A Meeting of the Minds:

Informal Agreements and Social Norms

Erin L. Krupka (School of Information, University of Michigan & IZA)[1]
Stephen Leider (Ross School of Business, University of Michigan)
Ming Jiang (School of Information, University of Michigan)

October 31, 2014[2]

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 their 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.

JEL Codes: C93, D23

[1] Corresponding author: ekrupka@umich.edu

[2] We wish to thank Tyler Fischer, Caitlin Holman, Jason Johnson, Felicia Kessler, Sally Meyers and Sarah Pipes for outstanding research assistance. In addition, we wish to thank Abigail Brown, Gary Charness, Rachel Croson, David Danz, Stefano DellaVigna, Martin Dufwenberg, Tore Ellingsen Simon Gächter, Ulrike Malmendier, Neslihan Uhler, Vera L. te Velde, Roberto Weber, participants of the School of
Information, Berkeley, INSEAD, University of Texas – Dallas and Cologne departmental seminars, the ESA Tuscon 2010 conference and the International ESA conference 2011.

1. Introduction

Many transactions are supported by verbal promises or other informal agreements rather than formal contracts.[3] For example, the motto of the London Stock Exchange is "My Word is My Bond." Several recent papers demonstrate that such promises have a substantial impact on an individual's behavior, 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).[4,5] These papers in economics build on an earlier literature is 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 frame-work. Further, the empirical work either offers only indirect evidence (c.f. Kessler and Leider 2012) or provides only a partial explanation for behavior (c.f. Charness and Dufwenberg 2006; Ellingson 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. Further, we compare social norms to guilt aversion and lying aversion and identify which mechanism (or combination of mechanisms) can best explain behavior in a setting both where informal agreements are present or absent.

[3] This type of agreement can be thought of as a form of 'cheap talk' since the parties engage in 'costless', 'non-binding' and 'non-verifiable' messages (see Farrell and Rabin 1996).

[4] 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).

[5] 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; Ozer et al. 2011).

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 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 (Lopez-Perez 2008; Kessler and Leider 2012).[6] Under the "social norms mechanism", changes in behavior stem from changes in an action's appropriateness when an informal agreement is present. One key feature that can distinguishes the social norms mechanism from lying aversion is that norms can also affect behavior when no agreement is present.[7] Social norms are distinguished from guilt aversion in that social norms are common across individuals, while guilt aversion depends on an individual's personal beliefs.

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 posits 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.[8] 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

[6] 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 give 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".

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

[8] 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.

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 argue that norms shape expectations, and deviation from those expectations generates guilt.[9] Echoing this intuition, 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[10] 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 was reached; in such cases, lying aversion is not an appropriate model while 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.

[9] To explain tipping behavior, Charness and Dufwenberg 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." (2006, p. 1596)

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

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 and a Bertrand Game (c.f. 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[11]. 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 zero. 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.[12] To implement the games in our *choice experiment*, we replicate the Kessler and Leider 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 (rather than just one) 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.[13] 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 which 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.

[11] The Double Dictator Game is therefore a two-person social dilemma game (see Dawes 1980 for an extensive survey).

[12] 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.

[13] Strategic complements should lead promises to have a larger impact on behavior. Miettinen (2008) studies a theoretical model of promise keeping that predicts promises will have a greater effect in games with strategic complements.

To collect data on the social norms, we conduct a separate *norms 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 pro-social action significantly and substantially changes the social norm: fulfilling the promise *increases* in appropriateness, while taking even very pro-social 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 social norms improve our explanatory power across games and agreement conditions and also capture 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 while the guilt aversion model does a better job of explaining behavior in one game but does not do well across both games. Further, 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, while we use largely the same methods as Krupka and Weber (2013) the goals and setting of this paper are different. Krupka and Weber is focused on demonstrating that a social norms framework is a viable explanation for (unilateral) dictator behavior, while in this paper we use the norms data to distinguish three separate mechanisms in strategic games.[14] Our results provide evidence on a

[14] 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.

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 non-binding 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 individual dislike disappointing others. Here a promise serves to change the beliefs of the other party, which then increases the costs of choosing actions less than that belief. Finally, one can directly assume a psychological cost for lying. Therefore, and 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 behavior, while 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.

1.3 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 (cf. Krupka and Weber 2013). Second, the "social" element of norms requires that they be jointly recognized, or collectively perceived, by members of a population.[15] These two

[15] 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

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).[16]

Further, 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 as prescribing what one "should do" or "should not do".[17] 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, then 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".[18] 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

(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 (cf. Bicchieri 2006; Young 2008).

[16] This is not to say that norms aren't 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.

[17] In the experiment we isolate the influence of descriptive norms on responses in the coordination game in two different ways that we describe in the appendix. 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.

[18] In this sense, the technique is very similar to hypothetical vignettes used in psychology to identify social norms. Recent examples include Conroy and Emerson 2006, Ergeneli 2005, McKinney and Moore 2007, Gino et al. 2008, Oumlil and Balloun 2009. However, the Krupka and Weber technique adds incentives and the coordination game structure. In this paper we add a proper scoring rule and we extend the protocol to elicit beliefs about the actual behavior of subjects playing these games.

socially appropriate or inappropriate, and to respond accordingly.[19] 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 sections 2 and 3 of the Appendix we describe several tests of this assumption).[20] 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 $N(a_k) > 0$.[21] 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

[19] 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).

[20] 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.).

[21] 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(.).

actions available to a decision maker that stems from the social norm the researcher is trying to measure.[22, 23]

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 a^{-i}), and the degree to which the action is collectively perceived as socially appropriate:

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

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 a standard social preference models (c.f. 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.[24] An individual entirely unconcerned with social norms (γ_i =0) will always select the payoff-maximizing action. On the other hand, as γ_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 in general not

[22] 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 over simplification of how norms appear to operate. In Krupka and Weber (2012), 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.

[23] In this paper we are focused on measuring the norm function N(.) 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.

[24] Several researchers have noted that there exists heterogeneity among individuals for the degree to which they care about complying with a social norm (cf. Ostrom 2000; Fisher and Huddart 2008) and such heterogeneity in pro-social concern is also common in most models of social preferences (Fehr and Schmidt 1999; Andreoni and Miller 2002; Benabou and Tirole 2006).

possible to separately identify both γ_i and N(.) from behavioral data alone. This is why it is necessary to have some independent means of identifying either γ_i or N(.). Our approach is to use a separate group of subjects to empirically identify N(.) using the *norm elicitation experiment*. We then combine the empirical measure of N(.) with the behavioral data from our *choice experiment* and are able to estimate γ.

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) find that subjects judge pro-social 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 pro-social 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 pre-play communication of various forms can increase the prosociality of individual behavior (see for example Dawes et al. 1977 for an early experiment, and Sally 1995 for an early survey). 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" while 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 to the no agreement case, an agreement will increase the appropriateness of the agreed upon action, and (weakly) decrease the appropriateness of all other actions.*

In the rest of the paper, we predict and explain behavior using elicited measures of social appropriateness ($N(a_k)$). In the *norm elicitation experiment*, we elicit the norms over possible action choices in the two games (Bertrand and in Double Dictator Game) and agreement conditions (when there exists an agreement to take the first best action and when no such agreement exists). In the final section of the paper, we use data collected in our *choice experiment* to test how well the elicited social norms, when integrated into the above simple utility framework, explain the actual choices made by subjects in these games.

