1999, Scott 2003) and can often be more efficient than other more formal contracts (Fehr and Falk 1999, Falk and Kosfeld 2006, Sliwka 2007, Rigdon 2009).

[3] More generally, many people prefer to make truthful statements even when they have a material incentive to lie (Gneezy 2005, Lundquist et al. 2009, Hurkens and Kartik 2009, Özer et al. 2011).

[4] Previous research on the effect of agreements on behavior has appealed to specific descriptions of the social norm such as a norm to reciprocate (Dufwenberg and Kirchsteiger 2000, Malhotra and Murnighan 2002, Dur et al. 2010, Englmaier and Leider 2012), a fairness norm (Fehr and Falk 1999), or a norm to honor obligations and entitlements (Hart and Moore 2008, Fehr et al. 2011). These can be thought of as a description of the social norm that gives a particular interpretation to the behavioral rule associated with the social norm. However, none of this previous work actually elicits the social norm, and as such, it is hard to say whether "reciprocity" or "honoring obligations" is a better description of the behavioral rule associated with promise making. In this paper, we empirically identify the social norm, characterize the behavioral rule, and interpret the rule as "obligation to honoring an agreement."

[5] For intuition on the difference between first- and second-order beliefs and social norms, imagine a typical ultimatum game setting in which the proposer receives an endowment of \$10 and must make a proposal for its division. A proposer might hold a first-order belief that the responder will accept offers of \$4 or higher. A proposer may hold a second-order belief that the responder expects the proposer to offer \$4. However, both the proposer and responder may believe that the prescriptive social norm is that one ought to offer \$5.

[6] To explain tipping behavior, Charness and Dufwenberg (2006) write, "Waiters and waitresses in the United States generally expect a 15% tip; this norm may shape everyone's expectations. Yet, guilt aversion may furnish an underlying motivation for why people behave accordingly. There is a norm, it shapes the server's expectation, and the customer lives up to this expectation because he would feel guilty if he did not" (p. 1596).

[7] Both Charness and Dufwenberg (2006) and Erat and Gneezy (2012) describe the desire for truthfulness—in our context, "keeping one's word"—as a social norm.

[8] Krupka and Weber (2013) provide evidence that social norms apply to situations even when actors have not had a chance to communicate.

[9] The double dictator game is therefore a two-person social dilemma game (see Dawes 1980 for an extensive survey).

[10] Many experiments involving promises (e.g., Charness and Dufwenberg 2006, Vanberg 2008) involve binary decisions. However, in these games there is essentially one moment of interest (the difference in the average choice with and without a promise), such that multiple mechanisms can have equal explanatory power (e.g., for any given difference in beliefs or norms between treatments there may be a coefficient that can justify the observed difference in the mean behavior). In games with many possible actions, there is a richer set of moments to explain, and this offers a better opportunity to test different models.

[11] Strategic complements should lead promises to have a larger impact on behavior. Miettinen (2013) studies a theoretical model of promise keeping that predicts promises will have a greater effect in games with strategic complements.

[12] In a gift-exchange setting with multiple employees, Gächter et al. (2013) use the Krupka and Weber methodology to test the relative explanatory power of distributional preferences and social norms and find that in their setting, distribution preferences have significant explanatory power, but social norms do not.

[13] At least implicitly, most definitions distinguish between social norms and personal norms. The former, which are our focus here, usually refer to a common understanding among members of a group. An individual member of a group has a belief that others in the group judge a particular behavior appropriate (or inappropriate) and that the others in the group assume the individual is aware of this judgment. In this sense, the individual and the group *share an understanding regarding the (in)appropriateness of behavior*, and this shared understanding is a social norm (see Bicchieri 2006, Young 2008).

[14] This is not to say that norms are not also attached to outcomes; rather, these definitions give particular prominence to the actions associated with achieving outcomes. What we find in this paper is that if we maintain this simple assertion (that norms apply to actions rather than outcomes), we can already do much by way of identifying their role in decision making.

[15] In the experiment, we isolate the influence of descriptive norms on responses in the coordination game in two different ways that we describe in Sections 1.1 and 3.1 of Online Appendix I. We show that injunctive social norms concerning the appropriateness of behavior one *ought to engage in* can explain a considerable amount of variation in behavior above and beyond the effect of subjects' beliefs about the descriptive norm.

