Honesty and informal agreements ☆

Martin Dufwenberg [a,b,c], Maroš Servátka [d], Radovan Vadovič [e,*]

[a] *University of Arizona, United States*
[b] *University of Gothenburg, Sweden*
[c] *CESifo, Germany*
[d] *MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, Australia*
[e] *Carleton University, Canada*

A R T I C L E   I N F O

A B S T R A C T

We develop, and experimentally test, models of informal agreements. Agents are assumed to be honest but suffer costs of overcoming temptations. We extend two classical bargaining solutions – split-the-difference and deal-me-out – to this informal agreement setting. For each solution there are two natural ways to do this, leaving us with 2×2 models to explore. In the experiment, a temptations-constrained version of deal-me-out emerges as the clear winner.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Traditional bargaining theory, *e.g.*, Nash (1950, 1953), focuses on binding contracts. Much less attention has been given to informal (non-binding) agreements. A likely reason is that if people maximize own income, a common assumption, then there is limited scope for informal agreements to have impact.[1] A selfish agent would simply renege if this were in his interest.

Humans have tendencies that curb such opportunism. Successful entrepreneur Karl Eller, for example, wrote his book *Integrity Is All You've Got* (2004) in which that message is clear. One can justify honesty with reference to repetition or reputation, but that cannot be the whole story. Eller writes about "the happiness that comes with knowing you'll never be ashamed to face yourself in the mirror" (p. 103). Indeed, experiments indicate that honesty matters even in non-repeat settings with anonymity guaranteed. For example, Malhotra and Murnighan (2002), Irlenbusch (2004), Ben-Ner and Putterman (2009),

and Kessler and Leider (2012) find that subjects who were offered an opportunity to enter an informal agreement often did so and then delivered although they could have profitably reneged.[2]

Honest individuals have much to gain by striking informal agreements. Binding contracts may be infeasible (*e.g.*, in developing countries with unreliable courts), illegal (*e.g.*, for cartelists), or costly (*e.g.*, nuptials). This begs questions regarding the shape and impact of informal agreements when honest folks interact. We develop, and then experimentally test, relevant theory. Throughout, we assume that individuals are completely honest in the sense that once they have struck an agreement they never renege. This is stark but once honesty is acknowledged to a degree, understanding the implications of complete honesty seems like a natural benchmark, so we start there.

For a variety of psychological reasons (discussed in section 2.2) honesty may (one way or another) be driven by psychic costs associated with breaking a promise or reneging. Such costs can only be incurred if the post-agreement behavior slips *off-the-agreement-path*. For honest individuals, this ensures adherence. Because the cost is counterfactual (never occurring) there is little reason to think its magnitude would affect how lucrative an informal agreement seems to an honest party. It is thus natural to wonder whether honesty implies that informal agreements work just like binding contracts.

A key insight of our analysis is that this is not necessarily the case. A binding contract differs from an informal agreement in that the latter involves materially profitable post-agreement reneging opportunities. Much evidence suggests that people generally struggle with resisting temptations.[3] Although an honest person fulfills the terms he agrees to, he need not be immune to the costs incurred when overcoming temptations to renege. We think of honest persons as having flashes of temptation such that they temporarily consider disregarding the off-path reneging costs and indulging the luring opportunistic gains that loom, even if ultimately the off-path costs win them over so that they never renege. These temptations occur *on-the-agreement-path* and affect the evaluation of informal agreements, including which ones are worth striking. This effect has no counterpart if a binding contract is considered, because binding contracts come with explicit material, rather than psychological and for a tempted party possibly oblivious, sanctions that make reneging not seem tempting.

We work with the following framework: (i) an informal agreement may be reached by two players about to play what we will call an "underlying game;" (ii) the object of negotiation concerns which strategy profile to play; (iii) whether or not an informal agreement is in place, no material sanctions punish off-path play. We extend two classical bargaining solutions – split-the-difference and deal-me-out – to this setting. With temptation costs in the picture, there are two seemingly natural ways to do this; beyond affecting participation constraints, temptation costs may or may not affect the predicted deals themselves. This leaves us with 2×2 models to explore. Our experiment may be seen as a horse race between these models.

Anchoring our analysis on an underlying game allows us to be explicit about the nature of the economic situation in which a deal is struck, and it allows informal agreements to be truly non-binding since post-agreement that game must still be played. The underlying game does not describe the haggling process. Agreement-formation is instead captured implicitly, through a solution-concept and predictions are formulated in terms of restrictions on strategies that players agree on. Although we focus on informal agreements, as will be seen later the approach is not limited to such contexts; it allows for analyzing and comparing binding contracts as well.

Our paper shares the focus on informal agreements in underlying game with the pioneering work by Miettinen (2006, 2013). He examines which deals players will honor if they have "costs of breaking agreements" (and he derives results that hinge on whether or not actions are strategic complements). Miettinen thus takes exogenously given informal agreements and asks if they will be honored, whereas we assume that informal agreements are honored and then offer theories that endogenize their shape.

Our contribution has two parts. #1 is formulating theory. #2 is testing that theory in an experiment. Sections 2, 3, and 4 present, respectively, theory, experiment, and concluding remarks including suggestions for follow-up research.

2. Theory

2.1. General framework

Our starting point is a two-player extensive game (form) Γ with dollar payoffs specified at end nodes. Let S_i be player i's set of strategies (taken to be singleton if i owns no information set), and $S = S_1 \times S_2$. Let $m_i : S \to \mathbb{R}$ be i's (dollar) payoff-from-strategy-profile function, derived from Γ.

This underlying game Γ describes the strategic structure of a situation where two persons just met and face opportunities of collaboration for mutual gain. Assume that Γ is a multi-stage game form with observed actions (including at endnodes),[4] so that all instances of imperfect information concern simultaneous choices. This simplifies the key definitions

---

[2] We view informal agreements, which involve a form of consensus, as conceptually distinct from promises, which may be unilateral. Several studies relatedly document a preference to keep one's word, *e.g.*, Kerr and Kauffman-Gilliland (1994), Ellingsen and Johannesson (2004), Gneezy (2005), Charness and Dufwenberg (2006), Vanberg (2008), and Servátka et al. (2011).

[3] See Benartzi and Thaler (2004), Brown et al. (2009), and Martinsson et al. (2012) for experiments, Thaler and Shefrin (1981), Gul and Pesendorfer (2001), Loewenstein and O'Donoghue (2005), Fudenberg and Levine (2006, 2011, 2012), Ozdenoren et al. (2012), and Lipman and Pesendorfer (2013) for theory.