2.3 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 to a non-promise 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:

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

In addition to testing for the effect of guilt on choices, we can also direct test the assumption that promises change first- and second-order beliefs, since this is a necessary pre-condition for the guilt aversion mechanism. We expect to find similar results as Charness and Dufwenberg 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, Ozer 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 in the difference between one's action and the promised action a_k^* (we will refer to this 'lying aversion' term as 'LA')[25, 26]:

[25] Assuming a linear cost of lying is common in the literature (e.g., Ozer 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 Appendix 1, table S12

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

$$U_i^{LA}(x,k) = V_i\left(x_i\left(a_{i,k}, a^{-i}\right)\right) - k\left|a_{i,k} - a_k^*\right|;\ \ i \neq j \qquad (3)$$

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* and a *choice experiment*.

In the *choice experiment*, we use a with-in subject design with 20 rounds of play. In the first ten rounds subjects make choices in the Double Dictator Game and in the second ten rounds they make choices in the Bertrand Game. In the Double Dictator Game subjects are randomly placed into AB pairs. Each subject in the AB 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 AB pairs. Each AB pair must simultaneously select 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 is equal to ½ 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.[27] 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

[27] In the original KL experiment all the possible agreements were fixed exogenously by the experimenter in order 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.

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" situation, 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 round for payment from each game.

Our *norms 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 thought the entire experiment consisted of 5 modules.[28] 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[29]:

> *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 × what A*

[28] The *norm elicitation experiment* contained 5 modules in total. However, the first module always elicited the injunctive social norms, is the focus of our analysis and the only data we use for the paper. Modules 2-5 always followed in the same order and are used for various robustness checks not reported in the paper. These three modules collect data on individual beliefs, personal characteristics, and re-measure the injunctive norm after subjects observe others' behavior. In 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 Appendix II.

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

sends) + (6 × what B sends). B's earnings are: 20 - (2 × what B sends) + (6 × 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 ½ 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,[30] subjects were asked to evaluate the "social appropriateness" of a number of the actions available to A[31] 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 with Agreement or No Agreement).[32]

We told subjects that by "socially inappropriate" we meant "consistent with what most people expect individual A ought to do".[33] We also told them that we would pay them not to reveal their own personal opinions but instead to try and match the

[30] 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.

[31] 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, 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.

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

[33] While Krupka and Weber (2013) use 4 categories of appropriateness, we expanded the categories to six (that ranged from very socially inappropriate, socially inappropriate, somewhat socially inappropriate, socially appropriate, somewhat socially appropriate and very socially appropriate).

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 norm elicitation protocol. This scoring rule elicits subjects' median belief[34] about the distribution of others' ratings by matching a subject with another subject and then paying them according to the following payoff function:

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

where π_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.[35] In order 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.[36]

Subjects' earnings for the *norm elicitation experiment* were calculated using the coordination payoffs described above. Their payment was calculated from three different components: (1) from one randomly selected action rating *either* from Module 1 -

[34] 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). Further, 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 pro-social action is unchanged when there is an agreement or not, the degree to which appropriateness ratings vary for actions that deviate from the most pro-social action may vary when an agreement is in place. This, in turn, will change the median rating.

[35] The formal proof of how this payoff function elicits a rater's guess about the median response can be found in Lambert and Shoham's 2009 paper. For our subjects, this payoff function pays them \$15 if they match the other person's ratings exactly. For each category by which they differ from their matched counter-part, subjects lose \$4. Thus, if they are off by 1 category in either direction, then they are paid \$11 and so on. If they are off by five categories (the most they can be wrong) then they pay the experimenter \$5 (ie, in the worst case they would lose their show-up fee). The intuition is that the symmetry of the penalty is sufficient to eliminate any bias in guesses since the rater has an equal incentive neither to be above nor below the median rating. Second, by making the penalty proportional to the difference in one's own rating and that of the other rater, we properly incent guessing about the median. Taken together, the symmetry of the penalty and an increase in penalty that is proportional to the degree of error, make this a proper scoring rule for eliciting a subject's guess about the median response. Finally, we choose a relatively large penalty for miss-coordination (\$4 per category difference) in order to reduce the potential for bias coming from risk aversion, where individuals could bias their ratings towards the middle rating in order to reduce the variance in their coordination payoffs.

[36] 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.

"Injunctive Norm, initial" - or from Module 4 - "Injunctive Norm, after" -, (2) their payoffs from the guesses about behavior in Module 2 and (3) their payoff from *either* the 'Advice Game' or the 'Helping Game' in Module 5 (randomly selected). 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 356 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 one hour and fifteen minutes. The average payoff for each subject was $29.72. In the *choice experiment*, there were a total of 62 subjects in 4 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 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 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 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 DDG).

Figure 1 about here.

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 to 9 ($p < 0.01$ for all). Second, sending the entire endowment of ten tokens is seen as more appropriate in the Agreement condition than in the No Agreement condition ($p < 0.01$). 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 ten tokens is not significantly higher than sending all nine tokens (signed-rank test: $p = 0.52$). By contrast, there is a very large difference in the Agreement condition where taking 'action 9' is rated as being roughly neutral (neither appropriate nor inappropriate) but taking 'action 10' is rated "very appropriate" (signed-rank test: $p < 0.01$).

Figure 2 about here.

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.[37] Hypothesis 2A is fully supported for the

[37] 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

Bertrand Game: taking action 100 is significantly more appropriate than any other action in the Agreement condition ($p < 0.01$ 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: $p = 0.57$), while in the Agreement condition choosing 'action 100' is significantly more appropriate than any other action ($p < 0.01$ for all comparisons).

Regression analysis reported in Table 1 supports these results[38]. 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 and 2, 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 ($b=0.078$, $p < 0.01$).

Table 1 about here.

---

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 actions 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.

[38] In all specifications we cluster the standard errors at the subject level.

However this specification is not flexible enough to capture the non-monotonicity and the sharp discontinuities in the Bertrand Game very well (c.f. Figure 2). 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 (hypothesis 2) 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 and $b=3.423$, $p < 0.01$ in the Bertrand Game) between the highest and next highest actions 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 (c.f. 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 and in subjects' first 10 rounds they 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. We find results that are in line with Kessler and Leider's 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.[39]

Table 2 about here.

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[40]. We see that the having an agreement in place significantly increase the chosen action in both games (β=2.686, $p<0.01$ in DDG; β=30.620, $p<0.01$ in BG). Columns (2) and (5) show that having an agreement significantly increases subjects' first order beliefs (β=3.301, $p<0.01$ in DDG; β=33.382, $p<0.01$ in BG) and second order beliefs (β=1.651, $p<0.01$ in DDG; β=12.511, $p<0.01$ in BG). The significant increase in both actions and first order beliefs is consistent with KL's 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 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 concern for norms ought to outperform models that do not.

[39] Kessler and Leider find that agreements increase the average action by 42% for the Double Dictator Game and by 44% for the Bertrand Game.

[40] In all specifications we cluster the standard errors at the subject levels.

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:

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

Our first specification assumes that utility only depends on own 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 beliefs about the other player's action a^{-i}. Thus, we estimate the weight, β, that individuals place on the money they receive from a particular choice as follows:

$$u_i(x, a_k) = \beta_1 x(a_{i,k}, a^{-i}) \quad [\textit{Selfish model}]$$

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 *choice experiment*.[41] We use a conditional logit regression (McFadden 1974)[42], 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 provides an estimate of the weight on social appropriates in equation (1), or γ.[43]

$$u_i(x, a_k) = \beta_1 x(a_{i,k}, a^{-i}) + \gamma N(a_k) \quad [\textit{Norms model}]$$

Table 3 reports the estimation results for the Double Dictator Game and the Bertrand Game. Because the average norm ratings are a measured quantity which may

[41] 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.