[16] In this sense, the technique is very similar to hypothetical vignettes used in psychology to identify social norms. Examples include Conroy and Emerson (2006), Ergeneli (2005), McKinney and Moore (2008), Gino et al. (2008), and Oumlil and Balloun (2009). However, the Krupka and Weber (2013) technique adds incentives and the coordination game structure. In this paper, we add a proper scoring rule, we extend the protocol to elicit beliefs about the actual behavior of subjects playing these games, and we expand the ratings from four to six appropriateness categories.

[17] Camerer and Fehr (2004) note that coordination games can be used with economic incentives to reveal shared understanding. They go on to suggest that experimental paradigms, such as simple coordination games, could prove useful for measuring dimensions of shared perception. See also Leider et al. (2009).

[18] Krupka et al. (2008) show that social norms elicited using the coordination exercise track ex ante identified social norms, and Burks and Krupka (2012) show that social norms elicited using the coordination game are distinct from personal opinions (which are elicited without the coordination game structure and without incentives), and they demonstrate the separate effect of personal opinions and social norms on behavior (see also Schwartz 1973).

[19] We take as a starting framework that all individuals in the group jointly agree on $N(a_k)$; however, it is clear that empirically there will likely be disagreement/miscoordination. In general, one would expect that the injunctive norm will have less influence on behavior when there is greater disagreement about $N(\cdot)$.

[20] That is, a norm is not necessarily a binary classification, such that a particular action (the "norm," e.g., "tip 20%" or "the 50–50 split") should be taken, by assumption leaving all remaining actions as those (equally inappropriate) actions that should not be taken. Such a definition is possible in our framework (by, for example, assigning $N(a_k) > 0$ to only one action (the "norm") and letting all other actions have a constant value of $N(a_k) < 0$) but is an oversimplification of how norms appear to operate. In Krupka and Weber (2013), the authors demonstrate that differences in the relative appropriateness of the other actions exert an important influence on behavior. Thus, we characterize the norm as it affects the appropriateness of the entire set of actions.

[21] In this paper we are focused on measuring the norm function $N(\cdot)$ in a particular setting. We do not propose a general model of what the norm function is likely to be in various settings, although this is certainly an important and interesting question for future research.

[22] Several researchers have noted that there exists heterogeneity among individuals for the degree to which they care about complying with a social norm (see Ostrom 2000, Fisher and Huddart 2008), and such heterogeneity in prosocial concern is also common in most models of social preferences (Fehr and Schmidt 1999, Andreoni and Miller 2003, Benabou and Tirole 2006).

[23] Assuming a linear cost of lying is common in the literature (e.g., Özer et al. 2011). We also consider disutility increasing in the square of the difference, as well as a constant penalty for lying (as in Ellingsen and Johannesson 2004) and find qualitatively similar results. See Online Appendix I, Section 4.

[24] In the No Agreement treatment, the lying aversion term is defined to be zero, as no action was promised.

[25] In the original Kessler and Leider (2012) experiment, all the possible agreements were fixed exogenously by the experimenter to increase the number of comparable observations across treatments. We follow this protocol, but see Dufwenberg et al. (2011) for an experiment that endogenizes the content of the unenforceable agreement.

[26] The norm elicitation experiment contained five modules in total. However, the first module always elicited the injunctive social norms, and it is the focus of our analysis and the only data we use for this paper. Modules 2–5 always followed in the same order and are used for various robustness checks not reported in this paper. These four modules collect data on individual beliefs and personal characteristics, and they remeasure the injunctive norm after subjects observe others' behavior. In Online Appendix I, we briefly outline Modules 2–5, their role in our empirical strategy, and our analysis of the results. A full set of instructions can be found in Online Appendix II.

[27] Both vignettes are abbreviated here for exposition purposes. The entire set of instructions is available and can be found in Online Appendix II.

[28] In addition, subjects were also tested on their comprehension of the situation with an interactive quiz, in which they calculated the payoffs of both players, A and B, in three hypothetical situations. They were not allowed to proceed until they got all the calculations correct.

[29] For the double dictator game, subjects were asked to rate all 11 possible actions. It was infeasible to ask subjects to rate all 101 actions in the Bertrand game, so instead we asked them to rate 21 actions (0, 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, and 100). We therefore see ratings that span the action space and get rich data on the ratings for the actions at the extreme ends of the action space.

[30] The decision screen is depicted in Online Appendix I, Figure S1.

[31] While Krupka and Weber (2013) use four categories of appropriateness, we expanded the categories to six (that ranged across very socially inappropriate, socially inappropriate, somewhat socially inappropriate, socially appropriate, somewhat socially appropriate, and very socially appropriate). We did so to be able to pick up finer gradations in the norm; this was especially important in the Bertrand game, which has a larger action space than the games studied by Krupka and Weber.

