Participation

Gary Charness & Martin Dufwenberg*

May 10, 2010

ABSTRACT: We show experimentally that whether and how communication achieves beneficial social outcomes in a hidden-information context depends crucially on whether low-talent agents can participate in a Pareto-improving outcome. Communication is effective (and patterns of lies & truth quite systematic) when this is feasible, but otherwise completely ineffective. We examine the data in the light of two potentially relevant behavioral models: cost-of-lying and guilt-from-blame.

KEYWORDS: Adverse selection, hidden information, participation, communication, cost-of-lying, guilt-from-blame, psychological games, trust, cooperation, lies, deception, social preferences, behavioral economics

JEL CODES: A13, B49, C72, C91, D63, D64, J41

ACKNOWLEDGMENTS: We would like to thank three referees for their very helpful comments, Geir Asheim, Pierpaolo Battigalli, Stefano DellaVigna, Tore Ellingsen, Shachar Kariv, Ulrike Malmendier, and Matthew Rabin for useful discussions, and the participants at the Arne Ryde Symposium on Communication in Games and Experiments at Lund University, the IWEBE conference in Lyon, the St. Andrews Applied Microeconomics Workshop, the Amsterdam Behavioral & Experimental Economics Workshop, the ESA conferences in Lyon and Innsbruck, and seminar participants at the University of Gothenburg, Brown University, the University of California at Berkeley, Columbia Business School, London School of Economics, Einaudi Institute for Economics and Finance in Rome, University of Amsterdam, University of Heidelberg, and University of Bonn for more helpful comments. The research was conceived with the support of the Russell Sage Foundation and completed with the support of the National Science Foundation (grants # SES-0617923, # SES-0921929).

* CONTACT: Gary Charness, Department of Economics, University of California at Santa Barbara, charness@econ.ucsb.edu; Martin Dufwenberg, Department of Economics, University of Arizona and Department of Economics, University of Gothenburg, martind@eller.arizona.edu.

Human collaboration has produced much in the world. Research in contract theory (often collaborative efforts!) explores which partnerships form, what contracts are signed, and what the consequences will be. Considerable attention has been given to settings with hidden action (where a party's future choice is not contractible) or hidden information (where a contract cannot be conditioned on a party's private information). When parties act opportunistically, these are hurdles that may preempt fruitful collaboration.[1]

In this paper, we investigate an environment with *hidden information*. Here, while the agent's effort choice is observable and contractible, his production also depends on his ability.[2] A crucial feature is that while the agent knows his ability, the principal does not. Our approach complements that of Gary Charness & Martin Dufwenberg (2006), who consider a hidden-action context. However, the games differ regarding the nature of the trust needed for efficiency to prevail. Under hidden action, a principal must rely on an agent to not act opportunistically but there is no doubt that the agent could deliver in principle. This is different from hidden information, where some agents (with low talent) simply cannot deliver as well as others. Hidden information involves an asymmetry that lacks a counterpart in the hidden-action case.

We consider the interaction of two important issues in our experimental design. The first issue is the extent to which an agent with low-talent can *participate* in an outcome that is a Pareto-improvement for both the principal and the agent. In one environment, there are two possible types of employment available, with more paid for the job requiring high talent; if the low-ability agent chooses the position not requiring high talent, both the agent and the principal

[1] For an entry to the literature, see Patrick Bolton & Mathias Dewatripont (2005). The gloomy outlook can be exemplified with reference to George Akerlof's (1970) classic work on hidden information: The seller of a used car knows its quality while the buyer does not. This creates an obstacle to reaching socially-attractive agreements, and market failure results. The terms hidden action and hidden information are often called, respectively, moral hazard and adverse selection. The "hidden" terminology seems more descriptive and less suggestive of the nature of outcomes.

[2] In this paper, we shall consider the principal to be female and the agent to be male.

are better off than if the principal chooses not to offer him employment. In the second environment, there is no low-skill position available. In both environments, a principal does poorly if matched with a low-talent agent who chooses the position requiring high ability. In both cases, principals must rely on low-talent agents to voluntarily accept less than could be obtained by acting selfishly and choosing the better-paid position, but in the second case low-talent agents who wish to avoid hurting the principal must step aside and decline the contract.

The second issue is whether *communication* can help to ameliorate the hidden-information problem. If agents have selfish preferences, the prediction in both environments is the same: A low-talent agent will choose the high-skill position and receive more income. Since a vast number of papers have shown that many people have social preferences, we would expect that not all low-talent agents make the selfish choice. But it also may well be the case that some aspect of communication will help to promote trust & cooperation. However, given the qualitative difference in the environments, the character and content of the messages sent are likely to also differ from those in the hidden-action environment.

We find that communication can be effective with hidden information, although this depends critically on low-talent agents having the possibility to participate in a Pareto-improving outcome. We proceed to discuss this result in the light of two behavioral models that can potentially explain such an effect and that have received some support in recent experimental research. One such model involves a cost-of-lying,[3] while the other is Pierpaolo Battigalli & Dufwenberg's (2007) model of guilt-from-blame, which has its intellectual home within the framework of psychological game theory (John Geanakoplos, David Pearce & Ennio Stacchetti

[3] Previous theoretical work considering various forms of cost-of-lying includes Tore Ellingsen & Magnus Johannesson (2004), Ying Chen, Navin Kartik & Joel Sobel (2008), Stefano Demichelis & Jörgen Weibull (2008), and Kartik (2009). For some related experimental results see Uri Gneezy (2005), Topi Miettinen (2008), Matthias Sutter (2009), Christoph Vanberg (2008), and Charness & Dufwenberg (forthcoming).

1989; Battigalli & Dufwenberg 2009). We present formal predictions for each model in our environments and discuss the extent that the models can capture the observed patterns of behavior.

Besides shedding light on the empirical relevance of some behavioral theory, we note that our results will reveal some seemingly rather stable patterns regarding how language is used strategically, and how words correlate with opportunism and trustworthiness. There may be ‘lessons-for-life’ to take away for both confidence tricksters who wish to improve their deceptive skills and for lie-detectors who wish to build better traps.

The remainder of the paper is organized as follows. Our hidden-information games are presented in section I. The experiment design is described in section II, and the experimental results are presented in section III. The two behavioral models are presented in section IV, and section V offers concluding remarks.

I. HIDDEN-INFORMATION GAMES

In this section we describe the games that we use in our treatments. The game (form) in Figure 1 models our benchmark scenario (which for reasons explained further below we shall call our (5,7)-game). A principal (player A) considers employing an agent (B) to form a partnership in which a project is carried out. If A passes on this option – an outcome corresponding to A's choice *Out* – then no contract is signed, no project is carried out, and the parties get their outside-option payoffs of 5 (dollars) each. The project is carried out if A chooses *In*, in which case A pays a fixed wage to B and then acts as residual claimant.

<<<Figure 1 about here>>>

Note that there is hidden information, since only the agent knows his own productivity (or talent). If B has low talent – which happens with probability 2/3 as indicated by the initial chance move – then he is only capable of performing a simple task such that if A pays B an appropriate low wage they split the gain and get 7 each. On the other hand, if B has high talent he could take on a more difficult and (in expectation) profitable task at which a low-talent agent would fail. Since only B knows his talent, only he can tell what is the best mutually beneficial contract, and the game in Figure 1 incorporates an opportunity for him to select it: choice *Don't* represents the low-wage simple task and choice *Roll* the high-wage difficult task.[4]

If a high-talent B chooses *Roll* then the outcome is potentially rewarding but risky: with probability 1/6 the project still fails (as it would for sure if low-talent B chose *Roll*). The chance move following path (*High*, *In*, *Roll*) captures this. The dotted line connecting A's payoffs of $0, following paths (*Low*, *In*, *Roll*) and (*High*, *In*, *Roll*, *Failure*) indicates an information set for A across end nodes.[5] This reflects how A is never told how her payoff of $0 came about.[6]

Why have we included this chance move that determines the project's success, rather than just replace it with its expected outcome (10, 10)? The answer is that this provides a *conceptual* justification for our claim that the game incorporates hidden information. This is a circumstance where a contract couldn't even in principle be conditioned on a party's private information; here this applies to the agent's talent. A typical justification for such a contractual limit, often stressed by contract theorists, is that the agent's type is not observable to the

[4] The labeling of players and strategies in Figure 1, which may appear somewhat artificial in light of the principal-agent story, anticipates the upcoming wording of our experimental instructions as described below.

[5] Information sets across endnodes are typically not given in standard game theory as they would have no bearing on equilibrium play. However, in psychological games such information can critically affect play (as our discussion in section IV will show). See Battigalli & Dufwenberg (2009, section IVb) for more discussion of this point.

[6] In principle, there should also be dotted lines connecting A's payoffs of $5 as well as A's payoffs of $7, but these are omitted for expositional clarity.

principal, or at least not verifiable in court. The chance move justifies a story where a low-type agent could falsely claim that he was in fact a high-type agent but that he had bad luck. Because of the chance move, it cannot be proven in court that he lied.

If the players are selfish and risk-neutral, the (5,7)-game of Figure 1 has a unique sequential equilibrium (henceforth, SE) as defined by Kreps & Wilson (1982): two steps of a backward induction argument yields that B chooses *Roll* independently of his talent, and A's best response is *Out* (this gives A a payoff of 5 whereas *In* would give A an expected payoff of $\frac{1}{3}(\frac{5}{6}\times 12+\frac{1}{6}\times 0)+\frac{2}{3}\times 0=\frac{10}{3}$). The players earn 5 each independently of B's talent. The outcome is inefficient, since A, a low-talent B, and a high-talent B would each receive more (in expectation) if A chose *In* and low-talent B chose *Don't* while high-talent B chose *Roll.*[7] This illustrates how hidden information may undermine efficient contracting.

We also consider a version of the game with an added communication opportunity; B can send a message to A just after chance has determined B's talent and just before A chooses *In* or *Out*. With standard preferences the prediction does not change relative to the no-communication game; words can't change the fact that B gets a higher dollar payoff from *Roll* than from *Don't*, and given this A chooses *Out*.

How should one react to these predictions? One possibility is to take the indicated problem at face value, and examine whether *other* contractual arrangements help overcome the problems. This sort of approach is typical in contract theory; the optimal choice of contract when a partnership is influenced by hidden information is a major issue, and the assumption that the principal and the agent are selfish is typically maintained. We do *not* follow that approach,

[7] A would get 8 = (1/3) · [(5/6) · 12 + (1/6) · 0] + (2/3) · 7; low-talent B would get 7; high-talent B would get 10.

as we are skeptical of this traditional premise, particularly when communication is involved. We stick with the game of Figure 1 with an open mind to whether or not the situation is problematic.