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

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

have sampling error, we use bootstrapped standard errors for the models containing the norm ratings.[44]

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) (β=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 non-zero 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) (β=0.254; p<0.01). For the Bertrand Game the payoff characteristic is positive and significant (column (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, which penalizes models for the number of parameters).[45]

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.[46] 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

[44] 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 re-estimate the choice model based on the sampled norm function. The distribution of the coefficients across replications generates the standard errors.

[45] 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 ($p < 0.01$).

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

the agreed upon action (‘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 3a-3d predict the behavior data from the coefficients on the ***Selfish*** and ***Norms*** models in Table 3. In the Double Dictator Game the ***Selfish*** model predicts the same distribution of actions for both the Agreement and No Agreement case (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 appears to provide not just 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 Alternate 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 while 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 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 in section 2.2). 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.

$$u_i(x, a_k) = \beta_1 x\left(a_{i,k}, a^{-i}\right) - \beta_2 \,\mathrm{GA} \qquad [GA\ model]$$

To test a model of lying aversion we the cost of lying increases linearly in the difference between one's action and the promised action (denoted by 'LA', see equation 3 in section 2.3) We estimate the lying aversion model as follows:

$$u_i(x, a_k) = \beta_1 x\left(a_{i,k}, a^{-i}\right) - \beta_4 \,\mathrm{LA} \qquad [LA\ model]$$

Finally, in the most general specifications, we assume that utility depends on own payoff, guilt or lying aversion and social norms.

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

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

In our data the Norms component can be separately identified from both the Guilt Aversion and Lying Aversion components. The values of N() are set based on the responses of the subjects in the *norms elicitation* experiment, who do not play the games, and N() 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() by construction: it is defined to be zero in the no promise case, and is defined to depend linearly on the action chosen in the promise case.

Table 4 reports for each game the results of the ***Guilt Aversion*** (columns (1) and (5)) and ***Lying Aversion*** models (columns (3) and (7)), 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 Bayesian Information Criterion (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=2759.37, *GA* BIC=2573.64, p = 0.213). The *LA* model has the worst fit of the three, and both *Norms* and *GA* are significant improvements (*LA* BIC=2927.72, $p < 0.01$ 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= 4995.71, *GA* BIC=5222.60, $p < 0.01$). As with the DDG, the *Lying Aversion* model is the worst overall fit, and is significantly worse than both *Norms* and *GA* (*LA* BIC = 5511.49, $p < 0.01$ 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), 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 and Norms* model for the Bertrand Game, the coefficient on the payoff is β=0.485, on GA is β=-0.024 while the coefficient on γ=0.959. Second, adding social norms to either ***GA*** or the ***LA*** model leads to a significant improvement in the BIC and, in the case of the Bertrand Game, substantially reduces the size of the coefficient on GA (from -0.063 to -0.024) and on LA (from -0.021 to +0.013). Using the BIC and comparing across all models in Table 4 and Table 3 we see that a combined model with *GA and Norms* is most preferred (it has the lowest BIC=2446.24 among all models tested). For both games, the *GA and 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 4a-4d 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 don't capture the small uptick at 'action 5' in the No Agreement treatment while the ***Norms*** model does. We see a similar short coming in the Bertrand Game. While 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 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 alternate 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. Further, the ***Norms*** model seems to also do a better job of predicting behavior when no agreement has been reached.

In aggregate it looks like ***GA*** 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 we look across Agreement *and* No Agreement conditions, ***GA*** 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 which 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.[47] Hence, models that

[47] 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 S12 of Appendix I, we estimate Lying Aversion with a fixed cost, a linear cost or 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 vs 645.43; Vuong test $p = 0.152$), while in the Bertrand Game the fixed cost of lying is a much better fit (BIC = 1805.20 vs 2300.39 and 2369.94; $p < 0.01$ for both). (cf. Vuong 1989)

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 your 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 each model relies upon. 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 in their 2006 paper), 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 upon 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, 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 doesn't rely on measured data on norms. And lastly, the elicited social norms can guide researchers in making apriori 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. Further, we show that the social norms we elicit capture key moments of the choice distribution compared to 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.

References:

Andreoni, J. and J. Miller. 2003. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, **70**(2): 737-53.

Benabou, R. and J. Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, **96**(5): 1652-1678.

Battigalli, P. and Dufwenberg, M. 2007. "Guilt in games." *American Economic Review, Papers and Proceedings*, **97**: 170-176.

Bettenhausen, K. and J. Murnighan. 1991. "The Development of an Intragroup Norm and the Effects of Interpersonal and Structural Challenges." *Administrative Science Quarterly*, **36**: 20-35.

Bicchieri, C. 2006. The Grammar of Society: the Nature and Dynamics of Social Norms. Cambridge University Press.

Bicchieri, C. and E. Xiao. 2009. "Do The Right Thing: But Only If Others Do So." *Journal of Behavioral Decision Making*, **22**(2): 191-208.

Burks, S. and E. Krupka. 2012. "Behavioral Economic Field Experiments Can Identify Normative Alignments and Misalignments within a Corporate Hierarchy: Evidence from the Financial Services Industry". *Management Science*, **58**(1):203-217.

Camerer, C. and E. Fehr. 2004. "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists." Foundations of Human Sociality -- Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Ed. J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis. Oxford, UK: Oxford University Press.

Cialdini, R., R. Reno and C. Kallgren. 1990. "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places." *Journal of Personality and Social Psychology*, **58**(6): 1015-26.

Charness, G. and M. Dufwenberg. 2006. "Promises and Partnerships." *Econometrica*, **74**(6): 1579-1601.

Conroy, S. and T. Emerson. 2006. "Changing Ethical Attitudes: The Case of the Enron and ImClone Scandals." *Social Science Quarterly*, **87**(2):395–410.

Chen, Y., N. Kartik and J. Sobel. 2008. "Selecting Cheap-Talk Equilibria." *Econometrica*, **76**(1): 117-136.

Dawes, R.M. 1980. "Social Dilemmas." *Annual Review of Psychology*. **31**: 169-193.

Dawes, R.M., J. McTavish and H. Shaklee. 1977. "Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation." *Journal of Personality and Social Psychology*, **35**: 1-11.

Dawes, R.M., J. Orbell, and A.J. van de Kragt. 1988. "Not me or thee but we: the importance of group identity in eliciting cooperation in dilemma situations." *Acta Psychologica*, **68** (1): 83–97.

Deutsch, M. and H. Gerard. 1955. "A Study of Normative And Informational Social Influences Upon Individual Judgment." *The Journal of Abnormal and Social Psychology*, **51**(3): 629-36.

Dufwenberg, M., & Gneezy, U. 2000. "Price competition and market concentration: an experimental study." *International Journal of Industrial Organization*, ***18***(1), 7-22.

Dufwenberg, Martin, Uri Gneezy, Jacob K. Goeree, and Rosemarie Nagel. 2007. "Price floors and competition." *Economic Theory* **33**(1): 211-224.

Dufwenberg, M. and G. Kirchsteiger. 2000. "Reciprocity and Wage Undercutting." *European Economic Review*, **44**(4-6):1069-1078.

Dufwenberg, M., M. Servatka and R. Vadovic. 2011. "ABC on Deals." Unpublished manuscript.

Dur, R., A. Non and H. Roelfsema. 2010. "Reciprocity and Incentive Pay in the Workplace." *Journal of Economic Psychology*, **31**(4):676-686.

Ellingsen, T. and M. Johannesson. 2004. "Promises, Threats and Fairness." *The Economic Journal*, **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.