[32] We chose to elicit an estimate of the median because (unlike a quadratic scoring rule to elicit the mean) this yields fewer extreme ratings when the distribution of the other's ratings is particularly skewed (as might be the case for actions that are, as an example, extremely self-regarding or other-regarding). Furthermore, while there may be no changes in the modal rating an action receives, the median rating can change between treatments. As an example, even if the modal rating for taking the most prosocial action is unchanged when there is an agreement or not, the degree to which appropriateness ratings vary for actions that deviate from the most prosocial action may vary when an agreement is in place. This, in turn, will change the median rating.

[33] In so doing we are imposing ratio scale characteristics on measurements that are in design ordinal. In some of what follows, this is merely for convenience, such as when we use a rank-order test for the equality of distributions. But on other occasions it implicitly adds extra assumptions upon which our analysis is then conditional, such as when we compare means.

[34] After the experiment, subjects were asked (via free response questions) to describe how they decided whether an action was appropriate or inappropriate. Their responses suggested that there may be two relevant norms in the Bertrand game with No Agreement: prosociality and risk avoidance. Many subjects described actions above 50 as being "too risky" while action 50 had an appropriate amount of risk. For example, one subject said that an action like 50 "is high enough where Individual B would not be upset with me low guessing and them losing all their tokens they sent. But it isn't too high where I am risking losing all of my tokens." Another said that 50 was the most appropriate "because I feel this is the best way of hedging my bet." (Additional responses are available upon request.) This notion of risk did not seem relevant when an agreement had been made.

[35] In these specifications, we cluster the standard errors at the subject level.

[36] Kessler and Leider (2012) find that agreements increase the average action by 42% for the double dictator game and by 44% for the Bertrand game. Furthermore, we ran all the subsequent analyses and specifications reported in this paper on the original Kessler and Leider data as well and found qualitatively similar results. We can share a copy of that analysis upon request.

[37] In these specifications, we cluster the standard errors at the session levels.

[38] For the Bertrand game, we use linear interpolation to determine the appropriateness of the actions that we did not explicitly measure. The programs that produce these interpolations are available.

[39] Conditional logit models are similar to multinomial logit models; however, conditional logit models emphasize the characteristics of the alternatives, whereas multinomial logit models depend on the characteristics of the individual making the choice. See Hoffman and Duncan (1988) for a comparison between these models.

[40] We restrict gamma to be the same for everyone ($\gamma_i = \gamma > 0$).

[41] To construct the bootstrapped standard errors, we conducted 500 replications. In each replication, we resample (with replacement) from the norm rating data (generated from the norm elicitation experiment) and construct an average norm function N( ). We then reestimate the choice model based on the sampled norm function. The distribution of the coefficients across replications generates the standard errors.

[42] A likelihood ratio test shows that in both games the model with social norms is significantly preferred over the Selfish model, consistent with the lower BIC ($p < 0.01$).

[43] We multiple by 0.15 because each token in the choice experiment is worth \$0.15.

[44] To keep the average marginal effects (AMEs) comparable, we consider a change of 1/1,000 of a standard deviation in the norm rating and in guilt for each individual action. We calculated the effect of such a change on the predicted probability of choosing that action. For example, in the GA + Norms model for the double dictator game, the AME for an increase in the norm rating for choosing action 10 is 0.15, whereas the AME for an increase in the guilt for action 10 is −0.01. Similarly, in the GA + Norms model for the Bertrand game, the AME for an increase in the norm for action 100 is 0.09, whereas the AME for an increase in the guilt for action 100 is −0.002. The full set of AMEs for all actions is available from the authors.

[45] The chosen functional form for lying aversion may impose some empirical restrictions. As an example, in our experiment, within the

Agreement treatment, the effect of a linear cost of lying is perfectly collinear with the underlying payoff structure. Additionally, a single functional form for lying aversion may not be the best fit for behavior across multiple games. In Table S8 of Online Appendix I, we estimate lying aversion with a fixed cost, a linear cost, and a quadratic cost of lying. In our data, the quadratic model is somewhat of a better fit for the double dictator game (BIC = 624.29 versus 645.43; Vuong test, $p = 0.152$), whereas in the Bertrand game the fixed cost of lying is a much better fit (BIC = 1,805.20 versus 2,300.39 and 2,369.94; Vuong test, $p < 0.01$ for both; see Vuong 1989).

## References

Andreoni J, Miller J (2003) Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* 70(2):737–753.

Battigalli P, Dufwenberg M (2007) Guilt in games. *Amer. Econom. Rev. Papers Proc.* 97(2):170–176.