We now move to the important issue of participation. The game in Figure 1 allows a way for each of the two types of the agent to have mutually profitable (Pareto-improving) dealings with the principal. A high-talent agent who chooses *Roll* moves himself and the principal from a payoff of 5 to a payoff of 10 (in expected terms), while a low-talent agent who chooses *Don't* moves the payoff from 5 to 7. Everyone gains. But note how the gains-from-trade are asymmetric as regards different types of agents. One may imagine a more extreme form of such asymmetry, where the low-talent agent is simply incapable of participating in making net additions to partnership profit. Perhaps they lack any helpful trait, or perhaps government taxation is so high that all gains from trade get wasted, or perhaps there is only one position to fill and many available agents so that the principal is only interested in hiring a high-talent agent. The game in Figure 2 incorporates such a change to the setting:

<<<Figure 2 about here>>>

We call the game in Figure 1 our (5,7)-game because 5 is the value of the outside option (*Out*) and 7 is the value of the low-wage simple-task outcome (path via *Don't*). Accordingly, the game in Figure 2 is our (5,5)-game. Parametrically, the change between games looks small: four 7's are replaced by four 5's. The interpretation of the *Don't* choice changes too, to reflect a 'step-aside' move. The prediction for selfish players does not change though: A chooses *Out*, and B chooses *Roll* independently of talent. And again, adding communication (in the same way as described for the (5,7)-game) would not change this dismal prediction.

We find it intuitive that when behavioral concerns are considered it will somehow be easier to foster trust & cooperation in the (5,7)-game than in the (5,5)-game – asking low-talent agents to accept a lesser gain seems easier than asking them to step aside. We explore this. It turns out that for theory-testing purposes we need a third game, a variant of the step-aside scenario called the (7,7)-game. We defer a discussion of the rationale and here just present it:

<<<Figure 3 about here>>>

II. EXPERIMENTAL DESIGN

In line with the presentation in section I, we have a 3×2 design. The first treatment variable concerns whether subjects played the (5,7)-game, the (5,5)-game, or the (7,7)-game. In each case we have one-shot interaction, to rule out any reputation or repeat-game effects. The second treatment variable concerns whether or not communication from B to A was allowed. We provided each potential sender with a blank piece of paper on which he could write any (anonymous) message instead of restricting the message space.

Participants were recruited at UCSB by sending out an e-mail message to the campus community. We conducted 18 sessions, three for each of our six treatments. Sessions were conducted in a large classroom that was divided into two sides by a center aisle, and people were seated at spaced intervals. The number of participants in a session ranged from 20 to 36, for a total of 510 people; each person could only participate in one of these sessions. Average earnings were about $14, including a $5 show-up fee; each session was one hour in duration.

In each session, participants were referred to as “A” or “B”. A coin was tossed to determine which side of the room was A and which side was B. Index cards with identification numbers were drawn from an opaque bag, and participants were informed that these numbers would be used to determine pairings (one A with one B) and to track decisions. Each B first learned his type, which was determined by his private draw. If his identification number was evenly divisible by three, B had high talent; otherwise he had low talent. Sample instructions are given in the online data file on the *American Economic Review* website.

In all of our treatments, we presented a Table to each of the participants, indicating the outcome for every combination of choices and die rolls. After answering questions, the experimenter chose individuals at random to state the outcome for each possible case, starting the session when it seemed clear that everyone understood the rules. In the message treatments, B had an option to send a free-form message to A prior to A's decision. B could also decline to send a message by circling the letter B at the top of the otherwise-blank sheet. Then A chose *In* or *Out*. Finally, B learned A's choice and, if A had chosen *In*, chose *Roll* or *Don't*.

Table 1 shows the experimental presentation of the (5,7)-game; this is identical for the (5,5)- and (7,7)-games, except that each "7" in the fourth and seventh rows is replaced with "5" and "7", respectively, and each "5" in the first row is replaced by "7" in the (7,7)-game.

Table 1: Payoff Outcomes in (5,7) Game

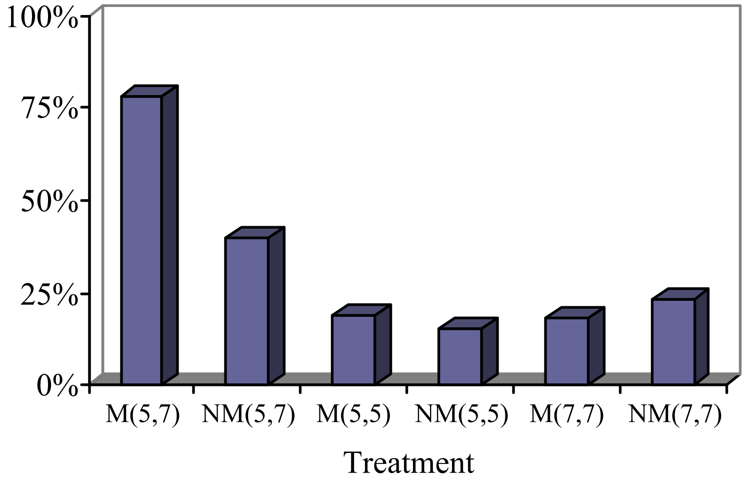
	A receives	**B receives**
A chooses OUT	**$5**	**$5**
A chooses IN and:		
B is LOW type and chooses DON'T ROLL	**$7**	**$7**
B is LOW type and chooses ROLL	**$0**	**$10**
B is HIGH type and chooses DON'T ROLL	**$7**	**$7**
B is HIGH type, chooses ROLL, die = 1	**$0**	**$10**
B is HIGH type, chooses ROLL, die = 2,3,4,5,6	**$12**	**$10**

III. EXPERIMENTAL RESULTS

IIIa. Data summary

Communication is totally ineffective when low-talent B's cannot participate, but has a dramatic effect on low-talent B choices (and leads to a modest increase in A's *In* rate) when participation is feasible. Figures 4 and 5 present low-talent B *Don't* rates and A *In* rates by treatment ("NM" means no message and "M" means message)[8] and Table 2 summarizes the effect of communication on behavior for the (5,7)-, (5,5)-, and (7,7)-games.

Figure 4 - Low-B's Don't Rate Across Treatments



[8] High-talent B choices are omitted, as they are invariably (63 of 63 times) *Roll* in our sessions.

Figure 5 - A's In rate Across Treatments

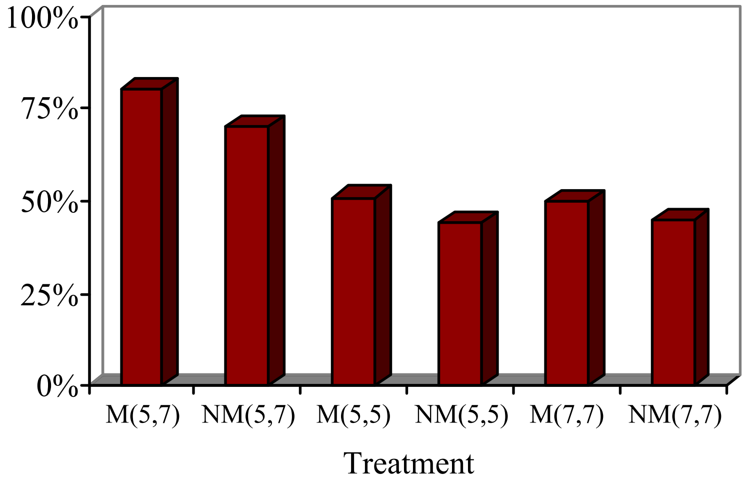


Table 2: Rates by Treatment and Tests for the Effect of Communication

	Low B's *Don't*			A's *In*		
Treatment	M	NM	*Z*-stat	M	NM	*Z*-stat
(5,7)	18/23 (78%)	8/20 (40%)	2.56***	33/41 (80%)	28/40 (70%)	1.09
(5,5)	3/16 (19%)	2/13 (15%)	0.24	24/47 (51%)	20/45 (44%)	0.64
(7,7)	2/11 (18%)	3/13 (23%)	-0.29	21/42 (50%)	18/40 (45%)	0.45

M/NM mean that no messages/messages were feasible. The *Z*-stat reflects the test of proportions across M and NM. *** indicates p < 0.01, one-tailed test.

Summarizing the results, the only case in which communication led to a significant increase was for low-talent B's in the (5,7) game, where the *Don't* rate nearly doubles, to 78%. Note that this rate is more than quadruple the *Don't* rates with communication in the two non-participation games, with statistical significance at $p < 0.001$ for each comparison.[9] The proportions of *Don't* are very close in the (5,5)- and (7,7)-games, whether or not there is

[9] Unless otherwise stated, the test used is the test of the difference of proportions (Douglas Glasnapp & John Poggio 1985); all *p*-values reflect two-tailed tests, unless otherwise stated.

communication. In general, it seems that low-talent B's refuse to step aside when there is no available Pareto improvement over A's outside option, but are often willing to accept lower payoffs than high-talent B's when participation is feasible.

Communication only affects A's behavior to a modest and insignificant degree, resulting a slight increase in the *In* rate in each of the three games. There is nevertheless a higher *In* rate in the (5,7)-game than in either of the other games both when communication is possible (Z = 2.88 and Z = 2.91, $p < 0.01$ in both cases) and when it is not (Z = 2.37 and Z = 2.26, $p < 0.025$ in both cases). The *Don't* rate in the (5,7)-game without communication is about twice as high as in the other games; however, the differences with respect to the other games is no more than marginally significant, perhaps due to the low number of observations. The test of proportions on the no-communication *Don't* rates gives Z = 1.50 and Z = 1.10, respectively, or Z = 1.55 for the pooled data from the (5,5)- and (7,7)-games; if we use one-tailed tests (which seem natural here), we get p = 0.067, p = 0.136, and p = 0.061 for these comparisons.

IIIb. Message content

What messages were sent? Free-form messages can potentially be classified in a variety of ways. To simplify the analysis, we assume that B can make a statement regarding his type (*Low* or *High*) and his choice (*Don't* or *Roll*), or stay silent. This produces five possible communication choices LD, LR, HD, HR, and S, where the notation in the first four cases refers to messages "I'm *Low* and I'll choose *Don't*", etc., with S representing silence. Ninety-three percent of all messages (121 of 130) can be assigned to one of these categories; in the other messages B stated that he was a low-talent B without implying an action. There is no doubt room for discussion in some cases regarding the classification; in any case, the precise messages are presented in the online data file on the *American Economic Review* website, where we also provide a richer classification scheme.

In Tables 3-5 below, we break down our results with communication according to the type of message sent and the actions that were observed thereafter. Notice that we *never* observe a LR- or HD-message.