Englmaier, F. and S. Leider. 2012. "Contractual and Organizational Structure with Reciprocal Agents." *American Economics Journal: Microeconomics*, forthcoming.

Erat, S. and U. Gneezy. 2012. "White Lies." *Management Science*, **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 and Perspectives in Management*, **2**:135-147.

Falk, A. and M. Kosfeld. 2006. "Distrust - The Hidden Cost of Control". *The American Economic Review*, **96**(5):1611-1630.

Farrell, J., and M. Rabin. 1996. "Cheap talk." *The Journal of Economic Perspectives* **10**(3): 103-118.

Fehr, E. and S. Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, **14**:159-81.

Fehr, E., O. Hart and C. Zehnder. 2009. "Contracts, Reference Points, and Competition - Behavioral Consequences of the Fundamental Transformation." *Journal of the European Economic Association*, **7**:561-572.

Fehr, E., O. Hart and C. Zehnder. 2011. "Contracts as Reference Points-Experimental Evidence." *American Economic Review*, **101**(2): 493-525.

Fehr, E. and A. Falk. 1999. "Wage Rigidity in a Competitive Incomplete Contract Market." *The Journal of Political Economy*, **107**(1): 106-134.

Fehr, E. and K. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics*, **114**(3):817-68.

Fisher, P. and S. Huddart. 2008."Optimal Contracting With Endogenous Social Norms." *American Economic Review*, **98**(4):1459-75.

Gächter, S., D. Nosenzo and M. Sefton. 2013. "Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?" *Journal of the European Economic Association*, **11**(3): 548-573.

Gino, F., D.A. Moore and M.H. Bazerman. 2008. "No Harm, No Foul: The Outcome Bias in Ethical Judgement." Harvard Business School Working Paper, No. 08-080.

Hart, O. and J. Moore. 2008. “Contracts as Reference Points.” *Quarterly Journal of Economics*, **123**(1):1-48.

Gneezy, U. 2005. "Deception: The role of consequences." *The American Economic Review* 95.1 (2005): 384-394.

Hart, O., and J. Moore. 2008. "Contracts as reference points." *The Quarterly Journal of Economics* 123.1 (2008): 1-48.

Hoffman, S.D., & Duncan, G.J. 1988. "Multinomial and conditional logit discrete-choice models in demography." *Demography*, **25**(3): 415-427.

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

Kandori, M. 1992. “Social Norms and Community Enforcement.” *Review of Economic Studies*, **59**:62-80.

Kessler J. and S. Leider, 2012 “Norms and Contracting.” *Management Science*, **58** (1):62-77.

Krupka, E. and R. Weber. 2013. “Identifying norms using coordination games: Why does dictator game sharing vary?” *Journal of the European Economic Association*, **11**(3): 495-524.

Krupka, E., R.Weber and R. Croson. (unpublished manuscript). “When in Rome: Identifying Social Norms as a Group Phenomenon.”

Krupka, E. and R.Weber. 2009. "The Focusing and Informational Effects of Norms on Pro-Social Behavior." *Journal of Economic Psychology*, **30**:307-20.

Lambert, N. and Y. Shoham. 2009. "Eliciting Truthful Answers to Multiple-choice Questions." *EC '09 Proceedings of the tenth ACM conference on Electronic Commerce.*

Leider, S., M. Möbius, T. Rosenblat and Q. Do. 2009. "Directed Altruism and Enforced Reciprocity in Social Networks." *Quarterly Journal of Economics*, **124**(4): 1815-1851.

Loomis, J.L. 1959. "Communication, the development of trust, and co-operative behavior." *Human Relations*, **12**: 305–l5.

López-Pérez, R. 2008. "Aversion to norm-breaking: A model." *Games and Economic Behavior* **64**(1):237-267.

Lundquist, T., T. Ellingsen, E. Gribbe, and M. Johannesson. 2009. “The aversion to lying.” *Journal of Economic Behavior and Organizations*, **70**(1–2):81–92.

Malhotra, D. and J. Murnighan. 2002. "The Effects of Contracts on Interpersonal Trust." *Administrative Science Quarterly*, **47**(3):534-559.

McFadden, D. 1974. “Conditional logit analysis of qualitative choice behavior”. In Frontiers in Econometrics. Ed. P. Zarembka. New York, NY: Academic Press.

McKinney, J., and C. Moore. 2008. “International Bribery: Does a Written Code of Ethics Make a Difference in Perceptions of Business Professionals.” *Journal of Business Ethics*, **79**:103-111.

Mehta, J., C. Starmer and R. Sugden. 1994. "The Nature of Salience: An Experimental Investigation of Pure Coordination Games." *American Economic Review*, **84**(3):658-73.

Miettinen, T. 2013. “Contracts and Promises - An Approach to Pre-play Agreements.” *Games and Economic Behavior*. **80**: 68-84.

Orbell, J.M., A.J. van de Kragt, A.J. and R.M. Dawes, R.M. 1991. "Covenants without the sword: the role of promises in social dilemma situations." in Social Norms and Economic Institutions. Ed. K. Koford and J. Miller. Ann Arbor, MI: University of Michigan Press.

Ostrom, E. 2000. "Collective Action and the Evolution of Social Norms." *Journal of Economic Perspectives*, **14**(3):137-58.

Ozer, O., Y. Zheng and K. Chen. 2011. “Trust in Forecast Information Sharing.” *Management Science*, **57**(6):1111-1137.

Oumlil A. and J. Balloun. 2009. "Ethical Decision-Making Differences Between American and Moroccan Managers." *Journal of Business Ethics*, **84**:457–478.

Rigdon, M. 2009. “Trust and Reciprocity in Incentive Contracting.” *Journal of Economic Behavior and Organization*, **70**:93-105.

Sally, D. 1995. “Conversation and cooperation in social dilemmas: Experimental evidence from 1958 to 1992.” *Rationality and Society*, **7**(1):58-92.

Scott, R. 2003. “A Theory of Self-Enforcing Indefinite Agreements.” *Columbia Law Review*, **103**(7):1641-1699.

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

Schwartz, S. 1973. “Normative Explanations for Helping Behavior: A Critique, Proposal and Empirical Test.” *Journal of Experimental Social Psychology*, **9**(4):349-364.

Sliwka, D. 2007. “Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes.” *American Economic Review*, **97**(3):999-1012.

Sugden, R. 1995. "A Theory of Focal Points." *The Economic Journal*, **105**(430):533-50.

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 Non-Nested Hypotheses." *Econometrica*, **57**(2):307-333.

Young, P. 1998. "Social Norms and Economic Welfare." *European Economic Review*, **42**:821-30.