Benabou R, Tirole J (2006) Incentives and prosocial behavior. *Amer. Econom. Rev.* 96(5):1652–1678.

Bettenhausen K, Murnighan J (1991) The development of an intragroup norm and the effects of interpersonal and structural challenges. *Admin. Sci. Quart.* 36(1):20–35.

Bicchieri C (2006) *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge University Press, New York).

Bicchieri C, Xiao E (2009) Do the right thing: But only if others do so. *J. Behav. Decision Making* 22(2):191–208.

Burks S, Krupka E (2012) A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry. *Management Sci.* 58(1):203–217.

Camerer CF, Fehr E (2004) Measuring social norms and preferences using experimental games: A guide for social scientists. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, eds. *Foundations of Human Sociality—Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies* (Oxford University Press, Oxford, UK), 55–95.

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

Cialdini R, Reno R, Kallgren C (1990) A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *J. Personality Soc. Psych.* 58(6):1015–26.

Conroy S, Emerson T (2006) Changing ethical attitudes: The case of the Enron and ImClone scandals. *Soc. Sci. Quart.* 87(2):395–410.

Chen Y, Kartik N, Sobel J (2008) Selecting cheap-talk equilibria. *Econometrica* 76(1):117–136.

Dawes RM (1980) Social dilemmas. *Annual Rev. Psych.* 31:169–193.

Dawes RM, McTavish J, Shaklee H (1977) Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *J. Personality Soc. Psych.* 35(1):1–11.

Dawes RM, Orbell J, van de Kragt AJ (1988) Not me or thee but we: The importance of group identity in eliciting cooperation in dilemma situations. *Acta Psych.* 68(1):83–97.

Deutsch M, Gerard H (1955) A study of normative and informational social influences upon individual judgment. *J. Abnormal Soc. Psych.* 51(3):629–36.

Dufwenberg M, Gneezy U (2000) Price competition and market concentration: An experimental study. *Internat. J. Indust. Organ.* 18(1):7–22.

Dufwenberg M, Kirchsteiger G (2000) Reciprocity and wage undercutting. *Eur. Econom. Rev.* 44(4–6):1069–1078.

Dufwenberg M, Servátka M, Vadovič R (2011) ABC on deals. Working paper, University of Arizona, Tucson.

Dufwenberg M, Gneezy U, Goeree JK, Nagel R (2007) Price floors and competition. *Econom. Theory* 33(1):211–224.

Dur R, Non A, Roelfsema H (2010) Reciprocity and incentive pay in the workplace. *J. Econom. Psych.* 31(4):676–686.

Ellingsen T, Johannesson M (2004) Promises, threats and fairness. *Econom. J.* 114(495):397–420.

Elster J (1989) *The Cement of Society: A Study of Social Order, Studies in Rationality and Social Change* (Cambridge University Press, Cambridge, UK).

Englmaier F, Leider S (2012) Contractual and organizational structure with reciprocal agents. *Amer. Econom. J.: Microeconom.* 4(2):146–183.

Erat S, Gneezy U (2012) White lies. *Management Sci.* 58(4):723–733.

Ergeneli A (2005) A cross-cultural comparison of ethical behavior in business related dilemmas: A comparison among Turkish, Egyptian, Kirghiz and Kazak marketing employees. *Problems Perspect. Management* 3(2):135–147.

Falk A, Kosfeld M (2006) Distrust—The hidden cost of control. *Amer. Econom. Rev.* 96(5):1611–1630.

Farrell J, Rabin M (1996) Cheap talk. *J. Econom. Perspect.* 10(3):103–118.

Fehr E, Falk A (1999) Wage rigidity in a competitive incomplete contract market. *J. Political Econom.* 107(1):106–134.

Fehr E, Gächter S (2000) Fairness and retaliation: The economics of reciprocity. *J. Econom. Perspect.* 14(3):159–181.

Fehr E, Schmidt K (1999) A theory of fairness, competition, and cooperation. *Quart. J. Econom.* 114(3):817–868.

Fehr E, Hart O, Zehnder C (2009) Contracts, reference points, and competition—Behavioral consequences of the fundamental transformation. *J. Eur. Econom. Assoc.* 7(2–3):561–572.

Fehr E, Hart O, Zehnder C (2011) Contracts as reference points-experimental evidence. *Amer. Econom. Rev.* 101(2):493–525.

Fisher P, Huddart S (2008) Optimal contracting with endogenous social norms. *Amer. Econom. Rev.* 98(4):1459–1475.