Table 3: Messages and Outcomes in (5,7)-Treatment

		LD	LR	HD	HR	S	Other	Total
Low B	*Out*	1	0	0	0	3	1	5
	In, R	0	0	0	5	0	0	5
	In, DR	13	0	0	1	4	0	18
	Total	14	0	0	6	7	1	28
High B	*Out*	0	0	0	1	2	0	3
	In, R	2	0	0	8	0	0	10
	In, DR	0	0	0	0	0	0	0
	Total	2	0	0	9	2	0	13

LD = *Low & Don't*; LR = *Low & Roll*; HD = *High & Don't*, HR = *High & Roll*, S = Silence

Table 4: Messages and Outcomes in (5,5)-Treatment

		LD	LR	HD	HR	S	Other	Total
Low B	*Out*	0	0	0	2	11	3	16
	In, R	1	0	0	4	8	0	13
	In, DR	1	0	0	0	2	0	3
	Total	2	0	0	6	21	3	32
High B	*Out*	0	0	0	6	1	0	7
	In, R	0	0	0	5	3	0	8
	In, DR	0	0	0	0	0	0	0
	Total	0	0	0	11	4	0	15

LD = *Low & Don't*; LR = *Low & Roll*; HD = *High & Don't*, HR = *High & Roll*, S = Silence

Table 5: Messages and Outcomes in (7,7)-Treatment

		LD	LR	HD	HR	S	Other	Total
Low B	*Out*	1	0	0	3	10	4	18
	In, R	1	0	0	5	3	0	9
	In, DR	1	0	0	0	0	1	2
	Total	3	0	0	8	13	5	29
High B	*Out*	1	0	0	0	2	0	3
	In, R	0	0	0	6	4	0	10
	In, DR	0	0	0	0	0	0	0
	Total	1	0	0	6	6	0	13

LD = *Low & Don't*; LR = *Low & Roll*; HD = *High & Don't*, HR = *High & Roll*, S = Silence

First, consider the messages of the low-talent B's. In the (5,5)-game, two chose LD, while six chose HR, and 21 chose Silence. The distribution of messages was similar for the low-talent B's in the (7,7)-game (the Chi-square test gives $\chi^2_2 = 1.91$, $p = 0.384$), where three chose LD, eight chose HR, and 13 chose Silence. However, the patterns are quite different in the (5,7)-game, where 14 low-talent B's chose LD, four chose HR, and seven chose Silence (the Chi-square test gives $\chi^2_2 = 16.19$, $p = 0.000$ and $\chi^2_2 = 10.23$, $p = 0.006$ for the respective comparisons). Overall, the rate of LD-messages from low-talent B's is much higher when they can potentially participate in a Pareto-improvement than when they cannot (50% versus 8%, $Z = 4.46$, $p = 0.000$), while the rate of Silence is much lower (25% versus 56%, $Z = -2.70$, $p = 0.007$).[10]

With respect to the responses of the A's to these messages, we see that 'promise' (HR and LD) messages induce *In* 53% of the time in the (5,5)-game, 72% of the time in the (7,7)-game, and 94% of the time in the (5,7)-game. As the rate in the (5,7)-game is significantly higher

[10] Regarding the messages of the high-talent B's, nine chose HR, two chose Silence, and two chose LD in the (5,7)-game; 11 chose HR and four chose Silence in the (5,5)-game; six chose HR, six chose Silence, and one chose LD in the (7,7)-game. The proportions of HR-messages in the three games do not differ significantly from any other.

than the rate in either of the other games ($Z = 3.33$, $p = 0.001$ and $Z = 2.01$, $p = 0.045$ for the respective comparisons), A’s seem to believe that promises are more credible in this case.

IIIc. Patterns of lies, truth & action

We now proceed to present some observations regarding the structure of lies, truth, and action in our data set. We shall find it useful to refer to ‘plans-of-action’, equivalence classes of strategies that specify a message plus subsequent *Don't* or *Roll* choice, as in the following examples that explain our associated notation:

LD-then-D = LD-message + *Don't* (in response to *In*)
HR-then-R = HR-message + *Roll*
S-then-R = Silence + *Roll*

There are striking patterns in the message-action combinations for the low-talent B’s. We focus on the messages LD and HR, which may be viewed as forms of promises each of which might induce A to choose *In*. Notice that given these two message options there are two possible ways for low-talent B’s to act ‘trustworthy’, either LD-then-D or HR-then-D; the first of these does not involve being exposed as having sent a deceitful message, while the second one does, but in each case the low-talent B at last makes the non-opportunistic choice. Overall, low-talent B’s choose LD-then-D 15 times, while choosing HR-then-D only once; a binomial test shows that this difference is not random ($Z = 3.21$, $p = 0.001$).

There are also two possible ways for low-talent B’s to act ‘opportunistically’, either HR-then-R or LD-then-R; once again, the first of these does not involve being exposed as having sent a deceitful message, while the second one does. Overall, LD-then-R occurred only twice, while HR-then-R occurred 14 times; a binomial test shows that this difference is not random (Z = 3.00, p = 0.003).[11]

We wish to highlight the remarkable degree of trust and trustworthiness behavior that is induced by LD-messages in the (5,7)-game. Not only is it the case that A responds with *In* 13 of 14 times when a low-talent B sends a LD-message, but it is also true that every (13 of 13) low-talent B who sends a LD-message chooses *Don't* when given the option. In fact, all five low-talent B's who chose *Roll* were amongst the six low-talent B's who had sent a HR-message. The difference in the *Don't* rates (100% versus 17%) is of course highly significant (Z = 3.83, p = 0.000).

These systematic patterns may offer some lessons-for-life regarding whom to trust and how to detect lies: Those people who confess to having low talent will perform up to the level of their ability. On the other hand, one should be skeptical of those who claim to be the best, as liars lurk among them.

IV. BEHAVIORAL THEORY

When players are selfish, inefficient outcomes are predicted. This conclusion is unchanged when agents communicate. Models of distributional preferences such as Ernst Fehr & Klaus Schmidt (1999), Gary Bolton & Axel Ockenfels (2000), and (part of) Charness & Matthew Rabin (2002) provide an alternative approach that can accommodate more cooperative behavior if players dislike payoff inequality or have tastes for social efficiency. However, these models can not explain why communication makes low-talent B's more likely to choose *Don't* in the (5,7)-game, as the material payoff distributions do not depend on the preceding words.

[11] High-talent B's chose LD-then-R twice, HR-then-R 19 times, and S-then-R seven times.

Instead, we examine two behavioral models that permit communication to foster trust & cooperation: cost-of-lying and guilt-from-blame. We are not claiming that these are the only relevant behavioral theories, only that recent developments (discussed in more detail below) suggest that they are worth scrutiny. Throughout we make the admittedly unrealistic assumption that apart from cost-of-lying or guilt-from-blame the only thing motivating a player is how much money that player gets. Moreover, we assume that the key psychological parameters involved (k in the case of cost-of-lying, θ in the case of guilt-from-blame) are commonly known among the players. As we deal with some fairly non-standard theory, we hope this approach helps highlight key insights regarding the psychological mechanisms at work, uncluttered by complicated signaling issues that otherwise would have to be addressed alongside.

IVa. Cost-of-lying

Vanberg (2007) presents evidence suggesting that a preference for promise-keeping may explain Charness & Dufwenberg's (2006) data, and the more general notion of cost-of-lying has been emphasized by several scholars (see footnote 3). The key idea is that a person who utters a lie experiences an associated cost $k>0$. If there can be no communication there can be no cost-of-lying, so in the no-communication games the predictions correspond to the case with selfish preferences described in section I: A chooses *Out* and B chooses *Roll* independently of talent.

In the communication games, however, the outcome may be improved. To see this let us first state precisely how we assume payoffs are affected. For player A (who cannot lie) payoffs will be as indicated in Figures 1-3 for any corresponding path of play. For each type of player B, payoffs will be as indicated in Figures 1-3 except that we must deduct k following paths that entail lies. For example, in the (5,7)-game, following path (*Low*, HD, *Out*) low-talent B's payoff is 5-*k* because he lied about his talent; following path (*Low*, LD, *In*, *Roll*) low-talent B's payoff is 10-*k* because he lied about his choice.[12]

Observation 1: In a (5,7)-communication game with cost-of-lying:

(i) If $k>3$ the strategy profile where A chooses *Out* and B chooses *Roll* independently of talent (and message) is not a SE.

(ii) If $k>3$ there is a SE where low-talent B uses plan-of-action LD-then-D, high-talent B uses HR-then-R, and A responds to messages LD and HR with *In*.

(iii) If $0<k<3$ the pattern of behavior described in (ii) can not appear in any SE.

The proof is in Appendix A.

Parts (i) and (ii) of Observation 1 imply that adding communication when players have high cost-of-lying *fundamentally alters the prediction* relative to the case with selfish preferences. (As we shall see below, the guilt-from-blame theory discussed below does not have the analogous property.) The SEs described are not unique.[13] However, the prediction described in part (ii) is most compelling because it could also be obtained via solution concepts that do not assume equilibrium behavior, e.g. iterated elimination of weakly dominated strategies (applied to the game's normal form, treating low- and high-talent B as separate players) or extensive-form rationalizability (Pearce 1984; see also Battigalli 1995). It may also be seen as capturing an idea from the literature on cheap talk (non-binding costless communication): Language conveys exogenously given meaning and players tend to believe what is said as long as such belief is

[12] A list of all cases where B's payoff is decreased by *k* comprises those end nodes reached by the following paths: (*Low*, HD, *Out*), (*Low*, HR, *Out*), (*High*, LD, *Out*), (*High*, LR, *Out*), (*Low*, LD, *In*, *Roll*), (*Low*, LR, *In*, *Don't*), (*Low*, HD, *In*, *Roll*), (*Low*, HD, *In*, *Don't*), (*Low*, HR, *In*, *Don't*), (*Low*, HR, *In*, *Don't*), (*High*, HD, *In*, *Roll*), (*High*, HR, *In*, *Don't*), (*High*, LD, *In*, *Don't*), (*High*, LD, *In*, *Roll*), (*High*, LR, *In*, *Don't*), and (*High*, LD, *In*, *Roll*).

[13] For example, with $k\in(3,5)$ pooling by low- and high-talent B's on message LD is sustainable in SE (say with out-of-equilibrium inferences assigning probability 1 to messages LR, HD, HR, and S coming from low-talent B). With $k<3$ there exist mixed strategy SEs where A chooses *Out* except in response to HR where he chooses *In* with probability $k/5$; low-talent B uses HR-then-R with probability 1/2 and S-then-R with probability 1/2; high-talent B uses HR-then-R.

consistent with rationality and the incentives given in the game.[14] Ponder the following story of *commitment* captured by the SE highlighted in part (ii): Each agent reveals his talent and cooperative choice-intention, and he neither lies nor reneges because that would trigger too much cost-of-lying.

The predictions for the (5,5)- and (7,7)-games are similar. We focus on the high *k* cases:

Observation 2: In a (5,5)- [(7,7)-]communication game with cost-of-lying, if *k*>5 [*k*>3] there is a SE where low-talent B uses plan-of-action LD-then-D, high-talent B uses HR-then-R, and A responds to messages LD and HR with *In*. There is also a SE where low-talent B uses S-then-R, high-talent B uses HR-then-R, and A chooses *Out* except in response to message HR.

The proof is in Appendix A.

Observation 2 does not single out a particular choice for a low-talent B. Low-talent B may in SE use either LD-then-D or S-then-R; A would respond with *In* or *Out*, respectively, and A and the low-talent B would both get the same payoff regardless so there are no welfare consequences. The essence of Observation 2 is that high-talent B can signal his presence and intention with message HR, which is credible since low-talent B won't copy as *k* is too high. Player A chooses *In* in response, and efficiency is obtained.[15]

The difference in dollar payoffs for a low-talent B between choices *Don't* and *Roll* is higher in the (5,5)-game than in the (5,7)- and (7,7)-games (10-5=5 instead of 10-7=3) and we

[14] For previous work that explores similar assumptions, see Rabin (1990), Joseph Farrell (1993), Farrell & Rabin (1996), Vincent Crawford (2003), Andreas Blume & Andreas Ortmann (2007), and Demichelis & Weibull (2008).