Figure 1: Average appropriateness ratings for the Double Dictator Game with and without agreement (data from Module 1*)*

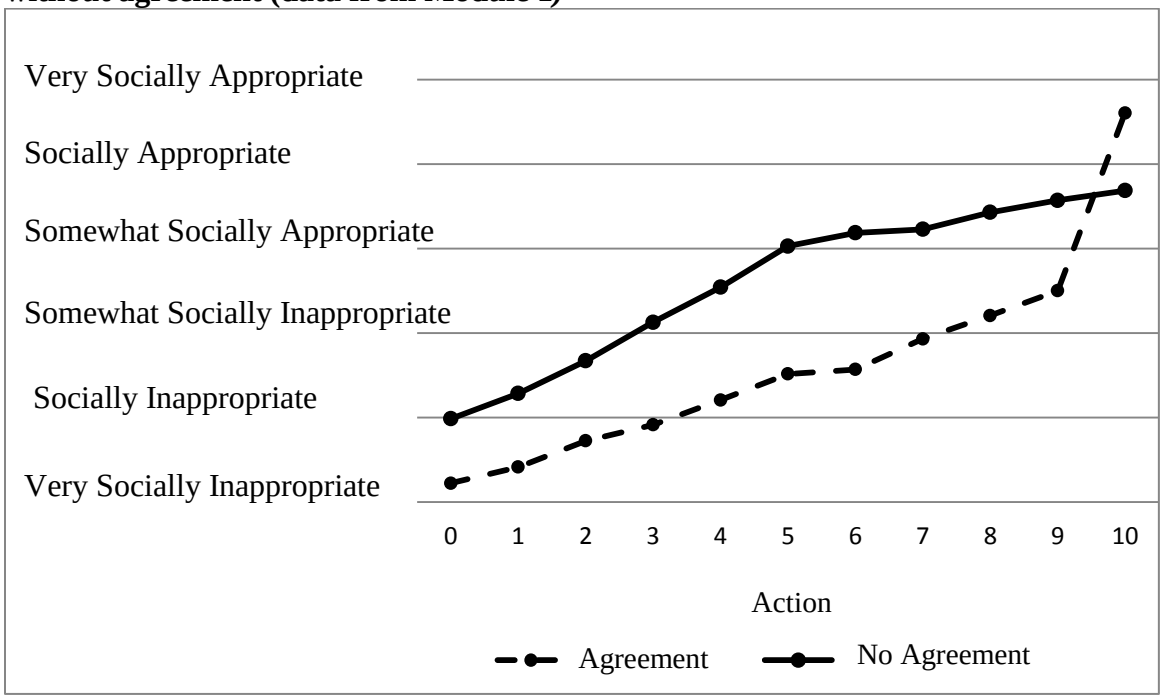

Figure 2: Average appropriateness ratings for the Bertrand Game with and without agreement (data from Module 1*)*

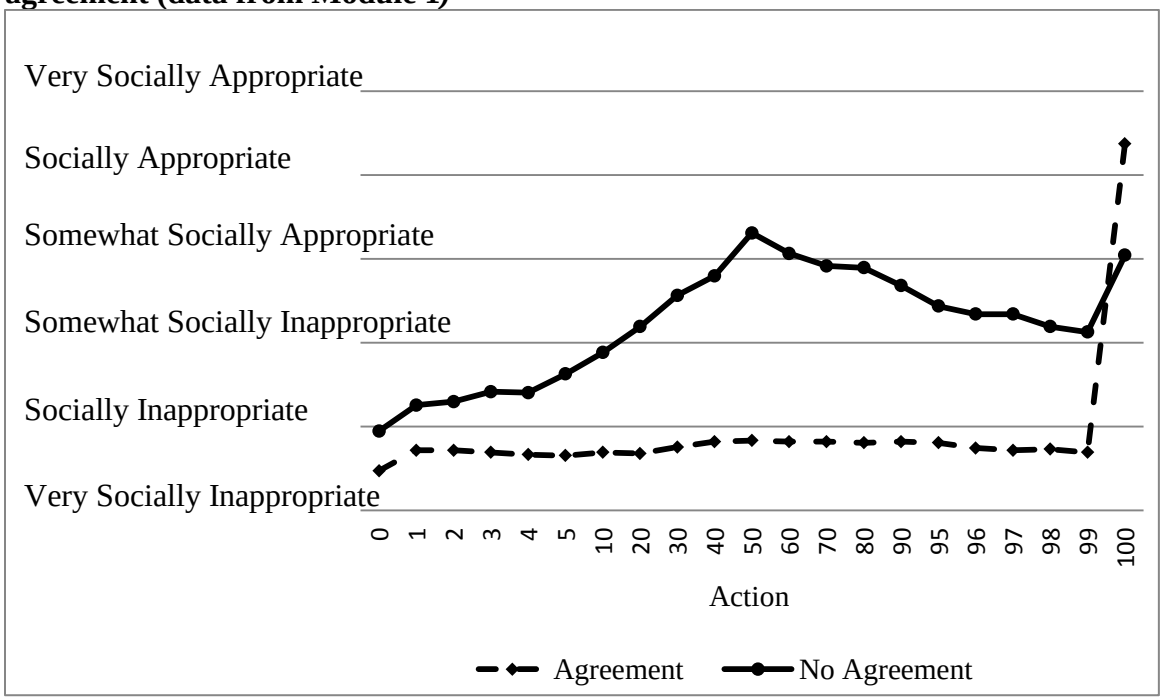

Figure 3a: Distributions and predicted distributions of actions taken in the Double Dictator Game (predictions based on *Selfish* model coefficients in Table 3, model 1)

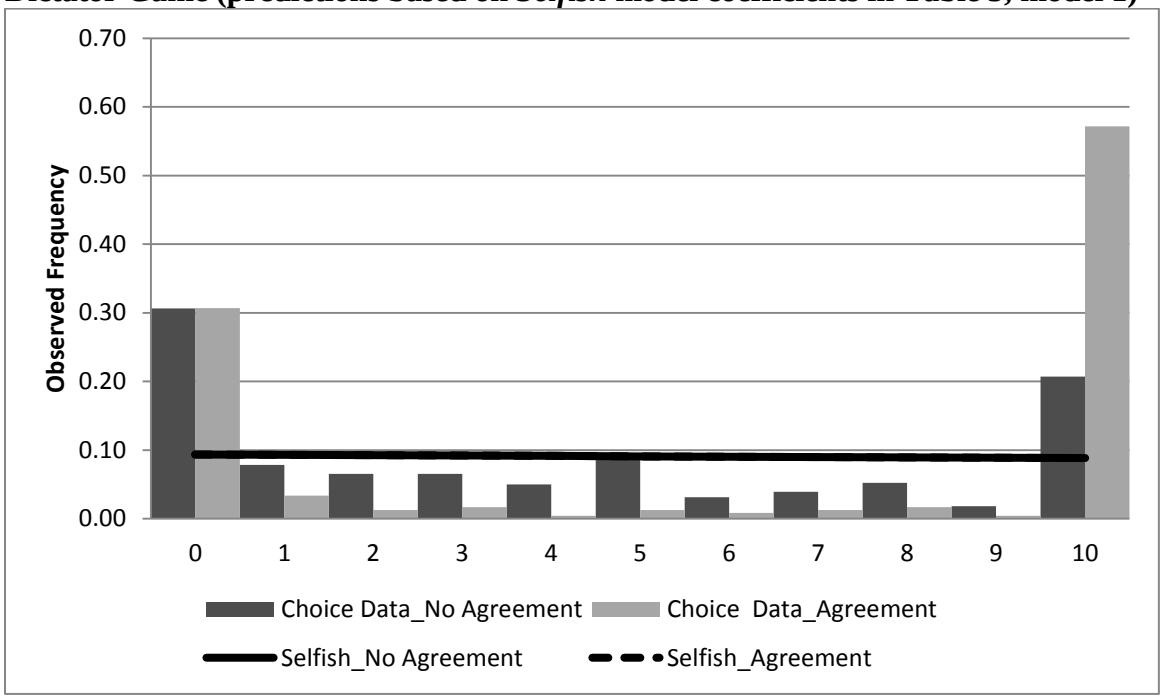

Figure 3b: Distributions and predicted distributions of actions taken in the Double Dictator Game (predictions based on *Norms* model coefficients in Table 3, model 2)