Gächter S, Nosenzo D, Sefton M (2013) Peer effects in pro-social behavior: Social norms or social preferences? *J. Eur. Econom. Assoc.* 11(3):548–573.

Gino F, Moore DA, Bazerman MH (2008) No harm, no foul: The outcome bias in ethical judgments. Working Paper 08-080, Harvard Business School, Boston.

Gneezy U (2005) Deception: The role of consequences. *Amer. Econom. Rev.* 95(1):384–394.

Hart O, Moore J (2008) Contracts as reference points. *Quart. J. Econom.* 123(1):1–48.

Hoffman SD, Duncan GJ (1988) Multinomial and conditional logit discrete-choice models in demography. *Demography* 25(3): 415–427.

Hurkens S, Kartik N (2009) Would I lie to you? On social preferences and lying aversion. *Experiment. Econom.* 12(2):180–192.

Kessler J, Leider S (2012) Norms and contracting. *Management Sci.* 58(1):62–77.

Krupka E, Weber R (2009) The focusing and informational effects of norms on pro-social behavior. *J. Econom. Psych.* 30(3):307–320.

Krupka E, Weber R (2013) Identifying norms using coordination games: Why does dictator game sharing vary? *J. Eur. Econom. Assoc.* 11(3):495–524.

Krupka E, Weber R, Croson R (2008) When in Rome: Identifying social norms as a group phenomenon. Working paper, University of Michigan, Ann Arbor.

Lambert N, Shoham Y (2009) Eliciting truthful answers to multiple-choice questions. *EC '09: Proc. 10th ACM Conf. Electronic Commerce* (ACM, New York), 109–118.

Leider S, Möbius M, Rosenblat T, Do Q (2009) Directed altruism and enforced reciprocity in social networks. *Quart. J. Econom.* 124(4):1815–1851.

Loomis JL (1959) Communication, the development of trust, and cooperative behavior. *Human Relations* 12(4):305–3l5.

López-Pérez R (2008) Aversion to norm-breaking: A model. *Games Econom. Behav.* 64(1):237–267.

Lundquist T, Ellingsen T, Gribbe E, Johannesson M (2009) The aversion to lying. *J. Econom. Behav. Organ.* 70(1–2):81–92.

Malhotra D, Murnighan J (2002) The effects of contracts on interpersonal trust. *Admin. Sci. Quart.* 47(3):534–559.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

McKinney JA, Moore CW (2008) International bribery: Does a written code of ethics make a difference in perceptions of business professionals? *J. Bus. Ethics* 79(1/2):103–111.

Mehta J, Starmer C, Sugden R (1994) The nature of salience: An experimental investigation of pure coordination games. *Amer. Econom. Rev.* 84(3):658–673.

Miettinen T (2013) Promises and conventions—An approach to preplay agreements. *Games Econom. Behav.* 80(1):68–84.

Orbell JM, van de Kragt AJ, Dawes RM (1991) Covenants without the sword: The role of promises in social dilemma situations. Koford K, Miller J, eds. *Social Norms and Economic Institutions* (University of Michigan Press, Ann Arbor), 117–134.

Ostrom E (2000) Collective action and the evolution of social norms. *J. Econom. Perspect.* 14(3):137–158.

Oumlil A, Balloun J (2009) Ethical decision-making differences between American and Moroccan managers. *J. Bus. Ethics* 84(4): 457–478.

Özer Ö, Zheng Y, Chen K-Y (2011) Trust in forecast information sharing. *Management Sci.* 57(6):1111–1137.

Rigdon M (2009) Trust and reciprocity in incentive contracting. *J. Econom. Behav. Organ.* 70:(1–2):93–105.

Sally D (1995) Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality Soc.* 7(1):58–92.

Schelling T (1960) *The Strategy of Conflict* (Harvard University Press, Cambridge, MA).

Schwartz S (1973) Normative explanations for helping behavior: A critique, proposal and empirical test. *J. Experiment. Soc. Psych.* 9(4):349–364.

Scott R (2003) A theory of self-enforcing indefinite agreements. *Columbia Law Rev.* 103(7):1641–1699.

Sliwka D (2007) Trust as a signal of a social norm and the hidden costs of incentive schemes. *Amer. Econom. Rev.* 97(3):999–1012.

Sugden R (1995) A theory of focal points. *Econom. J.* 105(430): 533–550.

Tirole J (1999) Incomplete contracts: Where do we stand? *Econometrica* 67(4):741–781.

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

Vuong Q (1989) Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* 57(2):307–333.

Young P (1998) Social norms and economic welfare. *Eur. Econom. Rev.* 42(5):821–830.