[15] As with Observation 1, the described SEs are not the only ones, just the plausible ones. There is also a SE where low-talent B uses S-then-R, high-talent B uses HD-then-D (!), and A assigns probability 1 to any messages except HD coming from low-talent B and responds to every message by *Out*. This pattern of behavior is, however, not plausible in the sense that it is again ruled out by iterated elimination of weakly-dominated strategies or extensive-form rationalizability, or the idea that players tend to believe what is said (here applied to message HR) as long as such belief is consistent with rationality and the incentives given.

need *k*>5 rather than *k*>3 to argue in favor of an efficient outcome. As we argued in section I, the (5,7)- and (5,5)-games compare well, in the sense that one moves from the former to the latter through a subtle change in the underlying economic story (moving from asymmetric-but-positive agent gains to a step-aside-completely scenario). We suggested that it was intuitive that that change alone may cause trust & cooperation to deteriorate. A comparison of Observations 1 & 2 highlights why, with respect to testing that idea experimentally, a comparison of the (5,7)- and (5,5)-games is confounded in that different costs-of-lying are needed to support efficient outcomes in the two cases.[16] This explains why we also consider the (7,7)-game, which avoids this confound.

Let us finally, then, recall the data from section III and reflect on how well the cost-of-lying model accommodates it. First, while cost-of-lying may help explain why communication fosters trust & cooperation in the (5,7)-game (Observation 1), it provides equally strong support for an efficiency-enhancing effect in the (7,7)-game. This prediction was not borne out by the data, as trust & cooperation are distinctly lower in the (7,7)-game than in the (5,7)-game. Cost-of-lying alone does not help us explain why it matters whether we have asymmetric-but-positive gains or a step-aside-completely scenario. Second, recall the results of section IIIc, concerning patterns of lies & truth. Interpret these as suggesting that trustworthy low-talent Bs (who choose *Don't*) prefer plan-of-action LD-then-D to HR-then-D while opportunistic low-talent Bs (who choose *Roll*) prefer HR-then-R to LD-then-R. The cost-of-lying theory handles the first preference well; LD-then-D involves a lie while HR-then-D does not. The theory fares less well with the second preference; HR-then-R and LD-then-R both involve lies, so low-talent B should

[16] A comparison of the two games would be similarly confounded were we to take distributional preferences into account. For example, if a low-talent B is inequity averse he is more prone to choose *Don't* in the (5,7)-game than in the (5,5)-game. And an analogous confound arises with guilt-from-blame (cf. below).

be indifferent between the two plans-of-action. This prediction is not easy to reconcile with our data, where only HR-then-R is used.[17]

IVb. Guilt-from-blame

For completeness and easy reference, we reproduce a condensed version of the Battigalli & Dufwenberg (2007) theory of guilt-from-blame in Appendix B. Here in the main text we instead address the reader who wants to get the intuition for the general theory through a verbal summary, followed by SE definitions that apply to our specific games.

The guilt-from-blame theory is developed within a framework which admits a description of players beliefs about beliefs about … choices. A one-sentence summary of the essence could be: Player *i* experiences guilt-from-blame depending on how much player *j* blames *i* for being willing to let down *j*. A more understandable multi-sentence summary breaks that down further: First, for each end node in a game tree, measure how let down *j* is by comparing how much material payoff *j* initially believed (at the root of the game tree) he would get to what he actually got at that end node. Second, calculate how much of that let down is caused by *i*, in the sense that it could have been averted had *i* chosen differently. Third, calculate *i*'s initial belief regarding the extent to which *i* will cause *j* to be let down. Fourth, for each end node, calculate *j*'s belief regarding *i*'s initial belief regarding the extent to which *i* would cause *j* to be let down. Say that this is how much *j* would blame *i* if he knew he were at that end node. Finally, assume that *i* suffers from guilt-from-blame to the extent that he believes he is blamed. *i*'s utility trades off his material gain against such guilt-from-blame.

[17] One way to fix this could be to assume that HR-then-R entails less cost-of-lying than LD-then-R, because the latter plan-of-action includes *two* lies (about talent and choice) while the former includes only *one* (about talent). The cost-of-lying theory we developed do not allow this possibility and we do not explore it further here.

We now apply these ideas formally to our (5,7)-game without communication. Summarize by p^{In}, p_L^R, and p_H^R the probabilities that A chooses *In*, low-talent B's choose *Roll*, and high-talent B's choose *Roll*, respectively. As indicated in the previous paragraph, guilt-from-blame calculations involve taking into account beliefs about beliefs about ... these numbers. However, two assumptions will allow us to side-step much of this complexity: First, since we shall focus on equilibria (SEs), we can assume that beliefs are correct.[18] Thus we can use p^{In}, p_L^R, and p_H^R rather than higher-order beliefs about these numbers, as long as we keep in mind the underlying interpretations. Second, we assume that A's and high-talent B's cannot feel guilt. This seems psychologically sensible in our specific context, as A's and high-talent B's have no choice that can in expectation hurt another player. Mathematically this allows us to simplify by assuming maximization of expected material payoff for these players.

Thus, since high-talent B's and A's are selfish and since in SE beliefs are correct, we have $p_H^R = 1$ while p^{In} must maximize A's subjectively expected material payoff, which we here denote by μ. With $p_H^R = 1$ we have:

$$\mu = (1 - p^{In}) \cdot 5 + p^{In} \cdot (\tfrac{2}{3} \cdot [(1 - p_L^R) \cdot 7 + p_L^R \cdot 0] + \tfrac{1}{3} \cdot [\tfrac{5}{6} \cdot 12 + \tfrac{1}{6} \cdot 0]) = (1 - p^{In}) \cdot 5 + p^{In} \cdot (8 - \tfrac{14}{3} \cdot p_L^R)$$

To state and explain low-talent B's utility we need μ as well as two more key variables, which we label λ and θ. λ is the probability A assigns to the leftmost node in the information set where she receives a 0 payoff. In SE, applying Bayes' rule, and using $p^{In} = 1$, we get:

$$\lambda = \frac{\tfrac{2}{3} \cdot p_L^R}{\tfrac{2}{3} \cdot p_L^R + \tfrac{1}{3} \cdot \tfrac{1}{6}} = \frac{12 p_L^R}{12 p_L^R + 1}$$

[18] Battigalli & Dufwenberg (2009) extend Kreps & Wilson's SE definition to psychological games.

We can now state low-talent B's utility and best response (and in the process introduce θ). We first discuss the case of an SE where A chooses *In* (p^{In}=1); Definition 1 below will consider also cases where p^{In}<1. In such an SE, Low-talent B experiences guilt-from-blame only if he chooses *Roll*, the choice that hurts player A and that might lead A to blame low-talent B. To determine his best response low-talent B compares the (guilt-from-blame-free) payoff of 7 from choosing *Don't* to the payoff from *Roll*, which is

$$10-\theta\cdot\lambda\cdot\min\{7,\mu\}$$

This expression describes utility as material payoff (=10) minus guilt-from-blame (=$\theta\cdot\lambda\cdot\min\{7,\mu\}$). We explain the latter term, walking through its factors from right to left. The expression $\min\{7,\mu\}$ measures how much A would blame low-talent B (and how much guilt-from-blame low-talent B would then experience) *were it known* that low-talent B chose *Roll*; μ is the difference between what A initially expected (=μ) and what he actually received (=0) due to low-talent B's opportunistic choice. The 7 is present in the expression because the blame/guilt is capped at 7, since this is the full payoff difference that low-talent B actually controls. Regarding λ, note that because of A's information set across the end nodes where he receives 0, he will actually never know for certain that low-talent B chose *Roll*. λ captures an assumption that a low-talent B is sheltered from guilt-from-blame to the extent that A isn't sure that B is blameworthy. Notice that A assigns probability 1-λ to the event that she received a payoff of 0 due to path (*High*, *In*, *Roll*, *Failure*), which would just be bad luck and no fault of a low-talent B. Finally, θ is a non-negative constant, describing how sensitive *i* is to feelings of guilt-from-blame. If θ = 0, a low-talent B would be selfish.

At this point we wish to make a comment about our approach: One may model guilt in many ways. Battigalli & Dufwenberg (2007) offer two models. In one variety (simple guilt) player *i* internalizes the emotion in the sense that he feels guilt when he believes he lets down *j*, regardless of what *j* believes *i*'s intentions are.[19] Guilt-from-blame is the other variety, where guilt is driven rather by what *i* believes *j* believes about *i*'s intentions as regards letting *j* down. The goal of our paper is *not* to test simple guilt against guilt-from-blame. Rather we focus only on the latter concept (which we compare with cost-of-lying), the reason being a recent string of papers (Jason Dana, Daylian Cain & Robyn Dawes 2006, Dana, Roberto Weber & Jason Xi Kuang 2007, Tomas Broberg, Ellingsen & Johannesson 2007, Steven Tadelis 2008, James Andreoni & B. Douglas Bernheim 2009, Edward Lazear, Ulrike Malmendier & Weber 2009,) that suggest in various ways that players are more prone to selfless choice to the extent that others will know about it.[20] Guilt-from-blame caters to such concerns through the way λ affects utility.[21]

Drawing on the above notations and calculations, we now state SE conditions formally:

Definition 1: A SE in the (5,7)-game, when low-talent B is sensitive to guilt-from-blame, is a triple (p^{In}, p_L^R, p_H^R) such that:

(i) p^{In} maximizes $\mu = (1-p^{In})\cdot 5 + p^{In}\cdot(8-\frac{14}{3}\cdot p_L^R)$ [player A best responds]

(ii) p_L^R maximizes (1- p_L^R) $\cdot 7 +$ $p_L^R \cdot (10 - p^{In}\cdot\theta\cdot\lambda\cdot\min\{7,\mu\})$, treating μ and λ as being fixed [low-talent B best responds]

[19] Although the distinction with guilt-from-blame had not yet been conceptualized when Charness & Dufwenberg (2006) was written, in retrospect we see that simple guilt was in focus in that paper.

[20] For example, Dana, Cain & Dawes let dictators either divide $10 (in which case the recipient learned of the dictator game and the dictator's choice) or choose to exit and take a smaller amount, in which case the would-be recipient would not learn of the dictator game. Many people choose to exit. In fact, 43% exit when the would-be recipient would learn of the dictator game without exit, but only 4% exit when the would-be recipient would never learn of the dictator game even if exit is forgone. Tadelis uses the same game as Charness & Dufwenberg (2006), but varies whether the principal will learn of the actual choice made by the agent. In two separate comparisons, he finds that *Roll* rates are nearly twice as high with this exposure than when the agent knows that the principal will not learn his choice.

[21] Some of our reviewers suggested that we should focus also on simple guilt and explore treatments that test the models against each other (for example via treatments where A always learns B's choice). We are certainly sympathetic to this suggestion, but we have chosen to leave these issues for us or others to explore in the future.