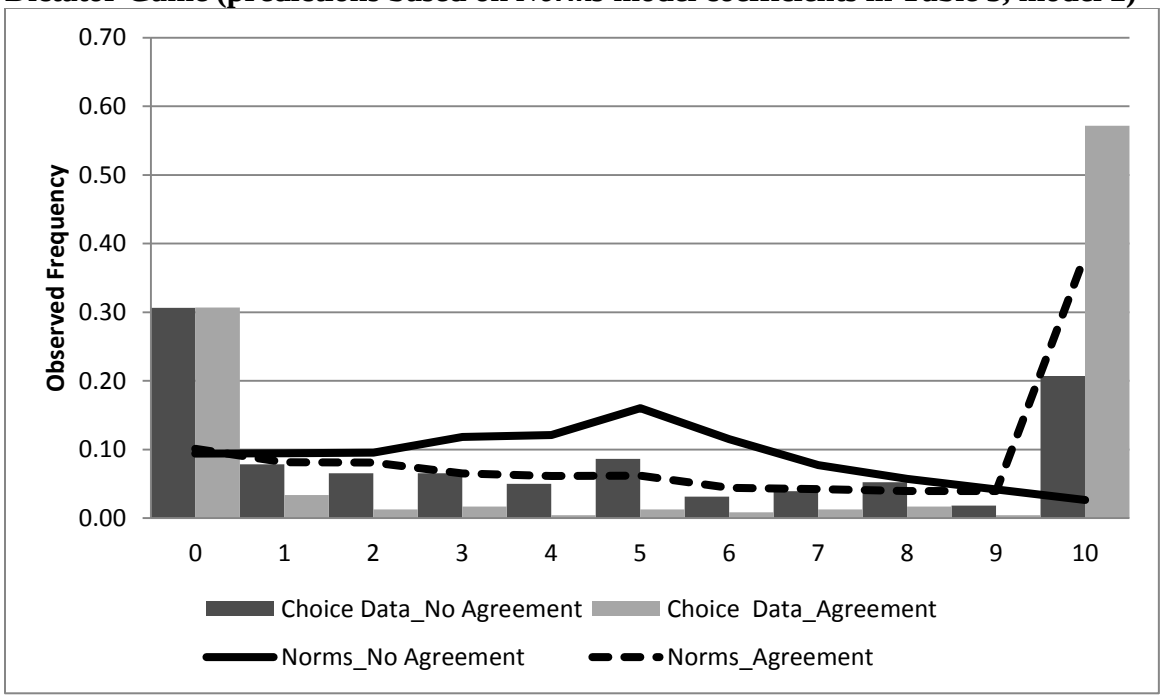

Figure 3c: Distributions and predicted distributions of actions taken in the Bertrand Game (predictions based on *Selfish* model coefficients in Table 3, model 3)

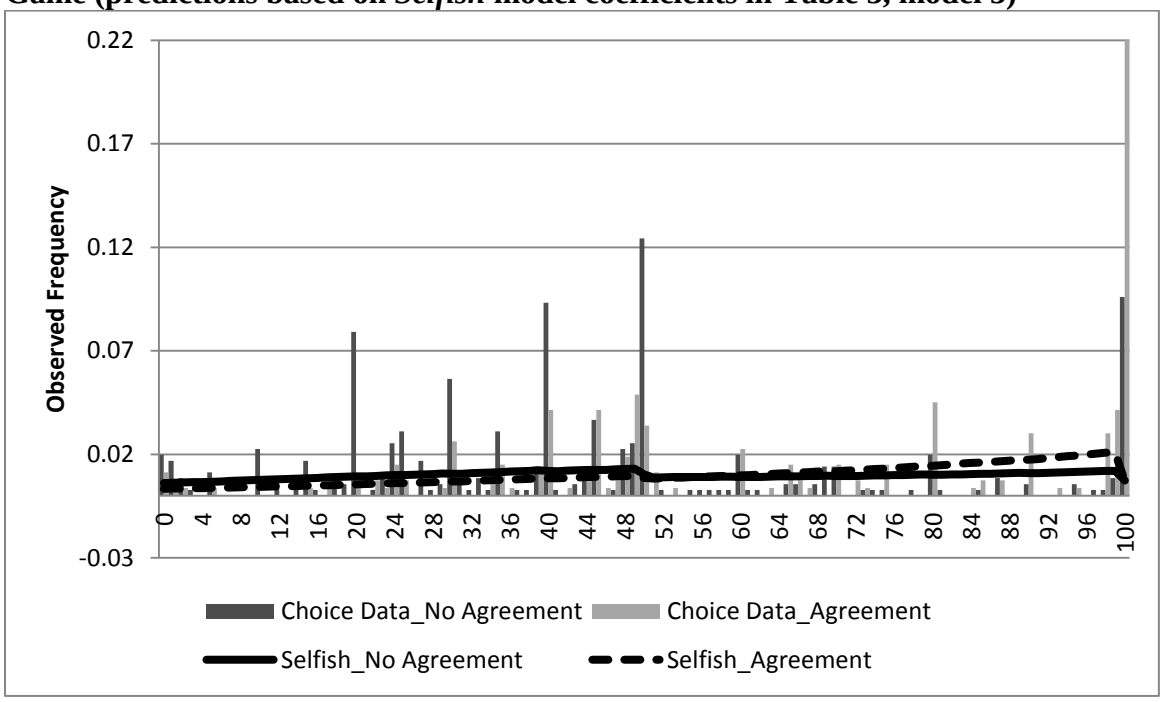

Figure 3d: Distributions and predicted distributions of actions taken in the Bertrand Game (predictions based on *Norms* model coefficients in Table 3, model 4)

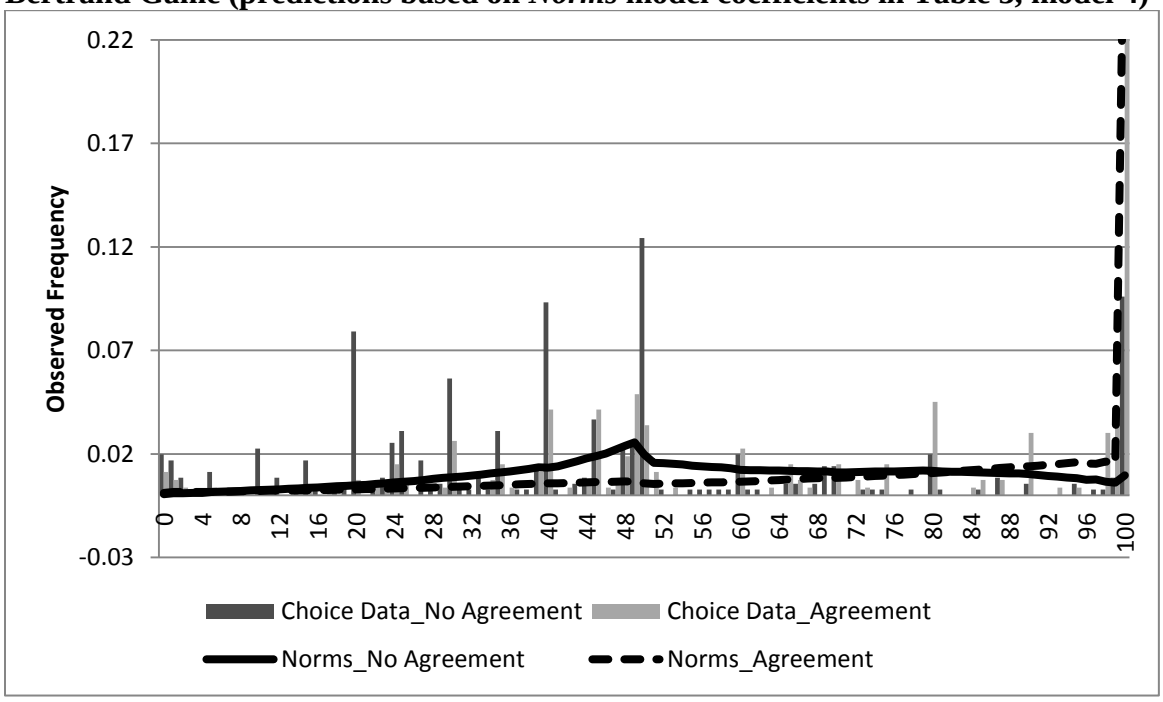

Figure 4a: Distributions and predicted distributions of actions taken in the Double Dictator Game (predictions based on the *Guilt Aversion* model coefficients in Table 4, model 1)