[4] See, *e.g.*, Fudenberg and Tirole (1991), chapter 3. The assumption of perfect information across end nodes is important for our upcoming comparison between informal agreements and binding contract; see section 2.3.
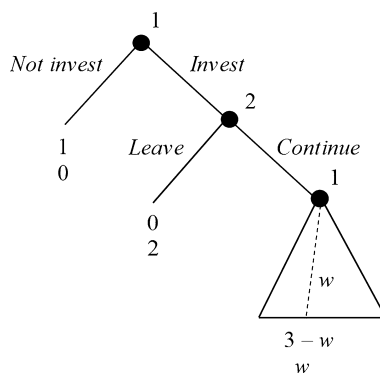
Fig. 1. Hospital-doctor game.

below by allowing us to refer to subgames in a useful way, without essentially compromising scope since most applied and experimental work is concerned with such games. The payoffs represent dollar increments relative to whatever wealth the players had before; a payoff of 0 means that a player's overall dollar wealth remains the same as if he had never met the other player.

Many economic situations involve payments so it is natural to consider underlying games reflecting that. For example, let $H > 0$ be the highest sum of the players' payoffs at any endnode. The games we focus on have the property that if $\$_1, \$_2 \geq 0$ and $\$_1 + \$_2 = H$ then Γ admits some endnode with payoffs $(\$_1, \$_2)$, thus allowing ways to equalize gains. However, it is not inconceivable that an economic situation somehow significantly constrains players' transfer opportunities away from allowing equal splits. Our definitions are intended to apply regardless and therefore stated without presumptions of transfer possibilities.

We envision the players as haggling over which strategy profile in Γ to play. Γ itself does not describe this process which is rather captured via a solution concept with a special structure: We select a triple of strategy profiles $a, b, c \in S$ such that a is the agreement, b is the behavior following the agreement, and c is what counterfactually would have happened if negotiations stranded. Predictions are formulated as restrictions on a, b, and c. Before we describe these, several clarifying comments are warranted:

First, since $a, b, c \in S$ are strategy profiles they describe off-path play which we interpret as reflecting the players' agreed upon understanding (presumably obtained through the haggling process or social norms) of what would happen following any deviation.[5] Second, we theorize only about what happens when negotiations generate agreements; c describes what would have happened had a not occurred, given that a does occur. It is possible that in some game no triple (a, b, c) exists that satisfies the postulated properties. The interpretation is that no agreement would be reached in that case. We offer no explicit prediction for play following such non-counterfactual negotiation-breakdown. Third, one could imagine a richer structure where c depends on how negotiations stranded (*e.g.*, which player caused the break down). We abstract away from such nuances.

Fourth, we elucidate why we do not explicitly model the strategic structure of pre-play negotiation. Consider Fig. 1, which comes with a story:

Player 1 is a hospital and player 2 an employed doctor. At the root 1 decides whether to *Invest* or *Not invest* in costly training for 2 to learn a new radiography technique. In the former case 2 becomes more productive but also more attractive to other hospitals; choice *Leave* with subsequent payoffs reflects what happens if 2 resigns and takes employment at Johns Hopkins. That would be bad for 1 who stands to gain if 2 instead *Continues* at the current job, in which case 1 can choose what wage $w \in [0, 3]$ to pay 2,[6] thereby affecting 2's life-time income.

This description is meant to appear somewhat realistic, yet it is overly barren as it incorporates no opportunities for haggling, promises, threats, etc. A more meaningful account might incorporate how the players meet and discuss whether 1 should pay for the training and what 2's pension should be. How should one model such considerations? One possibility is to modify the game, to include counter-offers, promises, threats, handshakes, signatures, etc., as explicit choices in a larger game. But such a game is likely to be unwieldy. It may be intractable to apply a solution concept. It is against this backdrop that we formulate our approach.

This connects us to some classics. Von Neumann and Morgenstern (1944) approach all games other than two-player zero-sum ones with this outlook (see *e.g.*, pp. 223–224). Nash (1953) assumes players strike binding contracts regarding how to play an underlying game, and before negotiations start they announce "threats" an "umpire" forces them to implement

---

[5] One could imagine alternatives, *e.g.*, as in a self-confirming equilibrium (Fudenberg and Levine, 1993; cf. Greenberg, 2000).

[6] The implicit assumption is that later in 2's life he has fewer outside opportunities and is therefore vulnerable to hold-up.

if they subsequently fail to reach an agreement.[7] We share the outlook that strategy profiles are objects of negotiation, but neither limit attention to binding contracts nor presuppose access to an umpire.[8]

2.2. Specific predictions

What psychological and economic principles determine the shape of $a, b, c \in S$? The framework of section 2.1 is useful for formulating answers. We develop two specific models that, apart from assuming honesty and temptation costs, connect closely to classical bargaining scholarship. These models differ on the specification of a but share a common specification of b and c. We start with the latter two.

As regards c, counterfactual post-negotiation-breakdown play, we make

Assumption 1. c is a subgame perfect equilibrium of Γ using $(m_i)_{i=1,2}$.

This modeling choice is a compromise. On the one hand, many studies show that players often act pro-socially. In many games (*e.g.,* prisoners' dilemma, public goods, or trust games) that exhibit a tension between individual and collective dollar-payoff-maximization, subjects manage to reach efficient outcomes,[9] suggesting that they appreciate collective well-being. On the other hand, it seems likely that players who do not manage to agree would end up being irritated with one another. Our assumption of selfish behavior takes a middle road.[10]

Next consider b, behavior following an informal agreement. Our assumption, key to everything to follow, is that players honor their agreements:

Assumption 2. $b = a$.

If an informal agreement $a = (a_i)_{i=1,2} \in S$ is struck, then each i subsequently chooses $b_i = a_i$. Such honesty may have multiple psychological foundations, like a preference to keep promises or (more generally) not to have lied (*e.g.,* Gneezy, 2005; Demichelis and Weibull, 2008; Vanberg, 2008; Kartik, 2009), obeying some social norm that one should honor agreements (*e.g.,* Malhotra and Murnighan, 2002; Miettinen, 2006, 2013; Kessler and Leider, 2012), or guilt aversion (*e.g.,* Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007) such that they live up to others' expectations as shaped through negotiations. One may additionally conceive ways to back up honesty via concerns of identity (*e.g.,* Akerlof and Kranton, 2000) or maintenance of self-esteem (*e.g.,* Benabou and Tirole, 2002). In principle it is of interest which story is more empirically relevant, but in this paper we do not aim to unpack the psycho-foundations of honesty. We simply assume that agreements are honored.

Next we turn to a, the informal agreement itself. Unlike binding contracts, which are enforced by explicit material sanctions, adherence to an informal agreement needs support by psychological costs of reneging. To an honest person, such cognitive costs are sufficiently large to prevent reneging. Since they occur *off-path* (as reneging never happens) arguably they shouldn't affect how attractive a deal seems at the point of agreement. However, another subtle difference between binding contracts and informal agreements may occur *on-path*, post-agreement. With an informal agreement a party may face materially lucrative opportunities to renege, which may be tempting if a player temporarily disregards the psychological off-path reneging costs and instead focuses on the opportunistic gain available. We propose that even players who are honest, in the sense that the off-path costs ultimately win them over such that they do not reneg, experience flashes of such temptations. There is a sizable literature on human tendency to resist temptations.[11] It is often argued that humans can overcome temptation, but that this comes at a cost. If honest player i considers such costs when evaluating an informal agreement, then his subjective gain (*i.e.,* net of the temptation cost) under an informal agreement will be lower than with the same strategy profile as a binding contract. These considerations may affect the shape of the agreement. If the temptation costs are big enough, they may even make i accept $a \in S$ as a binding contract but not as an informal agreement.