(iii) $p_H^R = 1$ [high-talent B best responds]

(iv) $\lambda = 12 p_L^R / (12 p_L^R + 1)$ [equilibrium expectations]

The only part of Definition 1 that has not already been motivated is (ii). We discussed the case of an SE where A chooses *In* (p^{In}=1) before, but (ii) handles also cases where p^{In}<1. As the theory is constructed, blame and guilt are only relevant to the extent that A believes low-talent B believes initially that he set out to let A down. This is captured through the presence of p^{In} in (ii), reflecting the understanding that in an SE low-talent B initially believes that the probability that A will choose *In* equals p^{In}. The lower is p^{In}, the less low-talent B initially believes (before updating based on A's observed choice) that he can influence A's payoff so the less relevant is blame (vanishing when p^{In}=0). Note also that the reason low-talent B treats μ and λ as being fixed in (ii) is that μ and λ depend on beliefs of A, which low-talent B cannot influence. Of course, in equilibrium, the players have correct beliefs (as reflected explicitly in (iv) and implicitly in (i) and (ii)).

We can state analogous definitions for the (5,5)- and (7,7)-games. For the (7,7)-game, the definition is *identical*, except that the two numbers "5" in Definition 1 should be replaced by "7". As regards the (5,5)-game the specification changes more:

Definition 2: A SE in the (5,5)- game, when low-talent B is sensitive to guilt-from-blame, is a triple (p^{In}, p_L^R, p_H^R) such that:

(i) p^{In} maximizes $\mu = (1 - p^{In}) \cdot 5 + p^{In} \cdot (\frac{20}{3} - \frac{10}{3} \cdot p_L^R)$

(ii) p_L^R maximizes $(1\text{-}\, p_L^R) \cdot 5 +\ p_L^R \cdot (10 - p^{In} \cdot \theta \cdot \lambda \cdot \min\{5, \mu\})$, treating μ and λ as being fixed

(iii) $p_H^R = 1$

(iv) $\lambda = 12 p_L^R / (12 p_L^R + 1)$

Applying these definitions we get multiple SEs once θ is high enough:

Observation 3: In both the (5,7)-game and the (7,7)-game, when low-talent B is sensitive to guilt-from-blame:

(i) For any $\theta \geq 0$, there is a SE with $(p^{In}, p_L^R, p_H^R) = (0, 1, 1)$

(ii) If $\theta \geq \frac{25}{42}$, there is a SE with $(p^{In}, p_L^R, p_H^R) = (1, \frac{1}{28\theta - 12}, 1)$

Observation 4: In the (5,5)-game, when low-talent B is sensitive to guilt-from-blame:

(i) For any $\theta \geq 0$, there is a SE with $(p^{In}, p_L^R, p_H^R) = (0, 1, 1)$

(ii) If $\theta \geq \frac{7}{6}$, there is a SE with $(p^{In}, p_L^R, p_H^R) = (1, \frac{1}{12\theta - 12}, 1)$

The proofs are in Appendix A.

Note several things: First, parts (i) of Observations 3 & 4 describe inefficient zero-trust play by A and no cooperation by low-talent B's. The intuition for why this pattern is allowed for any θ is that if low-talent B initially expects A to choose *Out*; then B believes that A then can't blame low-talent B, who therefore does not feel guilt. It is true that if A were to deviate, then a low-talent B would realize that he can affect A's payoff, so in principle one might imagine that guilt could come into play. However, as the theory is constructed (through the presence of factor p^{In} in parts (ii) of Definitions 1 & 2), blame and guilt are only relevant to the extent that A believes low-talent-B initially believes that low-talent B set out to let A down. Second, it is impossible in each of the games to have a full-trust-and-cooperation SE with $(p^{In}, p_L^R, p_H^R) = (1, 0, 1)$. In that case we would have $\lambda = 0$ and low-talent B would be entirely sheltered from blame and guilt and so choose *Roll,* i.e. $p_L^R = 1$, a contradiction. Instead, the SEs reflecting the most trust and cooperation involve mixing by low-talent B. For example, in the (5,7)-game, he chooses *Roll* with probability $p_L^R = \frac{1}{28\theta - 12}$. Note that $p_L^R \to 0$ as $\theta \to \infty$.

Third, the SEs described for the (5,7)- and (7,7)-games coincide (Observation 3; Appendix A also comments on some additional SEs for the (5,7)-game that are not covered in Observation 3). The (5,5)-game is different (Observation 4); because of the difference between parts (ii) of Definitions 1 & 2 there is a confound for comparing behavior in the (5,5)- and (5,7)-games analogous to what we discussed for cost-of-lying. So again our main comparison as regards whether the theory can explain why it matters whether we have asymmetric-but-positive gains or a step-aside-completely scenario will center on comparing the (5,7)- and (7,7)-games. Fourth, unlike in the case with cost-of-lying, rationalizability will not help pin down a clear prediction; one can show that for any $\theta \geq \frac{25}{42}$ each of A's and low-talent B's strategies is rationalizable (as defined by Battigalli & Dufwenberg 2009 who extend Pearce's extensive-form rationalizability notion to psychological games). Fifth, in light of the presence of multiple SEs when θ is large enough, we face an equilibrium-selection problem.

What happens when the communication stage is added (with messages LD, LR, HD, HR, and S, just as before)? The first thing to note is that (unlike in the case with cost-of-lying) we cannot hope to get an automatic move of the set of SEs in the direction of enhanced efficiency. In particular, there is no SE with full revelation + separation + honesty. To see this, imagine for example that low-talent B uses LD-then-D, high-talent B uses HR-then-R, and A chooses *In* if and only if she gets message LD or HR. The argument regarding why this cannot be part of a SE is analogous to that which ruled out, for the games without communication, an SE with $(p^{In}, p_L^R, p_H^R) = (1, 0, 1)$. If inferences were based on HR-messages never coming from low-talent B's, then a low-talent B would be safe choosing *Roll* as he wouldn't be blamed if he actually sent a HR-message and then chose *Roll*. Following HR, we'd have $\lambda = 0$ and a complete shelter for low-talent B's feelings of guilt-from-blame.

On the other hand, every pattern of SE play described for the games without communication is also attainable via some SE in the games with communication. For example, consider the (5,7)-game and suppose $\theta \geq \frac{25}{42}$. The SE with (p^{In}, p_L^R, p_H^R) = (0, 1, 1) in part (i) of Observation 3 could be matched if low- and high-talent talent both use HR-then-R while A responds to any message message with *Out*. The SE with (p^{In}, p_L^R, p_H^R) = (1, $\frac{1}{28\theta-12}$, 1) in part (ii) of Observation 3 could be matched if low-talent B uses HR-then-R with probability $\frac{1}{28\theta-12}$ and otherwise LD-then-D; high-talent B uses HR-then-R; A responds to both LD and HR with *In* but would respond to any other message with choice *Out*. Note that, in this SE, a low-talent B's randomization occurs at the message stage only; once the message is sent (LD or HR) the subsequent pure choice is given by the plan-of-action in question.

Communication may, however, help the players coordinate on a favorable SE. One-way communication has been found to lead to coordination on a strictly Pareto-superior equilibrium in papers such as Charness (2000). This could be relevant to the two SEs for the (5,7)-game with $\theta \geq \frac{25}{42}$ described in the previous paragraph, which are indeed strictly Pareto-ranked. But this idea does not extend to the (5,5)- and (7,7)-games. While these games also exhibit multiple SEs, no strict Pareto-gains are available. In particular, a low-talent B lacks a strict incentive to sway A away from his choice *Out* by promising A he will choose *Don't*. The low-talent B gets exactly the same payoff from choosing *Don't* after A chooses *In* as when A chooses *Out*, as does A.

Let us finally, then, recall the data from section III and reflect on how well the guilt-from-blame theory accommodates it. First, even without communication non-selfish choice is possible if players are motivated by guilt-from-blame, so guilt-from-blame can help explain why in the experimental (5,7)-game we saw considerable deviations from the selfish SE. Second, guilt-from-blame may help explain why communication fosters additional trust and cooperation in the (5,7)-game as well as the observed differential effect of communication in this game in comparison with the (5,5)- and (7,7)-games, if we add the idea (admittedly from outside the guilt-from-blame theory proper) that one-sided communication helps players coordinate on a strictly Pareto-superior SE. However, this equilibrium selection argument hinges heavily on the idea that communication alters both the behavior of A and the beliefs of B about A's behavior (as reflected by the presence of the factor p^{In} in Definition 1(ii)). One may call to question how compelling this is, in light of the observation that while communication seems to considerably affect player B, the effect on player A is weaker (as reported in section IIIb).

Third, regarding the patterns of lies & truth reported in section IIIc, we already described an SE that accords well with the observed patterns (low-talent B randomizes between plans of actions HR-then-R and with probability $\frac{1}{28\theta-12}$ and LD-then-D; high-talent B uses HR-then-R; A responds to LD and HR with *In*). However, this SE does not rule out other SEs (for example, babbling equilibria or SEs that permute which messages are used). Nevertheless, we note that the following pattern of inferences would naturally produce the result that no low-talent B uses LD-then-R: Suppose B uses LD-then-R and A receives 0. A knows B lied, but whether A blames B depends on whether she thinks B had low or high talent. On the presumption that high-talent B's always choose HR-messages, A would interpret an LD-message as coming from a low-talent B. Given that inference, a low-talent B would refrain from LD-then-R, in line with the data.

We close this section with a comment on the methodology of testing models grounded in psychological game theory, for example models of guilt aversion. Previous work (starting with Dufwenberg & Gneezy 2000) has elicited or induced beliefs, which is often useful for direct tests. A recent paper by Ellingsen, Johannesson, Sigve Tjøtta & Gaute Torsvik (2009) calls to question the accuracy of some of these methods and the conclusions drawn. The issues are hardly settled,[22] but there is a concern to acknowledge. Our paper illustrates that to some extent the problems can be side-stepped, as models of belief-dependent utility may generate predictions that can be tested without belief data. The guilt-from-blame predictions we took to the data in this section concerned choices only.[23]

V. CONCLUSION

Samuel Goldwyn quipped "an oral contract isn't worth the paper it is written on." Contract theorists mainly agree, if not explicitly in writing, at least in the spirit of their work. Their basic models typically possess a unique equilibrium, which cannot be upset by the addition of communication. Yet, the human side of contracting seems a bit less dismal.

In this paper we explore whether and how communication can achieve beneficial social outcomes in a hidden-information context. It turns out that whether communication affects behavior depends crucially on whether low-talent agents can participate in an outcome that, compared to no contractual agreement, is a Pareto-improvement for the principal and the agent. When low-talent agents can participate in this way, communication is quite effective; the great majority of these agents behave cooperatively, foregoing the additional earnings that could be pocketed. However, when participation for low-talent agents is infeasible, selfish behavior on the part of these agents predominates, whether or not communication is feasible.

[22] Ernesto Reuben, Paola Sapienza & Luigi Zingales (2009) present evidence that to some extent goes against Ellingsen *et al*'s.