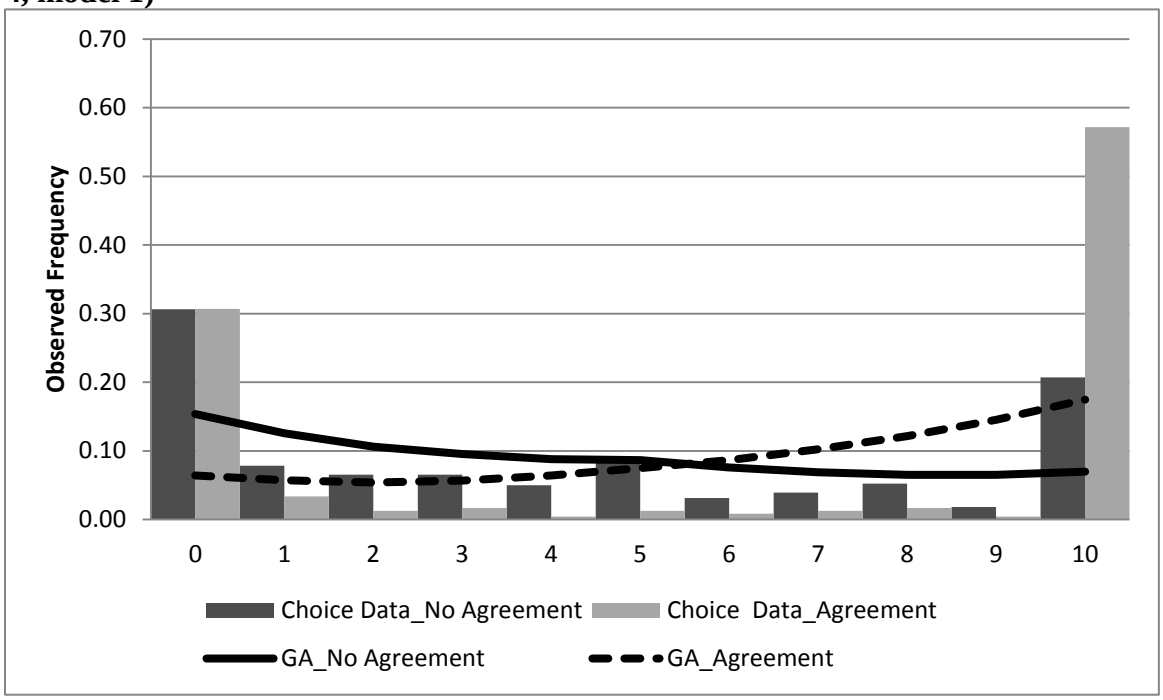

Figure 4b: Distributions and predicted distributions of actions taken in the Double Dictator Game (predictions based on the *Lying Aversion* model coefficients in Table 4, model 3)

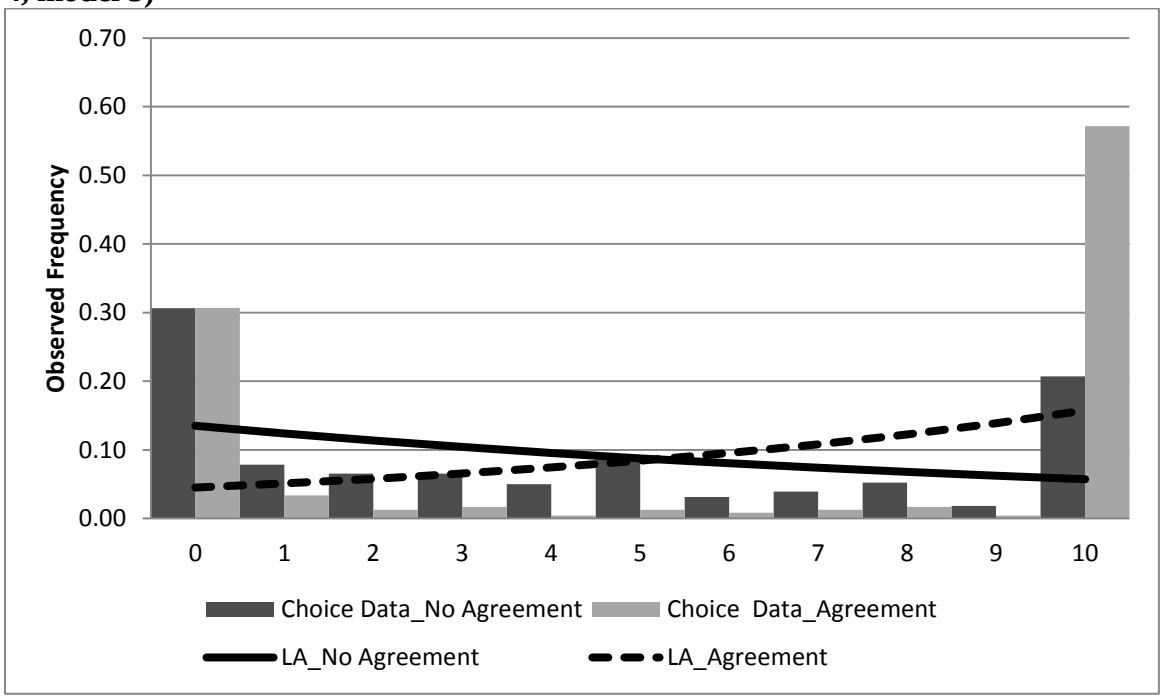

Figure 4c: Distributions and predicted distributions of actions taken in the Bertrand Game (predictions based on *Guilt Aversion* model coefficients in Table 4, model 5)