How should one calculate costs of overcoming temptation? Are they linear or perhaps convex in how much a player may gain (cf. Fudenberg and Levine, 2006, 2011, 2012)? Are they stochastic (cf. Dekel and Lipman, 2012)? Do they depend on how many times a player is tempted (cf. Ozdenoren et al., 2012), or only on the maximum temptation along the path? Are

---

[7] See Kalai (1977) and Kalai and Tauman Kalai (2010) for more work in this vein.

[8] The cheap talk literature (*e.g.,* Crawford and Sobel, 1982; Farrell and Rabin, 1996) also studies the effect of communication in games. Unlike our approach, cheap talk is modeled as explicit choices and, most importantly, presumed not to affect preferences (over strategy profiles) in the underlying game. In our approach players have a preference for playing as they agree, so talk is not cheap. There is also the game-theoretic literature on communication equilibria (*e.g.,* Forges, 1986; Myerson, 1986), which (like us) captures the effect of messages through solution concepts but (like the cheap talk literature) assumes communication does not affect preferences over strategy profiles.

[9] For surveys of the evidence, see *e.g.,* Camerer (2003) or Fehr and Schmidt (2002).

[10] Note that there is scant evidence to guide our modeling choice: existing data on the relevance of social preferences typically concerns play after neither actual nor counterfactual negotiation breakdown.

[11] See the references in footnote 3. The literature focuses on single decision maker settings (Loewenstein and O'Donoghue's section 6 is an exception), not temptation to renege and hurt a co-player, but that extension seems plausible to us. Indeed, Martinsson et al. reports support for "the proposition that individuals may experience a self-control conflict between the temptation to act selfishly and the better judgment to act pro-socially."

they moderated if reneging hurts others (cf. Gneezy, 2005), or via some notion of "empathy" (Loewenstein and O'Donoghue, 2005)? The answers are not obvious. We work with the following formulation. Given an informal agreement $a = (a_i, a_j) \in S$, i's cost of overcoming a temptation associated with a is a real-valued, continuous, strictly increasing, weakly convex function $\gamma_i : S \to \mathbb{R}_+$ defined by $\gamma_i(\max_{s_i \in S_i} m_i(s_i, a_j) - m_i(a))$ such that $\gamma_i(0) = 0$.

Drawing on γ_i, we now define three concepts which are key components and which stay constant across all versions of our third assumption:

Three definitions:

$a' \in S$ allows *strict gains* if $[m_i(a') - \gamma_i(a')] - m_i(c) > 0$ for $i = 1, 2$.

$a' \in S$ is *efficient* if there exists no $a'' \in S$ such that $[m_i(a'') - \gamma_i(a'')] - m_i(c) > [m_i(a') - \gamma_i(a')] - m_i(c)$ for $i = 1, 2$.

$a' \in S$ is *c-based* if off-its-path a' specifies the same choices as c.

To understand strict gains, note that $m_i(a) - \gamma_i(a)$ is the value of the deal net of temptation cost while $m_i(c)$ is the value of the forgone opportunity. The difference $[m_i(a') - \gamma_i(a')] - m_i(c)$ may be interpreted as i's subjective gain-from-trade, which we shall require to be strictly positive since it seems plausible that people agree only to deals where they improve. The efficiency requirement is analogous to what is typically assumed in theories of binding contracts, and seems equally plausible in our context. The c-based requirement is an assumption about how the play proceeds after a player reneges. The motivation is analogous to what we proposed for c (including that c is independent of how negotiations stranded or the nature of reneging).

With this in place we state four competing versions of our third and final assumption. Each predicts that a particular equal split of something valuable will occur, if this is possible while respecting strict gains, efficiency, and being c-based. The versions differ on exactly what is being split. We draw on classical bargaining scholarship and extend two models to our setting: *split-the-difference* and *deal-me-out*. Under the former the value split is measured relative to the parties "threat points" $m_i(c)$ for each i.[12] Under the latter the value is measured without regard to $m_i(c)$.[13] We consider versions where the value split reflects, or does not reflect, temptation costs.

We will offer further comments, but doing so is easier if we can refer to the definitions. These are (with *ESIP* mnemonic for *equal-split-if-possible*):

Assumption 3. Version (i): $a \in S$ allows strict gains, is efficient, and is c-based. Moreover, it satisfies the following ESIP(i) condition: Let $E^{(i)}$ be a set of all $a' \in S$ that allow strict gains, are efficient, are c-based, and satisfy $[m_i(a') - \gamma_i(a')] - m_i(c) = [m_j(a') - \gamma_j(a')] - m_j(c)$ for $i, j = 1, 2$. If $E^{(i)}$ is nonempty, then $a \in E^{(i)}$.

Version (ii): $a \in S$ allows strict gains, is efficient, and is c-based. Moreover, it satisfies the following ESIP(ii) condition: Let $E^{(ii)}$ be a set of all $a' \in S$ that allow strict gains, are efficient, are c-based, and satisfy $m_i(a') - \gamma_i(a') = m_j(a') - \gamma_j(a')$ for $i, j = 1, 2$. If $E^{(ii)}$ is nonempty, then $a \in E^{(ii)}$.

Version (iii): $a \in S$ allows strict gains, is efficient, and is c-based. Moreover, it satisfies the following ESIP(iii) condition: Let $E^{(iii)}$ be a set of all $a' \in S$ that allow strict gains, are efficient, are c-based, and satisfy $m_i(a') - m_i(c) = m_j(a') - m_j(c)$ for $i, j = 1, 2$. If $E^{(iii)}$ is nonempty, then $a \in E^{(iii)}$.

Version (iv): $a \in S$ allows strict gains, is efficient, and is c-based. Moreover, it satisfies the following ESIP(iv) condition: Let $E^{(iv)}$ be a set of all $a' \in S$ that allow strict gains, are efficient, are c-based, and satisfy $m_i(a') = m_j(a')$ for $i, j = 1, 2$. If $E^{(iv)}$ is nonempty, then $a \in E^{(iv)}$.

These solutions depend on the γ_i's, as if those functions were known. In bargaining theory, assuming players to have common knowledge of features of one another's preferences (*e.g.*, discount rates) is not unusual. How compelling is it to extend the idea to temptation costs? The assumption may be plausible for people who know each other very well. But perhaps more importantly, since our players are honest, it seems reasonable to assume that they do not pretend to have different γ_i.