[23] This is of course not to suggest that we couldn't have conducted sharper tests had we had access to beliefs data (in particular beliefs about p^{In}, □, and □). However, we chose not to elicit these beliefs because we were concerned that eliciting conditional beliefs about beliefs up to the fourth order would have been confusing and time-consuming.

Our data exhibit some systematic patterns regarding how people lie and tell truth, in the game where there are gains for all (the (5,7)-game). Liars claim to be better than they are, as if they meant to suggest that the subsequent bad outcome was due to bad luck rather than opportunistic choice. Trustworthy people, on the other hand, truthfully reveal their level of talent and can then be relied upon to do as well as they can. These results provide some 'useful lessons' that, on extrapolation, may offer useful guidance for those who wish to tell if someone else is being honest. A claim that the agent has high talent should be viewed with some suspicion, as it often 'the big lie'. However, when participation is possible regardless of the agent's talent, the claim that someone has low talent but will do his best turns out to be completely reliable, and is in fact almost always believed by the principal; it seems that one can trust people who confess imperfections.

We present the predictions from two relevant behavioral models, one that involves a cost-of-lying and one that involves guilt-from-blame. When communication is allowed, both theories offer some scope for trust and cooperation, although the mechanisms differ. Under (high enough) cost-of-lying, incorporating communication leads to new equilibria that embody Pareto-gains (predictions that also obtain with solution concepts like iterated weak dominance or extensive form rationalizability). The cost-of-lying theory does not, however, predict a difference depending on whether or not Pareto-gains are feasible for both talent levels for B, as all those who gain can unilaterally credibly separate.

With guilt-from-blame on the other hand, allowing communication does not add new patterns of equilibrium play. There are multiple equilibria both with and without communication. Communication may, however, enhance trust & cooperation not by expanding the possible patterns of equilibrium play, but rather by facilitating equilibrium coordination.

Previous experimental studies (e.g. Charness 2000) have suggested that communication has this efficiency-enhancing property when equilibria can be strictly Pareto-ranked. The coupling of this idea with the guilt-from-blame theory may shed some light on why it is easier to obtain efficient outcomes when everyone gains than when some are excluded. Those who are excluded have nothing to gain by using communication to change the outcome in a way that is favorable to the other party, so they may lose credibility. Our experimental data is consistent with this story. However, it is somewhat puzzling that in the case when equilibria can be strictly Pareto-ranked communication seems to affect agents much more than principals.

As regards the stable patterns of lies & truths we observe, each of the two theories cost-of-lying and guilt-from-blame can capture these patterns to some degree. Since on balance both theories capture many but not all aspects of the data, we shall not declare a winner. Moreover, we do not claim that there may not be other theories that may describe the relevant motivational forces at work.[24] Our goal was not to formulate a new theory for explaining the data, but rather to check the performance of two specific theories we had reason to believe would have something to say. We hope our discussion may inspire efforts to further develop behavioral theory that can shed light on how communication and gain-for-all opportunities may shape trust and cooperation in partnerships.

Do the effects we have documented in a laboratory environment extend to the field? If so, then people may be substantially more prone to be cooperative when they can participate by having a voice and choosing an action that yields improvements in material payoffs for all

[24] For example, in sociology and social psychology there is the notion of impression management, which is the process through which people attempt to influence how other people perceive them. The earliest reference in this area is Erving Goffman (1956); for related contributions see Barry Schlenker (1980), John Tedeschi & Michael Riess (1981), and R. Lynn Hannan, Frederick Rankin & Kristy Towry (2006). As impression management has not been formalized mathematically, we chose not to analyze the predictions of this theory. (Or perhaps we do, if avoiding guilt-from-blame, as described in section IVb, may be seen as one form of impression management.)

parties involved than when the only way to gain is at the expense of others. This may have bearing on the understanding of many market situations, which differ as regards the availability of Pareto-improvements. For example, e-commerce furnishes settings in which the quality of the good traded is not readily observable, and it may or may not be the case that all sellers have the ability to provide a good that has value to buyers.[25]

REFERENCES

Akerlof, George. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics,* 84(3): 488–500.

Andreoni, James and B. Douglas Bernheim. 2009. "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," *Econometrica*, 77(5): 1607-36.

Battigalli, Pierpaolo. 1997. "On Rationalizability in Extensive Games," *Journal of Economic Theory,* 74(1): 40–61.

Battigalli, Pierpaolo and Martin Dufwenberg. 2007. "Guilt in Games," *American Economic Review Papers and Proceedings*, 97(2): 170-176.

Battigalli, Pierpaolo and Martin Dufwenberg. 2009. "Dynamic Psychological Games," *Journal of Economic Theory*, 144(1): 1-35.

Blume, Andreas and Andreas Ortmann. 2007. "The Effect of Costless Pre-play Communication: Experimental Evidence for Games with Pareto-ranked Equilibria," *Journal of Economic Theory*, 132(1): 274–90.

Bolton, Patrick and Mathias Dewatripont. 2005. *Contract Theory*. Cambridge, MA: MIT Press.

Bolton, Gary and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1): 166-93.

Broberg, Tomas, Tore Ellingsen and Magnus Johannesson. 2007. "Is Generosity Involuntary?" *Economics Letters*, 94(1): 32-37.

[25] We thank Ulrike Malmendier for the following example. If an internet seller expects buyers to only be interested in a brand-new item, he is likely to claim that the item for sale is new, whether or not it is. However, if the seller believes that there is a market for used items, perhaps he is much more likely to confess the item is used. Of course, this argument requires that online reputation systems are less-than-perfect, but people do to some extent game this system by tactics such as changing online identities after misbehavior.

Charness, Gary. 2000. “Self-serving Cheap Talk and Credibility: A Test of Aumann’s Conjecture,” *Games & Economic Behavior*, 33(2): 177-194.

Charness, Gary and Martin Dufwenberg. 2006. “Promises and Partnership,” *Econometrica*, 74(6): 1579-1601.

Charness, Gary and Martin Dufwenberg. Forthcoming. “Bare Promises: An Experiment,” *Economics Letters*.

Charness, Gary and Matthew Rabin. 2002. “Understanding Social Preferences with Simple Tests,” *Quarterly Journal of Economics*, 117(3): 817-69.

Chen, Ying, Navin Kartik and Joel Sobel. 2008. “Selecting Cheap-Talk Equilibria,” *Econometrica*, 76(1): 117-136.

Crawford, Vincent. 2003. “Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions,” *American Economic Review*, 93(1): 133-149.

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

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

Demichelis, Stefano and Jörgen Weibull. 2008. “Language, Meaning, and Games: A Model of Communication, Coordination, and Evolution,” *American Economic Review*, 98(4): 1292–1311.

Dufwenberg, Martin and Uri Gneezy. 2000. “Measuring Beliefs in an Experimental Lost Wallet Game,” *Games & Economic Behavior*, 30(2): 163-82.

Ellingsen, Tore & Magnus Johannesson. 2004. “Promises, Threats, and Fairness”, *Economic Journal*, 114(495): 397-420.

Ellingsen, Tore, Magnus Johannesson, Sigve TjØtta and Gaute Torsvik. 2010. “Testing Guilt Aversion,” *Games and Economic Behavior*, 68(1): 95-107.

Farrell, Joseph**.** 1993. “Meaning and Credibility in Cheap-Talk Games,” *Games and Economic Behavior*, 5(4), 514 –31.

Farrell, Joseph and Matthew Rabin. 1996. “Cheap Talk,” *Journal of Economic Perspectives*, 10(3): 103 –118.

Fehr, Ernst and Klaus Schmidt. 1999. “A Theory of Fairness, Competition, and Cooperation,” *Quarterly Journal of Economics*, 114(3), 817-68.

Geanakoplos, John, David Pearce and Ennio Stacchetti. 1989. “Psychological Games and Sequential Rationality,” *Games and Economic Behavior*, 1(1), 60-80.

Glasnapp, Douglas & John Poggio. 1985. *Essentials of Statistical Analysis for the Behavioral Sciences*. Columbus: Merrill.

Gneezy, Uri. 2005. "Deception: The Role of Consequences," *American Economic Review*, 95(1): 384-394.

Goffman, Erving. 1956. "Embarrassment and Social Interaction," *American Journal of Sociology*, 62(3): 264-271.

Hannan, R. Lynn, Frederick Rankin and Kristy Towry. 2006. "The Effect of Information Systems in Managerial Reporting: A Behavioral Perspective," *Contemporary Accounting Research*, 23(4): 885-918.

Kartik, Navin. 2009. "Strategic Communication with Lying Costs," *Review of Economic Studies*, 76(4): 1359-1395.

Kreps, David and Robert Wilson. 1982. "Sequential Equilibrium," *Econometrica*, 50(4): 863-894.

Lazear, Edward, Ulrike Malmendier and Roberto Weber. 2009. "Sorting and Social Preferences," mimeo.

Miettinen, Topi. 2008. "Contracts and Promises - An Approach to Pre-play Agreements,, SSH/EFI Working Paper No 707, Stockholm School of Economics.

Pearce, David. 1984. "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52(4): 1029–50.

Rabin, Matthew**.** 1990. "Communication Between Rational Agents," *Journal of Economic Theory*, 51(1): 144-70.

Reuben, Ernesto, Paola Sapienza and Luigi Zingales. 2009. "Is Mistrust Self-Fulfilling?" *Economics Letters*, 104(2): 89-91.

Schlenker, Barry. 1980. *Impression Management: The Self-Concept, Social Identity, and Interpersonal Relations*. Monterey/California: Brooks/Cole.

Sutter, Matthias. 2009. "Deception Through Telling the Truth?! Experimental Evidence from Individuals and Teams," *Economic Journal*, 119(534): 47-60.

Tadelis, Steven. 2008. "The Power of Shame and the Rationality of Trust," mimeo.

Tedeschi, John and Michael Riess. 1981. "Identities, the Phenomenal Self, and Laboratory Research". In *Impression Management Theory and Social Psychological Research*, ed. John Tedeschi, 3-22, New York: Academic Press.

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

Appendix A: Proofs

Proof of Observation 1

(i) In SE, if a low-talent B sends message LD he must follow up with choice *Don't* (since 7>10-*k*). (Note that it now also follows that A must respond by *In* to message LD; in SE high-talent B chooses *Roll* after message LD so by choosing *In* after LD player A gets at least 7>5).

(ii) The described SE profile describes sequentially rational play, as no player has a profitable unilateral deviation. To pin down a SE just add an appropriate specification for out-of-equilibrium beliefs, e.g. probability 1 to low-talent B following messages LR, HD, or S, and choices following those messages for each type of player B.

(iii) A low-talent B would have a unilateral incentive to deviate to HR-then-R.

Proof of Observation 2:

It is straightforward to assert that no player has a unilateral deviation incentive. We leave the specification of complete strategies and out-of-equilibrium inferences for the reader.