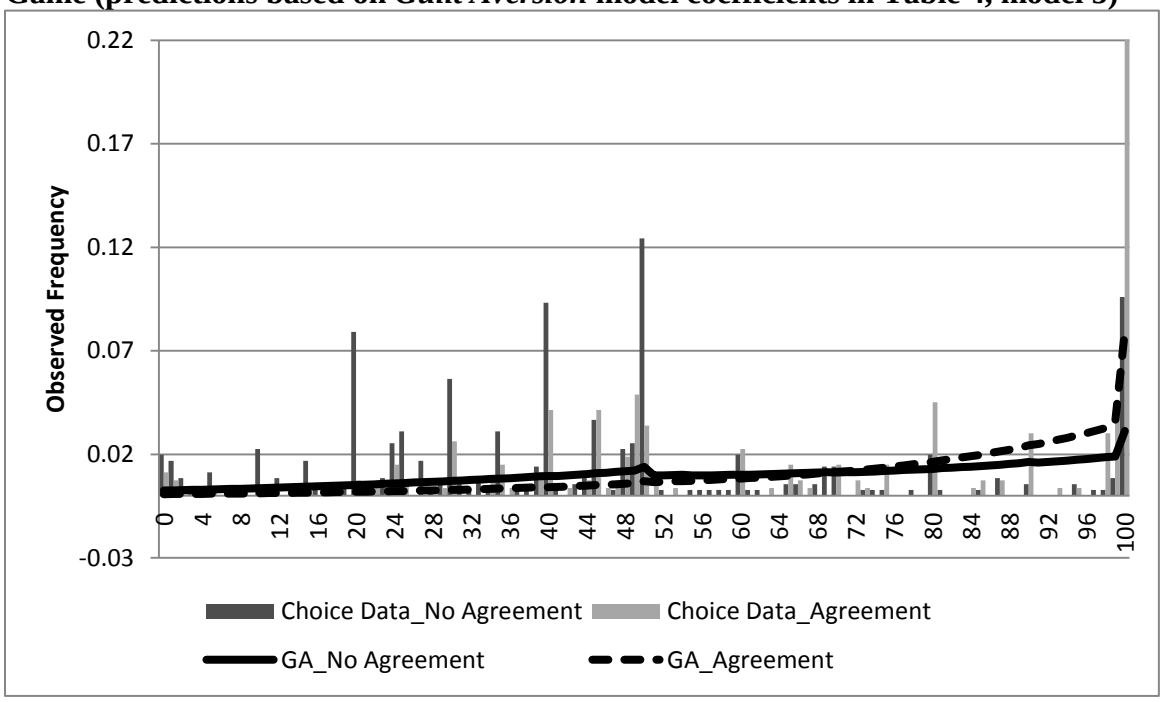

Figure 4d: Distributions and predicted distributions of actions taken in the Bertrand Game (predictions based on *Lying Aversion* model coefficients in Table 4, model 7)

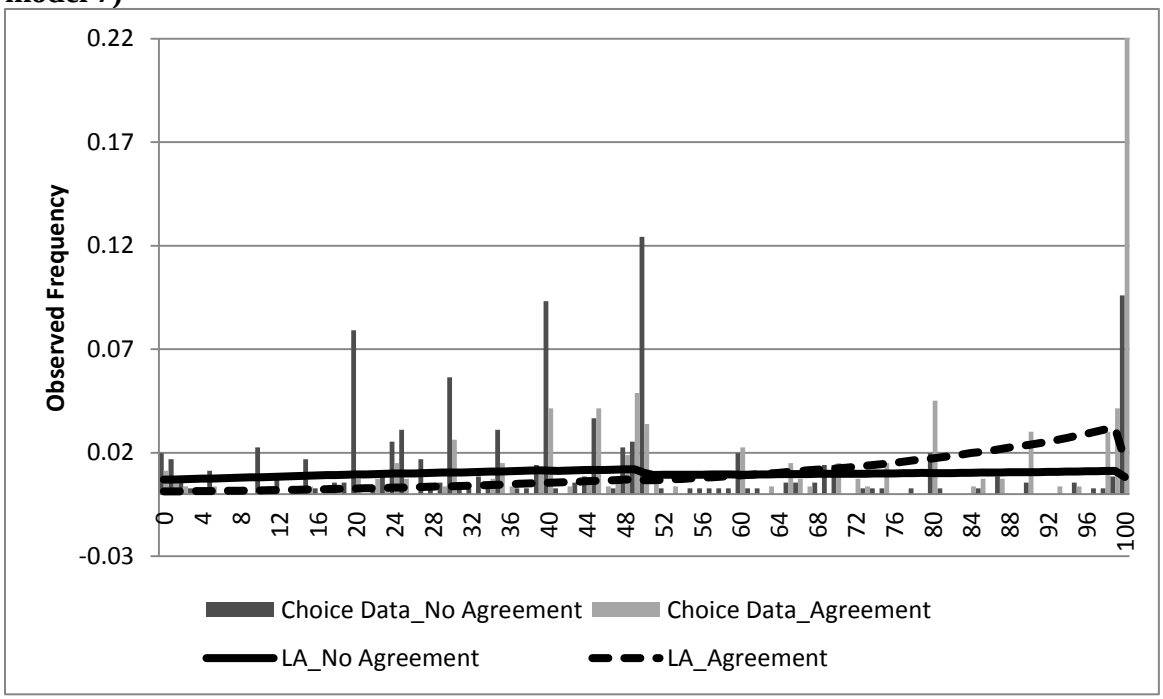

Table 1: OLS regressions on appropriateness ratings for the Double Dictator Game and the Bertrand Game (ratings elicited in Module 1 of the experiment)

	DDG		BG	
VARIABLES	(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	1914	1914	3864	3864
# 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.

Table 2: Statistical tests of the effect of having an agreement on behavior, first and second order beliefs in the Double Dictator and Bertrand Game

	DDG			**BG**		
DV:	*Chosen action*	*FOB*	*SOB*	*Chosen action*	*FOB*	*SOB*
	(1)	(2)	(3)	(4)	(5)	(6)
Have Agreement	2.686***	3.306***	1.652***	30.620***	33.382***	12.512***
	(0.427)	(0.378)	(0.328)	(3.563)	(3.187)	(2.626)
FOB			0.614***			0.630***
			(0.060)			(0.049)
Constant	3.913	5.062	2.160	43.726	54.677	24.413
	(0.355)	(0.298)	(0.423)	(2.658)	(2.945)	(3.460)
Model	*OLS, RE*					
Observations	620	620	620	620	620	620
# of Subjects	62	62	62	62	62	62
R-Squared	0.054	0.146	0.553	0.1878	0.238	0.604

Notes: The dependent variable is the chosen action (Columns 1 and 3), first order belief (Columnds 2 and 4) and second order belief (Columns 3 and 6) in the Double Dictator Game (columns 1 to 3) or in the Bertrand Game (columns 4 to 6); standard errors clustered by subject are reported in parentheses; *** p<0.01, ** p<0.05, * p<0.1.

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	
	Selfish	*Norms*	*Selfish*	*Norms*
VARIABLES	(1)	(2)	(3)	(4)
Action Payoff (β)	0.003	0.254***	0.025***	0.032***
	(0.006)	[0.019]	(0.002)	[0.002]
Norm Rating (γ)		1.449**		1.202***
		[0.263]		[0.0752]
Monetary Value (0.15γ/β)		0.855***		5.634***
		[0.059]		[0.569]
Observations	620	620	620	620
Log Likelihood	-1486.59	-1370.86	-2770.3	-2486.81
Bayesian IC	2982.0	2759.37	5551.66	4995.71

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 reported in parentheses with bootstrapped standard errors in brackets for specifications with norm ratings; *** p<0.01, ** p<0.05, * p<0.1. 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 DG 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, "very socially appropriate"=6. The ratio 0.15γ/β 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 experiments was worth $0.15.

Table 4: Conditional logit estimation of choice determinants for the Double Dictator Game and Bertrand Game using alternate mechanisms

	DDG				BG			
	GA	*GA+Norms*	*LA*	*LA+Norms*	*GA*	*GA+Norms*	*LA*	*LA+Norms*
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Action Payoff (β)	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 (γ)		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 (0.15γ/β)		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	-1277.99	-1209.87	-1455.03	-1370.19	-2600.29	-2469.10	-2744.70	-2479.46
Bayesian IC	2573.64	2446.24	2927.72	2766.86	5222.6	4971.35	5511.49	4992.06

Notes: The dependent variable is the chosen action in the Double Dictator Game (columns 1 to 4) or in the Bertrand Game (columns 5 to 8) ; standard errors are reported in parentheses with bootstrapped standard errors in brackets for specifications with norm ratings; *** p<0.01, ** p<0.05, * p<0.1. 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 DG 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, "very socially appropriate"=6. The ratio 0.15γ/β 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 experiments was worth $0.15.