Our definitions reflect two distinct ways to operationalize that idea. Assumption 3 versions (i) & (ii) involve default deals that factor in temptation costs directly. During the course of negotiations players' offers and actions reveal their γ_i's to one another! This is stark. Sticking to an agreement may be one thing, revealing private information quite another.[14] Nevertheless, the assumption is consistent with (extreme) honesty, a testable benchmark worth considering. Assumption 3 versions (iii) & (iv), by contrast, involve default deals that refer to material rewards only (again, with or without $m_i(c)$ in

---

[12] When selfish, risk-neutral players divide money, a number of bargaining solutions make the same prediction as split-the-difference, including the Nash (1950) and Kalai and Smorodinsky (1975) solutions.

[13] Several experiments tested split-the-difference vs. deal-me-out models in various bargaining scenarios (e.g., Hoffman and Spitzer, 1982; Binmore et al., 1998; Feltovich and Swierzbinski, 2011; Anbarci and Feltovich, 2013). It seems subjects are largely less sensitive to nonbinding disagreement outcomes than predicted by split-the-difference. In many instances they simply split the pie down-the-middle.

[14] In addition, in line with Loewenstein's (1996) contention that "people underestimate the impact of visceral factors on their own future behavior" (his Proposition 5), we note that assuming that i knows γ_i is not entirely innocuous.
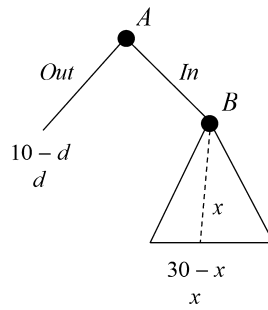
Fig. 2. Lost wallet game.

the picture). Temptation costs now matter only through the strict gains and efficiency conditions. Players merely say "yes" or "no" to the default deals as given by ESIP(iii) or ESIP(iv).

For many underlying games, if the default scenario with equal-gains/shares obtains, the prediction is unique. However, if equal-gains/shares is incompatible with strict gains, efficiency, and being c-based, then any deal that satisfies the latter three restrictions may obtain.[15] Existence is not guaranteed, as it may be impossible to satisfy those requirements. If the temptations that weigh on the players become very strong, both may demand "compensation" (relative to equal-gains/shares) beyond what's feasible. We illustrate these possibilities in section 3.

2.3. Binding contracts

Our main interest concerns informal agreements but our framework admits the case of binding contracts as a modified case. To cover that, assume that, independent of whether a party is honest, deals are honored because high explicit sanctions would punish breach of contract. Players are then not tempted to renege. Hence, binding contracts may be studied by applying Assumptions 1, 2, and 3($\cdot$), under the new assumption that $\gamma_i(s) = 0$ for all $s \in S$, $i \in N$.

This interpretation presumes there are no issues of moral hazard, such that some choices would be non-verifiable to a contracting party (or a court) and a binding contract regulating that choice infeasible. This is justified through our assumption of observed actions (also at endnodes; cf. section 2.1). A large contract-theoretic literature explores moral hazard – see *e.g.,* Bolton and Dewatripont (2005) for an entry. We restrict attention to underlying games where the issue is irrelevant, in order to highlight differences between informal agreements and binding contracts other than feasibility.

3. An experiment

We proposed four extensions of classical bargaining theory to informal agreements. Are they empirically relevant? We designed an experiment to shed light on the issue.

3.1. Experimental games and predictions

We use the lost wallet game (Dufwenberg and Gneezy, 2000), presented in Fig. 2, where $d \in \{0, 5\}$ is a parameter which varies by treatment. The game presents several advantages: It is easy to explain to subjects and implement, yet rich enough to allow a deal with equal payoffs. The theory generates sharp comparative statics predictions across our four treatments (discussed shortly). As regards $c \in S$, there is a unique subgame perfect equilibrium using $(m_i)_{i=1,2}$, namely $(s_1, s_2) = (s_1, x) = (Out, 30)$. Our solutions predict that only player B may face a temptation to renege, which simplifies the analysis (relative to, say, the hospital-doctor game of section 2.1).

The assumptions of section 2.2 imply that $c = (Out, 30)$ and $b = a$, but rule out $a = (Out, x)$ for any x. Too see this, note that the profile is not c-based unless $x = 30$, in which case strict gains fails for A. Consider instead $a = (In, x)$. Strict gains for A implies $(30 - x) - \gamma_A(a) - (10 - d) > 0$ implying $(30 - x) > (10 - d)$ implying $\gamma_A(a) = 0$. For B, strict gains implies that $x - \gamma_B(30 - x) - d > 0$. Combining inequalities, we get the *strict-gains constraint*: $20 + d > x > d + \gamma_B(30 - x)$.

Next consider Assumption 3. Under 3(i), ESIP(i) implies that if possible $x = x^{(i)}$ satisfies $(30 - x^{(i)}) - 0 - (10 - d) = x^{(i)} - \gamma_B(30 - x^{(i)}) - d$, or equivalently $x^{(i)} = 10 + d + \frac{1}{2}\gamma_B(30 - x^{(i)})$. Under 3(ii), ESIP(ii) implies that if possible $x = x^{(ii)}$ satisfies $(30 - x^{(ii)}) - 0 = x^{(ii)} - \gamma_B(30 - x^{(ii)})$, or equivalently $x^{(ii)} = 15 + \frac{1}{2}\gamma_B(30 - x^{(ii)})$. Since γ_B is strictly increasing and continuous each of the two equations has a unique solution. However, in each case, the solution is not guaranteed to always satisfy the strict-gains constraint. If it is satisfied, then the corresponding set $E^{(j)}$, $j \in \{i, ii\}$, is nonempty and we have found our solution: agreement $a = (In, x^{(j)})$.[16] Otherwise, $E^{(j)}$ is empty. Then, using the strict-gains constraint, we get a solution

---

[15] This differs from Binmore et al. (1989), where a departure from the equal split gives a constrained player exactly what makes him indifferent between agreeing or not, hence the solution is unique but there is no strict gain which we find unintuitive.

[16] It is easy to verify that a is c-based and efficient.

Table 1
Predictions: *x*.

Binding contract (BC): $x^{(j)} =$				
Assumption 3: $(j) =$	(*i*)	(*ii*)	(*iii*)	(*iv*)
$d = 0$	10	15	10	15
$d = 5$	15	15	15	15

Informal agreement (IA): $E^{(j)}$ is nonempty, $x^{(j)} =$				
Assumption 3: $(j) =$	(*i*)	(*ii*)	(*iii*)	(*iv*)
$d = 0$	$10 + \gamma_B(30 - x^{(i)})/2$	$15 + \gamma_B(30 - x^{(ii)})/2$	10	15
$d = 5$	$15 + \gamma_B(30 - x^{(i)})/2$	$15 + \gamma_B(30 - x^{(ii)})/2$	15	15

Informal agreement (IA): $E^{(j)}$ is empty, $x^{(j)} \in$	
Assumption 3: $(j) =$	(*i*) & (*ii*) & (*iii*) & (*iv*)
$d = 0$	$(\gamma_B(30 - x^{(j)}), 20)$
$d = 5$	$(5 + \gamma_B(30 - x^{(j)}), 25)$

set: $\{x \mid 20 + d > x > d + \gamma_B(30 - x)\}$. If γ_B is so "steep" that $\gamma_B(30 - x) \geq 20$ for all $x < 20 + d$ then this set is also empty, illustrating the potential for non-existence.[17] Otherwise, all the elements are *c*-based and efficient, so all the corresponding $a = (In, x)$ satisfy our assumptions.