Proof of Observation 3:

(i) As seen in parts (ii) of Definitions 1 & 2, the guilt-from-blame term vanishes when p^{In}=0. A low-talent B thus chooses *Roll* since 10>7, so p_L^R=1. We know from Definition 1(iii) that $p_H^R = 1$, and we see that A's best response is *Out*. All in all, $(p^{In}, p_L^R, p_H^R) = (0, 1, 1)$.

(ii) A low-talent B randomizes and so must be indifferent between *Don't* and *Roll*: 7 = $10 - p^{In} \cdot \theta \cdot \lambda \cdot \min\{7, \mu\}$. This equation can be simplified:

- $p^{In} = 1$ is given by the SE
- $\mu = (1 - p^{In}) \cdot 5 + p^{In} \cdot (8 - \frac{14}{3} \cdot p_L^R) = 8 - \frac{14}{3} \cdot p_L^R$ since $p^{In} = 1$
- $p_L^R = \frac{1}{28\theta - 12}$ is given by the SE and since $\theta \geq \frac{25}{42}$ we get $p_L^R \leq \frac{3}{14}$, which in turn

  implies that $\mu = 8 - \frac{14}{3} \cdot p_L^R \geq 7$, so that $\min\{7, \mu\} = 7$.

Thus, we have that 7 = $10-\theta\cdot\lambda\cdot 7$. Plug in $\lambda=\frac{12p_L^R}{12p_L^R+1}$ and solve for p_L^R as a function of θ to verify that $p_L^R = \frac{1}{28\theta-12}$. We know from Definition 1(iii) that p_H^R = 1, and we see that A's best response is indeed *In*. All in all, (p^{In}, p_L^R, p_H^R) = (1, $\frac{1}{28\theta-12}$, 1).

Comment on Observation 3: All of the SEs described under part (ii) give A a payoff of at least 7, but in the (5,7)-game (and not the (7,7)-game) there are also SEs where A chooses *In* and receives a payoff in the range (5, 7). In these cases, $\frac{3}{14}<p_L^R\leq\frac{9}{14}$, and $\min\{7,\mu\}$=$\min\{7,8-\frac{14}{3}\cdot p_L^R\}$= $8-\frac{14}{3}\cdot p_L^R$. The dependence on p_L^R means that when calculating the SEs one must solve the quadratic equation $7=10-\theta\cdot\frac{12p_L^R}{12p_L^R+1}\cdot(8-\frac{14}{3}\cdot p_L^R)$, which describes the relevant indifference condition for a low-talent B. Additional SE appear for values of θ slightly lower than 25/42 (down to slightly more than 0.54), as well as much higher values of θ (<61/18). Some manipulations show that relevant roots satisfy p_L^R = $\frac{(24\theta-9)}{28\theta}\pm((\frac{24\theta-9}{28\theta})^2-\frac{3}{56\theta}))^{\frac{1}{2}}$ as well as $\frac{3}{14}<p_L^R<\frac{9}{14}$.

Proof of Observation 4: The technique is analogous to that in the proof of Observation 3 and is therefore omitted.

Appendix B: Guilt-from-blame and sequential equilibrium

The presentation is condensed from Battigalli & Dufwenberg (2007). Consider finite extensive game forms with features as follows: N, T, $t^0$and Z are, respectively, the player set, the set of nodes, the root, and the set of end nodes. $T \setminus Z$ is partitioned into subsets X_i of decision nodes for each $i \in N$ and the set of chance nodes X_c. $\sigma_c \,|(\cdot \,|\, x) > 0$ denotes the chance probabilities choices for each $x \in X_c$.

It is crucial to represent players' information also at nodes where they do not make choices. Thus, player i's information structure is a partition H_i of T that contains as a subcollection the standard information partition of X_i. H_i satisfies perfect recall and is a refinement of $\{\{t^0\}, X \setminus \{t^0\}, Z\}$; players know when they are at the root and when the game is over. Material (dollar) payoffs are given by functions $m_i : Z \to \Re$, $i \in N$ such that $m_i(z') \neq m_i(z'')$ implies $H_i(z') \neq H_i(z'')$; i observes his material payoff.

A pure strategy S_i specifies a contingent choice for each $h \in H_i$ with $h \subset X_i$. It is convenient to refer to pure strategies also of chance, i.e. functions $s_c : X_c \to T$ that select an immediate successor of each chance node; such strategies are chosen at random according to $\sigma_c = [\sigma_c(\cdot \,|\, x)]_{x \in X_c}$. The set of i's pure strategies is S_i and $S = S_c \times \prod_{i \in N} S_i$, $S_{-i} = S_c \times \prod_{j \neq i} S_j$. For any $h \in H_i$ and i, $S_i(h)$ is the set of i's pure strategies allowing h, and $S_{-i}(h) \subset S_{-i}$ is the set of profiles S_{-i} allowing h. $s \in S$ yields an end node $z(s)$. A behavioral strategy for i is an array σ_i of probability measures $\sigma_i(\cdot \,|\, h)$, $h \in H_i$, $h \subset X_i$, where $\sigma_i(a \,|\, h)$ is the probability of choice a at h. Given σ_i and perfect recall, one can compute conditional probabilities $\Pr_{\sigma_i}(s_i \,|\, h)$, $h \in H_i$ (even if $\Pr_{\sigma_i}(S_i(h)) = 0$).

For each $h \in H_i$ player i holds a conditional belief $\alpha_i(\cdot \,|\, h) \in \Delta(S_{-i}(h))$ about the co-players' strategies; $\alpha_i = (\alpha_i(\cdot \,|\, h))_{h \in H_i}$ is the system of first-order beliefs of i. Player i also holds, at each $h \in H_i$, a second-order belief $\beta_i(h)$ about the first-order belief system α_j of each co-player j, a third-order belief $\gamma_i(h)$ about the second-order beliefs, and so on. When focusing on SEs, as we shall do, one may assume that higher-order beliefs are degenerate point beliefs; identify $\beta_i(h)$ with a particular array of conditional first-order beliefs $\alpha_{-i} = [\alpha_j(\cdot \,|\, h')]_{j \neq i, h' \in H_j}$. A similar notational convention applies to other higher-order beliefs. The beliefs i would hold at different information sets are not mutually independent; they must satisfy Bayes' rule and common certainty that Bayes' rule holds (cf. Battigalli & Dufwenberg 2009). Players *initial* beliefs are those held at the information set $h^0 = \{t^0\}$.

Given s_j and $\alpha_j(\cdot \mid h^0)$ player j forms an initial expectation about his material payoff: $\mathrm{E}_{s_j,\alpha_j}[m_j \mid h^0] = \sum_{s_{-j}} \alpha_j(s_{-j} \mid h^0)\mathrm{m}_j(\mathrm{z}(s_j,s_{-j}))$. For any $z \in Z$ consistent with s_j, the expression $D_j(z,s_j,\alpha_j) = \max\{0, \mathrm{E}_{s_j,\alpha_j}[m_j \mid h^0] - \mathrm{m}_j(z)\}$ measures how much j is 'let down'.

If at the end of the game i knew z, $s_{-i} \in S_{\text{-i}}(z)$ and $\alpha_j(\cdot \mid h^0)$, then he could derive how much of $D_j(z,s_j,\alpha_j)$ is due to his behavior: $G_{ij}(z,s_{-i},\alpha_j) = D_j(z,s_j,\alpha_j) - \min_{s_i} D_j(\mathrm{z}(s_i,s_{-i}),s_j,\alpha_j)$. And given s_i, $\alpha_j(\cdot \mid h^0)$, and $\beta_i(\cdot \mid h^0)$, i can compute how much he initially expects to cause j to be let down: $G^0_{ij}(s_i,\alpha_i,\beta_i) = \mathrm{E}_{s_i,\alpha_i,\beta_i}[G_{ij} \mid h^0] = \sum_{s_{-i}} \alpha_i(s_{-i} \mid h^0) G_{ij}(\mathrm{z}(s_i,s_{-i}),s_{-i},\beta^0_{ij}(h^0))$, where $\beta^0_{ij}(h^0))$ denotes the initial (point) belief of i about $\alpha_j(\cdot \mid h^0)$.

Now suppose $z \in Z$ is reached. The conditional expectation $\mathrm{E}_{\alpha_j,\beta_j,\gamma_j}[G^0_{ij} \mid H_j(z)]$ measures j's inference regarding how much i intended to let j down, or how much j 'blames' i. Player i is affected by guilt-from-blame if he dislikes being blamed; i's preferences are represented by $u_i^{GB}(\mathrm{z},\alpha_{\text{-}i},\beta_{\text{-}i},\gamma_{\text{-}i}) = \mathrm{m}_i(\mathrm{z}) - \sum_{\mathrm{j}\neq \mathrm{i}} \theta_{ij} \mathrm{E}_{\alpha_j,\beta_j,\gamma_j}[G^0_{ij} \mid H_j(\mathrm{z})]$, where $\theta_{ij} \geq 0$ are exogenously given parameters reflecting i's guilt-from-blame sensitivity with respect to j.[26] Append the functions $(u_i^{GB})_{i \in N}$ to a given extensive game form to obtain a psychological game with guilt-from-blame.

An assessment is a profile $(\sigma,\alpha,\beta,...) = (\sigma_i,\alpha_i,\beta_i,...)_{i \in N}$ specifying behavioral strategies, first- and higher-order beliefs. Assessment $(\sigma,\alpha,\beta,...)$ is consistent if there is a strictly positive sequence $\sigma^k \to \sigma$ such that for all $i \in N$, $h \in H_i$, $s_{-i} \in S_{-i}(h)$,

[26] Since the players need not know which strategy profile is played, or the co-players' beliefs, and may not even observe z, u_i^{GB} does not represent a utility 'experienced' by i. What is assumed is that, given his beliefs (up to the fourth order), i tries to make the expected value of u_i^{GB} as large as possible.

$\alpha_i(s_{-i} \mid h) = \lim_{k\to\infty} \frac{\Pr_{\theta_c}(s_c)\prod_{j\neq i}\Pr_{\theta_j^k}(s_j)}{\sum_{s'_{-i}\in S_{-i}(h)}\Pr_{\theta_c}(s'_c)\prod_{j\neq i}\Pr_{\theta_j^k}(s'_j)}$ and higher-order beliefs at each information set are correct: for all $i \in N$, $h \in H_i$, s_{-i}, $\beta_i(h) = \alpha_{-i}$, $\gamma_i(h) = \beta_{-i}$, $\delta_{-i}(h) = \gamma_{-i}$, and so on.

Definition 3: Fix a profile of utility functions $(u_i^{GB})_{i\in N}$. A consistent assessment $(\sigma,\alpha,\beta,...)$ is a sequential equilibrium (SE) if for all $i \in N$, $h \in H_i$, and $s_i \in S_i(h)$ we have $\Pr_{\sigma_i}(s_i \mid h) > 0 \Rightarrow s_i \in \text{argmax}_{s'_i\in S_i(h)} \mathrm{E}_{s'_i,\alpha,\beta,...}[u_i^{GB} \mid h]$.[27]

Definition 3, when applied to our specific games, implies the predictions given by Definitions 1 & 2 in the main text. There we assume that $\theta_{ij} = 0$ except when i = low-talent B and j = A (in which case $\theta_{ij} = \theta \geq 0$).

[27] If the payoff functions depend only on the end node, this definition is equivalent to the classical SE notion for standard games. Adapting an existence proof from Battigalli & Dufwenberg (2009), one can show that every psychological game with guilt-from-blame has a SE.

Figure 1: The (5,7)-game

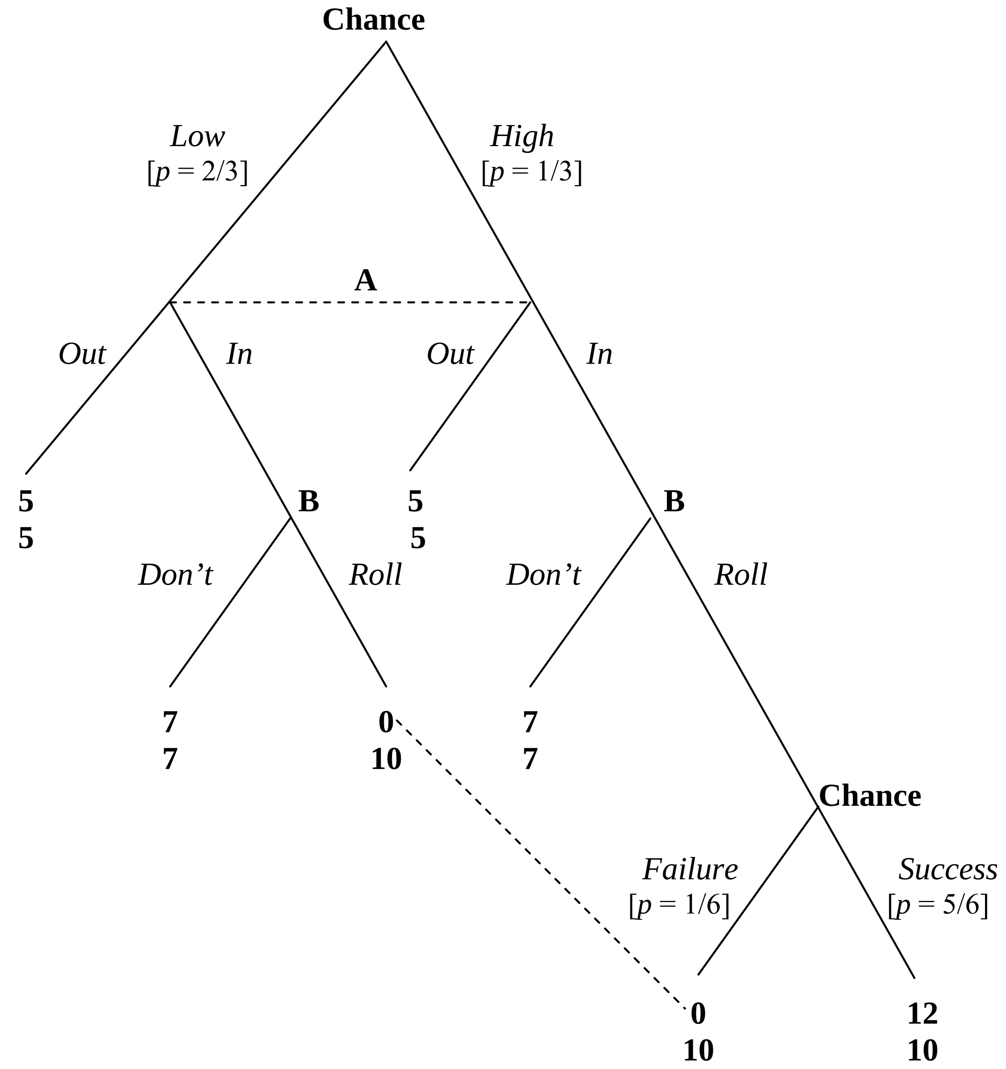


Figure 2: The (5,5)-game

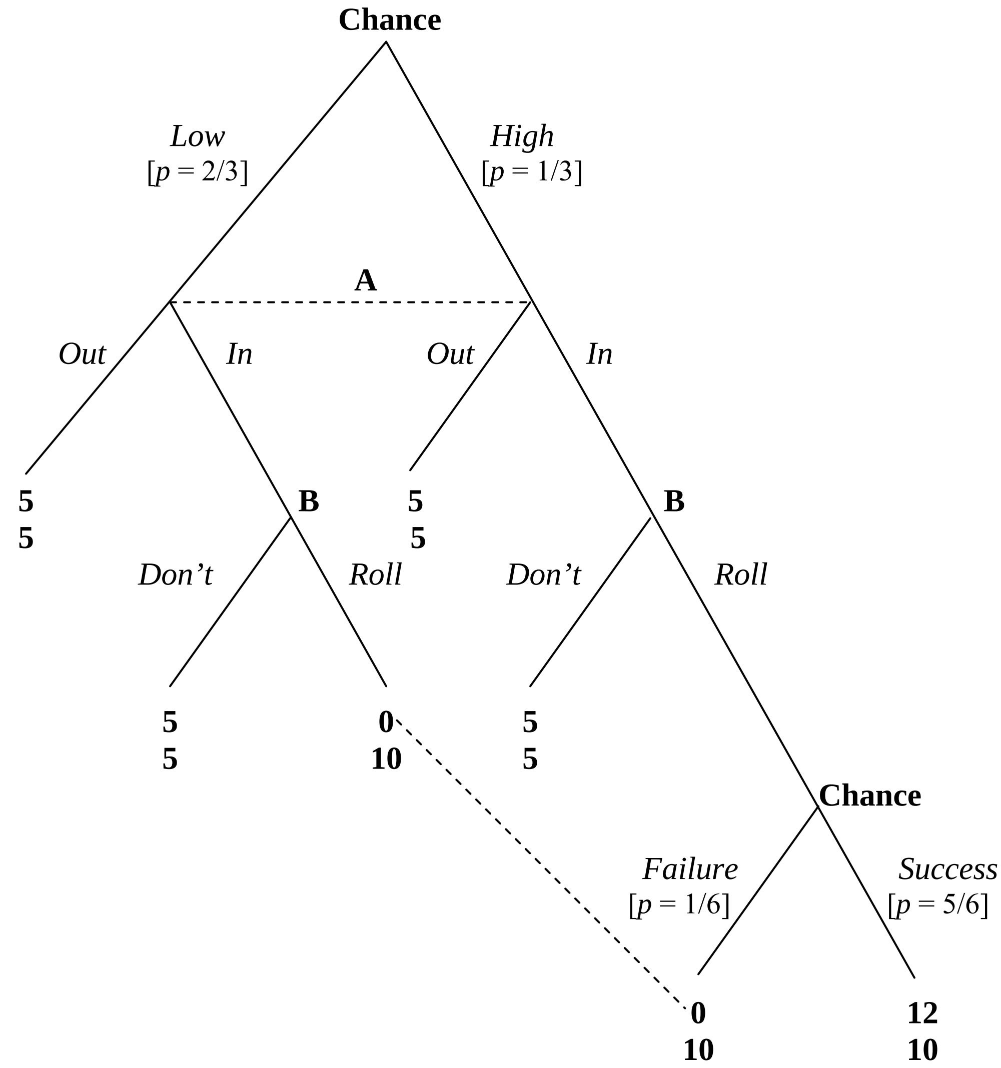
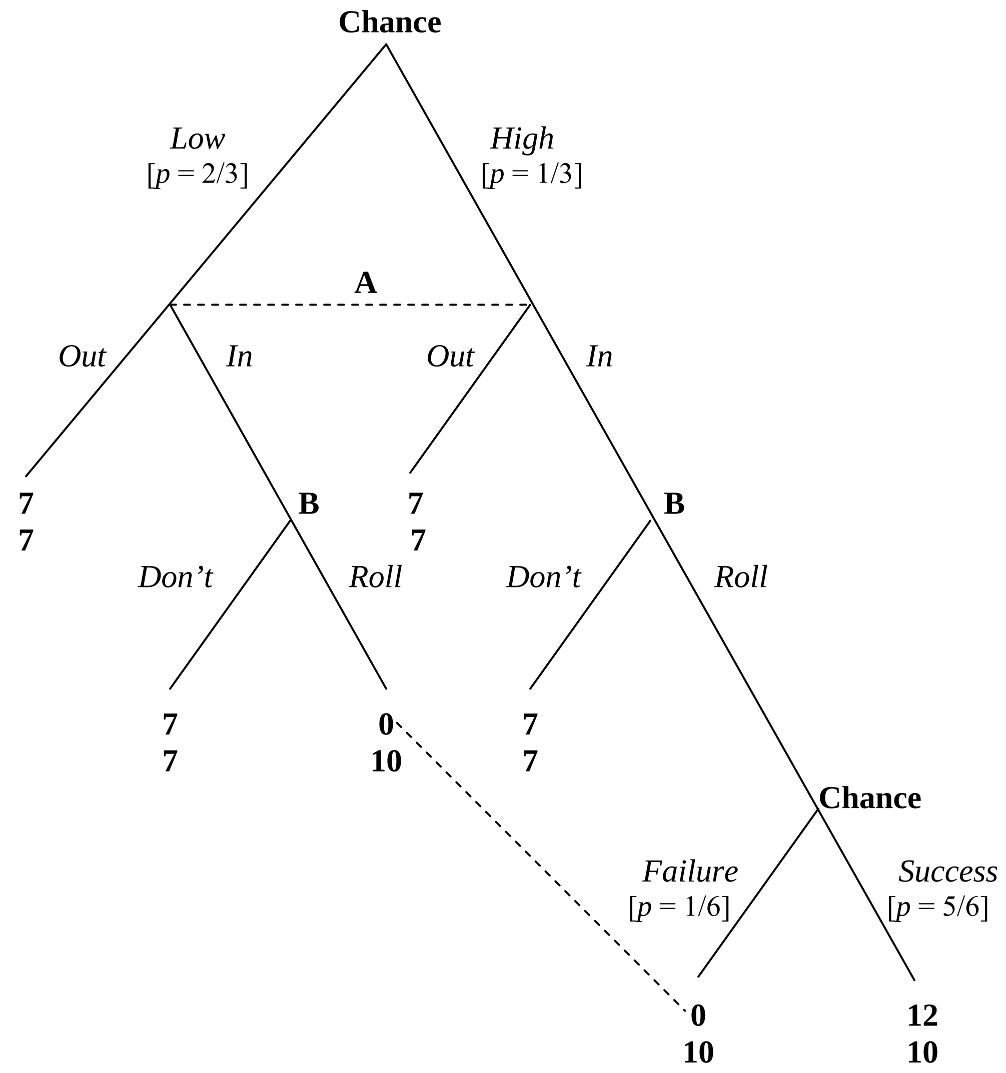


Figure 3: The (7,7)-game


Chance
Low
[p = 2/3]
High
[p = 1/3]
A
Out
In
Out
In
7
7
B
7
7
B
Don’t
Roll
Don’t
Roll
7
7
0
10
7
7
Chance
Failure
[p = 1/6]
Success
[p = 5/6]
0
10
12
10