Now move to Assumption 3 versions (iii) and (iv). ESIP(iii) implies that if possible $x = x^{(iii)}$ satisfies $(30 - x^{(iii)}) - (10 - d) = x^{(iii)} - d$, or $x^{(iii)} = 10$; ESIP(iv) implies that if possible $x = x^{(iv)}$ satisfies $(30 - x^{(iv)}) = x^{(iv)}$, or $x^{(iv)} = 15$. Claims that follow are parallel to those in the preceding paragraph. If any of the solutions satisfies the strict-gains constraint, then the corresponding set $E^{(j)}$, $j \in \{iii, iv\}$, is nonempty and we have found our solution: agreement $a = (In, x^{(j)})$. Otherwise, all elements of the solution set $\{x | 20 + d > x > d + \gamma_B(30 - x)\}$ are informal agreements; or, if the set is empty, there is no agreement.

For binding contracts, A and B agree on $a, b, c \in S$ according to the assumptions of section 2.3. ESIP(i) and ESIP(ii) collapse to ESIP(iii) and ESIP(iv), respectively. Moreover, the binding contract version of the strict-gains constraint $20 + d > x^{(j)} > d$ always holds ($j \in \{i, ii, iii, iv\}$).

The following tables summarize the predictions for the cases when the agreement exists. An agreement always involves A choosing *In*. B keeps an amount x that corresponds to the appropriate case as shown.

The experiment uses a 2×2 between-subjects design that closely follows the discussion above.[18] Between treatments we varied payoff parameter $d \in \{0, 5\}$ and whether the agreement was informal (IA) or a binding contract (BC). The four treatments are labeled: BC-0, BC-5, IA-0, and IA-5. At a later date we added two control treatments NN-0 and NN-5 that involved no pre-play negotiations.[19]

For binding contracts we obtain point predictions. Each pair of subjects agrees that A chooses *In* and B keeps the amount as specified in the top panel of Table 1. For informal agreements the predictions are summarized in the middle and the lower panels of Table 1. Again, every agreement involves A choosing *In*. However, the distribution of x's will depend on the (unobservable) distribution of γ_B's in the subject population. To keep the discussion organized we separately discuss two cases: one, where γ_B is rather flat so that predictions are contained within the middle panel of Table 1, and another, in which γ_B is sufficiently steep so that the predictions fall within the lower panel of the table.

First, consider that case when γ_B is not very steep. According to Assumption 3(i) x's will be distributed on the support bounded by 10 and 20 in the IA-0 and by 15 and 25 in the IA-5 treatment. Moreover, the distribution in IA-5 should stochastically dominate that in IA-0. This follows directly from comparing the two rows ($d = 0$ vs. $d = 5$) in Assumption 3(i) of the middle portion of Table 1. Under Assumption 3(ii) the distribution in IA-0 is bounded by 15 and 20 in IA-0 and by 15 and 25 in IA-5. Where the two supports overlap (on 15 to 20) the conditional distributions should be the same. This is clear from the column two of the middle panel of Table 1. The remaining two versions of the Assumption 3, versions (iii) and (iv), make exact point predictions, so we should observe data concentrated on the values as shown in the table.

The second case refers to a γ_B that is sufficiently steep that the equal-split-of-gains is incompatible with strict gains, but not too steep as otherwise no agreement would be feasible. In this case the prediction is set-valued: A and B are predicted to agree on some x belonging to the set given in the bottom panel of Table 1. From the table it is clear that as we move from IA-5 to IA-0 the boundaries shift to the right (by 5). Hence, the strict-gains constraint is more likely to bind in IA-5 than IA-0. This has identifiable implications for the theory under Assumption 3 versions (iii) and (iv). Notice that in these

---

[17] Reflect on the intuition: B is so easily tempted that the compensation he craves, relative to equal-split-of-gains, is incompatible with strict gains for A.

[18] We restricted x to be a whole \$-amounts: $x \in \{0, 1, ..., 30\}$. There is some issue when the predicted x is not an integer, but any finer scale for x would not fix that and bring more complexity to the experiment. The theory of section 3 is thus taken to make an approximate prediction for our finely discretized lab implementation.

[19] The role of these treatments is not in testing our theory of how agreements form but rather in evaluating the impact of agreements on eventual outcomes. The discussion of the data from these treatments is deferred until section 3.4.

two cases the theory makes point predictions *unless* (some) subjects have high enough γ_B's. For these instances we expect the agreed on *x*'s to compensate (favor) players *B*.

Finally, if the subject's γ_B is so steep that the if-equal-split-not-possible set is empty, then there is simply no room for players to agree. This would happen when $\gamma_B(30 - x) < 20$ for all $x < 20 + d$. Since the condition is tighter in the $d = 0$ case, if there are any disagreements we would expect them to be more likely in IA-0 than in IA-5.

3.2. Procedures

The experiment was computerized and conducted at the University of Arizona's Economic Science Laboratory. The software was written in Visual Basic 6. In total, 308 undergraduate students participated as subjects. The sessions and participation is summarized in Table 9 in Appendix. Subjects played one game – no repetitions – and were then privately paid. The average final payment was $19.17, including a $5 show-up fee. On average, sessions lasted about 50 minutes.[20]

Once all subjects were seated at computer terminals separated by privacy dividers, hard copies of instructions were handed out (see Appendix) and subjects were given 10–15 minutes to read them. When everyone had finished reading, the instructions were also read out loud. After this, the experimenters answered any questions individually. The software then started up with a set of comprehension questions. Every subject had to get all answers correct before the experiment proceeded further.

Our theory presumes pre-play negotiation but leaves the strategic details of this process implicit, reflected only through the solution concept. In the lab, however, one has to offer some specific format for the haggling. We chose an alternating-offer structure. After being acquainted with game details, and learning their respective roles as player *A* or *B*, subjects could send proposals back and forth and agree on how to play. One person from each pair was randomly selected to make an opening proposal. Each proposal specified whether player *A* would choose *In* or *Out*, and, conditional on *In*, the amount that player *B* would keep. The party who received a proposal could accept it, make a counter-proposal, or disagree and quit negotiating. Acceptance of a proposal led to an agreement. This ended the negotiations and a message saying either "Player *A* chooses OUT" or, *e.g.*, "Player *A* chooses IN and Player *B* keeps $18 and gives $12 to Player *A*" appeared on the pair's computer screens. A counter-proposal reversed the negotiation roles while a disagreement terminated the negotiation process. There was no limit imposed on the length of negotiations or on the time within which a subject had to submit his decision.

3.3. Main results

In what follows *y* is the agreed-upon *x*, *z* is the post-agreement choice of *x*. Table 2 presents raw data on negotiated agreements and paths of play.[21] In the IA-treatments, the agreement/path of play is described by the amount for player *B* ($= y, z$), implying that *A* chooses *In*, or by indicating that the choice for *A* is *Out* (and hence that *B* had no decision to make). In the BC-treatments, $z = y$ by definition/design.

Agreement formation

Table 2 shows that apart from two cases in BC-5 all other pairs of subjects reached an agreement. All BC's involved player *A* choosing *In*. In the IA-treatments 100% of our subject-pairs formed an agreement. From the vantage point of the theory, this would suggest that the subjects' γ_B's are not high enough to induce disagreements. All but one agreement involved player *A* choosing *In*.[22] These patterns can, largely, be accommodated by all models.

Do players honor agreements?

Table 3, distilled from Table 2, provides a summary of reached agreements and subsequent behavior. The first column, *Obs*, denotes the number of subject pairs who participated in a given treatment. The second column, *Agr*, provides the count of reached agreements which we further split (in subsequent columns) into what these agreements prescribe that players *A* and *B* do. For *A*'s we compare the number of subjects who agreed on *In* (see column "Agreed to *In*") with those who agreed and chose *In* (column "Chose *In*"). For example, in IA-0, twenty-four *A*'s agreed to choose *In* and twenty-three of them subsequently honored that agreement. In IA-5 all twenty-six *A*'s agreed and chose *In*.[23] Finally, in the rightmost part of the table we list the number of *B*'s for whom the amount kept (*z*) was smaller than, equal, or greater than the agreed upon amount (*y*).[24]

---

[20] This is from the time of arrival until the last subject was paid out and left. The actual experiment (reading instructions, decisions and questionnaire) took about 30 minutes.

[21] Appendix contains descriptive statistics (Table 10).

[22] One pair in IA-5 agreed on *Out*, then *A* chose *In* and *B* kept 15.

[23] Observation 27 involves the subjects who agreed on *Out*, then chose *In* and 15.

[24] In the BC-treatments, presented in the two bottom rows, by definition there is no variation between the agreement and the observed behavior of either player.

Table 2
Raw data on agreements and path of play.

IA-0		IA-5		BC-0	BC-5	NN-0	NN-5
y	z	y	z	y = z	y = z	z	z
0	15	15	15	14	14	15	15
10	20	15	15	15	15	15	15
13	18	15	15	15	15	15	15
15	15	15	15	15	15	15	15
15	15	15	15	15	15	15	15
15	15	15	15	15	15	15	15
15	15	15	15	15	15	15	20
15	15	15	15	15	15	15	20
15	15	15	15	15	15	15	20
15	15	15	16	15	15	19	20
15	15	15	17	15	15	20	20
15	15	15	25	15	15	20	20
15	15	15	20	15	15	23	20
15	15	15	30	15	15	28	20
15	15	15	30	15	15	30	21
15	15	15	30	15	15	30	25
15	15	16	16	15	15	30	25
15	15	16	20	15	15	Out	25
15	20	17	15	15	16	Out	25
15	30	18	18	15	17	Out	30
15	30	18	18	15	20	Out	30
15	Out	20	20	20	20	Out	30
		20	20		20	Out	Out
		24	22		Disagr.	Out	Out
		Out	15		Disagr.		

Note: *y* refers to the agreed-upon amount that player *B* would keep and *z* to the amount *B* actually kept. Observations that differ from 15 are italicized. In BC-5 following the disagreement both *A* players chose *In* and their matched *B* players chose to keep 20 and 30 respectively.

Table 3
Agreements and honesty.

Treatment	Observations	Agreements	Player A		Player B		
			Agreed to *In*	Chose *In*	z < y	z = y	z > y
IA-0	24	24	24	23	0	17	6
IA-5	27	27	26	27	2	16	8
BC-0	24	24		24		24	
BC-5	27	25		25		25	

Note: In IA-5 one pair has agreed on player *A* choosing *Out*. Following this, player *A* chose *In* and player *B* kept 15.

Table 3 shows that a majority of agreements were honored. In all cases where *A* agreed to choose *In*, that choice was subsequently made. *B*'s, unlike *A*'s, cost themselves a lot of money (usually $15) by not reneging. Nevertheless, the proportion of *B*'s who honor the agreement is rather high. In IA-0 74% of *B*'s did exactly what they agreed to. In IA-5 this proportion was slightly lower at 64%. Out of the sixteen subjects who reneged (about 10% of *B*'s), five kept everything (31% of those who reneged) while the remaining eleven gave their paired player *A*'s a non-zero amount.[25] While our assumption of honesty does not garner universal support, it approximates the data reasonably well.

Equal splits & compensated deals: do temptation costs matter?

As regards agreements formation and honesty, the aspects of the data discussed so far match up with the theory rather well. In this section we proceed to test the implications of the various versions of Assumption 3 on the distribution of *y*'s across treatments, performing several associated tests.

Table 4 presents the data in the condensed form. In all treatments we find substantial concentrations of observations on equal splits ($y = 15$). It is useful to categorize the data with respect to equal-splits vs. other agreements. We break up the data into three separate blocks: first we only present BC treatments, then IA-treatments including all observations, and then we show data for IA-treatments restricted to only subjects who did what they agreed on.

---

[25] We discuss these 'selfish' and 'semi-honest' subjects further in section 3.5.

Table 4
Agreements.

Treatment	$y < 15$	$y = 15$	$y > 15$
BC-treatments			
BC-0	1	22	1
BC-5	1	20	4
IA-treatments: All data			
IA-0	3	21	0
IA-5	0	18	8
IA-treatments: Honest subjescts ($y = z$)			
IA-0	0	17	0
IA-5	0	11	5

In the BC-treatments (see the top panel of Table 4), high frequencies of equal-splits support versions (ii) and (iv) of the Assumption 3.[26] In particular, data from BC-0 convincingly reject versions (i) and (iii) that predict all y's at 10. In fact, there is only a single $y = 10$ in the data. However, one may be wondering whether the four cases in BC-5 where $y > 15$ exert significant pull on the distribution. We cannot reject the equality of the two y-distributions in BC-treatments (two-sided Fisher's exact test has p-value $= 0.58$).

When it comes to IA-treatments (the middle panel of Table 4) we again observe distinct patterns in the data. Two observations are the key to testing various versions of the Assumption 3.

Observation 1. In both IA-treatments, we notice large frequencies of equal splits, $y = 15$. In IA-0 the share is 86% and in IA-5 it is 69%.

Observation 2. Deviations from the equal split (either $y < 15$ or $y > 15$) are significantly different between the two treatments. There are three such cases in IA-0 (12.5% of the sample agreed on $y < 15$) and eight opposite cases in IA-5 (30.7% of the sample agreed on $y > 15$). We can reject the equality of the two y-distributions in the IA-treatments at 1% level (two-sided Fisher's exact test has p-value $= 0.001$).[27]

Those observations have the following implications for versions (i)–(iv) of the Assumption 3:

Assumption 3(i) is inconsistent with Observation 1. Generating such high frequencies of equal splits in both treatments would require a large group of subjects in IA-0 with a γ_B for which $5 = \gamma_B(15)$ and another large group of subjects in IA-5 with a different γ_B for which $0 = \gamma_B(15)$. This is inconsistent with participants having comparable underlying characteristics across treatments. Moreover, in IA-5 this violates monotonicity of γ_B.

Assumption 3(ii) is inconsistent with both Observations 1 and 2. The last argument of the previous case applies here as well. To explain the high shares of equal splits in both treatments we would need to have large proportions of subjects in both treatments with γ_B's for which $0 = \gamma_B(15)$; this would violate the monotonicity assumption on γ_B. But perhaps this argument is too strict. It could be that γ_B is indeed quite flat and instead of agreeing on predicted $y = 16$ or 17 subjects naturally gravitate toward 15, despite the positive temptation costs. But if γ_B's are rather flat, then the distributions of y's in the two IA-treatments should be the same. This is rejected by Observation 2.

Assumption 3(iii) is inconsistent with Observation 1. In IA-0, A3(iii) predicts that all data be concentrated at $y = 10$ unless γ_B's are steep enough so that $y = 10 < \gamma_B(30 - y)$, i.e., $E^{(iii)}$ is empty. In that case the agreement could be any $y \in \{\gamma_B(30 - y), ..., 20\}$. The theory does not specify any particular y from this set so it is safe to assume that as long as 15 is included, it would be picked as the "focal" agreement. The fact that there is just one agreement at $y = 10$ indicates that for most subjects γ_B's are steep enough so that $\gamma_B(30 - y) > 10$. This, however, implies that $\gamma_B(30 - y) + 5 > 15$, and so in IA-5 most agreements should be compensated: $y > 15$. Although, there is a significant number of compensated agreements, the frequency of equal splits is too high to be in line with this prediction.

Assumption 3(iv) is (largely) consistent with both Observations 1 and 2, and organizes the data quite well. It predicts an equal split in both IA-treatments if γ_B is not very steep (so that the strict gains constraint is slack). Otherwise, for steep γ_B's (such that the strict gains constraint would be violated at the equal split), $E^{(iv)}$ is empty and agreements are predicted to compensate players B, i.e., $y > 15$. This is more likely to happen in the IA-5 than in the IA-0, hence, if there

---

[26] Recall, this is supporting the classical notion of deal-me-out.
[27] If one runs a test on just the observations for which $y \neq 15$, the results are virtually the same. For BC-distributions the p-value is 0.524, for IA-distributions it is 0.006.

Table 5
Hypotheses regarding the outcomes of play.

Test	Q1:		Q2:	
	Proportion *In*	*p*-Value	Average of *z*'s	*p*-Value
IA-0 vs. NN-0	0.96 vs. 0.65	0.011	16.87 vs. 19.71	0.084
IA-5 vs. NN-5	1.00 vs. 0.92	0.226	18.41 vs. 21.35	0.031
BC-0 vs. NN-0	1.00 vs. 0.65	0.002	15.17 vs. 19.71	0.001
BC-5 vs. NN-5	1.00 vs. 0.92	0.226	16.19 vs. 21.35	0.000

are any compensated agreements we would expect that they are more frequent in IA-5 than in IA-0. This is indeed the observed pattern. Taking a further step and restricting attention to those subjects in the position of player *B* who did what they agreed on – *i.e.*, who were the revealed honest – in the bottom part of Table 4, we find the same qualitative pattern and a significant difference between the two frequency distributions of *y*'s (two-sided Fisher's exact test[28] has *p*-value = 0.018).[29]

Square deals

The data in the BC-treatments are consistent with versions (ii) and (iv) of Assumption 3, and the data in IA-treatments with version (iv). Consequently, we find that the overall evidence supports version (iv). It deserves a nice name, indicating how the notions of honesty and equity shape an informal agreement. We baptize any deal satisfying the combination of Assumptions 1, 2, and 3(iv) a *square deal*. According to dictionaries, "square" can mean "straightforward and honest" as well as (in math) that "all sides are equal." Since we predict, and find, that many players are honest and that many deals involve straightforward equal splits, the terminology seems appropriate.

We hope future research will test square-deal predictions in other games. In this connection, we have a comment to add: So far we emphasized how the solution (often) involves a particular equal split. Another interpretation is feasible though. Perhaps a square deal is best thought of not as a theory of splitting gains but rather as a theory based on focal points? Splitting monetary gains, rather than overall gains (that include temptation costs), would be focal. This idea connects to a line of thought that goes back to Thomas Schelling. In two intriguing recent papers, Isoni et al. (2013, 2014) discuss Schelling's (1960) idea that outcomes under tacit bargaining (where communication is incomplete or impossible) as well as binding contracts may depend on focal points which in turn may depend on cues such as object proximity, existing location of bargaining parties, salience of geographical boundaries (*e.g.,* a river), precedence of supply chains, or a historical consumer base. Isoni et al. experimentally test Schelling's theory and find some support. We suggest that these ideas naturally extend beyond tacit bargaining and bargaining with binding contracts, to explicit haggling and informal agreements, which is our main focus. It would seem an exciting long-run goal to merge the ideas of Schelling + Isoni et al. with our framework, although in this paper we focus on a simpler norm which may nevertheless be very relevant in many contexts (that perhaps lack salient locations, rivers, or historical antecedents): 50/50 splits.[30]

3.4. No negotiation

Are outcomes the same with and without pre-play negotiation? In order to perform clean tests, we ran additional lost wallet games without pre-play negotiation: NN-0 and NN-5. For each of our four original treatments, we ask:

Q1: Does the proportion of *In*-choices by *A* players differ relative to the corresponding NN-setting?

Q2: Does the distribution of *B* players' choices (*z*'s) differ relative to the corresponding NN-setting?

A priori one may be skeptical of Q2's relevance. Conceivably, pre-play negotiation could select a biased sample of *B* players whose paired *A* players chose *In*. Such bias is obviously ruled out in NN-0 and NN-5. However, this worry is overcome because, in our data with pre-play negotiation, almost all *A*'s chose *In* (the single exception occurred in IA-0). Pre-play negotiations virtually did not restrict which *B*'s got to move.

To answer Q1 we run Fisher's exact tests based on proportions of *In*-choices, distilled from Table 2 and reported in Table 5. To answer Q2 we run Mann–Whitney tests based on entire distributions of *z*-choices, given in Table 2 whereas in Table 5 we report only averages. Table 5 reports the *p*-values of two-sided tests of the relevant null hypotheses of no

---

[28] Applying Fisher's test to sample restriction is not without caveats as marginals are no longer exogenous by design. Subjects selected into the restricted sample, making Fisher's test more conservative. In our case, we can reject at a high significance level even despite this potential issue.

[29] There are three deviations from equal split in IA-0 that go in the "wrong direction" (*y* < 15 instead of predicted *y* > 15). Notice however that all those agreements were broken; *B* reneged, violating our (extreme) assumption of honesty.

[30] Andreoni and Bernheim (2009, p. 1607) reference a variety of studies documenting prevalence of equal splits of dollar gains (*e.g.,* joint ventures between corporations, share tenancy in agriculture, bequests to children, negotiation and arbitration, business partners splitting earnings from joint projects, or friends splitting tabs).

treatment effect.[31] As seen, with the exception of the Q1-tests involving NN-5, where the proportion of *In*-choices is very high even without negotiation ($23/25 = 0.92$), and the IA-0 vs. NN-0 test for Q2 (with $p = 0.084$), all tests deliver significant differences at the 5 percent level.

One might ask (as a referee did) whether models where players' preferences depend only on the distribution of material payoffs (*e.g.*, inequity aversion) can capture the observed subject behavior in our experiment? If so, the answer to Q2 should be no, which however is not the case.

One might also ask (as another referee did) whether models of reciprocity (*e.g.* Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Cox et al., 2008) could do a good job. The answer is no. The precise argument differs between models, but the following property (which again relates to Q2) is shared in theory and violated in the data: As regards player *B*, the expected behavior is uniquely defined and should be the same in IA-*d* as in NN-*d* for $d \in \{0, 5\}$.[32]

3.5. Additional observations

We close this section by briefly discussing three systematic and intriguing patterns in the data that are either at odds with or not explicitly predicted by any of the models discussed so far.

Reneging and semi-honesty

Our first comment belongs to subjects who reneged but did not keep all 30. Such behavior lies outside the tight boundaries of our theoretical model. It is not easy to judge whether these subjects acted in an honest or a dishonest manner. On one hand, they did break the agreement; on the other hand, they still showed concern for their respective player *A*'s by sending them some money. Perhaps one might refer to them as semi-honest.

What should we think of their behavior? In IA-0 three pairs negotiated agreements in which players *B* were supposed to keep $y = 0, 10,$ and 13. However, each of the *B* players reneged by "shading" the agreed-on amount by some fraction and keeping $z = 15, 20,$ and 18, respectively. It seems as if these semi-honest *B*'s had different terms in mind – ones where $30 > y \geq 15$.[33] The remaining data are in line with this story. In IA-5 we observe eight agreements compensating *B*'s but this time only three of them reneged ($y = 16, 17, 24$ and $z = 20, 15, 22$ respectively). Each of the three subjects gave his matched player *A* a positive amount. In two of those cases *B* actually gave *A* more money than what they agreed on!

The selfish fringe

Next let us examine the behavior of player *B* subjects who kept all 30 for themselves. Only five subjects fall into this category. It might nevertheless be interesting to look at their negotiation patterns. If their behavior was planned, then they knew at the point of the agreement that they were going to renege. One would think that their main objective would then be to maximize the chances that their paired player *A* chooses *In*. What is the most likely behavior to do the job?

All five *B*'s in question ended up agreeing on an equal split. Three of them accepted the opening equal split proposals made by their respective *A*'s. One of them proposed an equal split which was accepted. And the last one initially proposed 25 for himself but that was rejected and countered with an equal split. This proposal was accepted by *B*.

Beware of people who do not goof around! The selfish fringe hide among the subjects who strike 50/50 deals. We find it intriguing that there seem to be conformity in the community of confidence tricksters. An analogous finding, for a different strategic setting with asymmetric information, is reported by Charness and Dufwenberg (2011; see Section III.C)

Bargaining delay

Our next remark concerns a systematic pattern of bargaining delay. Most of the time the parties agree quickly, but in almost all instances where the negotiating proceeds several rounds this happens in the BC-treatments and involves a player who demands more than 15. In Table 6 we list the sequences of proposal exchange for all deals that gave player *B* more money in treatments BC-5 and IA-5. BC-5 agreements that favored player *B* ($y > 15$) involved a struggle between the paired subjects, with one pair negotiating as long as fifteen rounds! By contrast, similar agreements in IA-5 were settled easily with only a few rounds of offer-exchange. We find a similar pattern for other departures from equal split in BC-0 and IA-0 (see Table 7 which presents data on the length of negotiations broken down by final agreement *y*). Agreements that depart from equal splits in the BC-treatments were hard bargains while this is not the case in the IA-treatments.

---

[31] One-sided tests (of the hypotheses that there is more giving and more *In*-choices in the IA- and BC-treatments than in the NN-treatments) reduce the *p*-values further (*e.g.*, Mann–Whitney cuts them in half).

[32] More details: D&K (extending Rabin, 1993) define an equilibrium such that *B* (purely or in expectation) chooses $x = \max\{d + 2/Y_{BA}, 30\}$, where $Y_{BA} \geq 0$ is *B*'s reciprocity sensitivity (a parameter in the model). F&F's theory also delivers a unique prediction (with $x > 15$), independently of *d*. In CF&S' theory a notion of first-mover choices that are (weakly!) "more generous than" (MGT) is central to second-mover reciprocation. In our games, choosing *In* is MGT choosing *Out*. That per se has no behavioral implications (since *B* has no choice following *Out*). In each case – D&K, F&F, or CF&S – no reference is made to whether NN-*d* or IA-*d* is considered.

[33] One of our referees suggested that guilt aversion (Battigalli and Dufwenberg, 2007) might play a role, while noting that it would take a modified version such that *B*'s guilt were convex in how much *A* is hurt relative to his expectations.

Table 6
Sequences of proposals.

	IA-5								BC-5			
Observations:	1	2	3	4	5	6	7	8	1	2	3	4
First proposer:	B	B	A	A	B	B	A	A	B	B	B	B
Round: 1	16	20	17	18	16	20	20	24	20	20	20	20
2		15			18				15	15	15	
3		18							17	18	20	
4		16							16	16	10	
5										18	20	
6										17	15	
7										18	20	
8										15	15	
9										18	20	
10										16	15	
11										18	20	
12										17	15	
13										18	20	
14										*Out*		
15										17		
Agreement:	16	16	17	18	18	20	20	24	16	17	20	20

Note: Row "First proposer" lists the player (*A* or *B*) who opened the negotiations. Sequences of proposal-exchange run from top to bottom; *e.g.*, sequence 2 in IA-5 reads as follows: player *B* made the first proposal to keep 20; player *A* countered with 15; player *B* rejected this and suggested he keeps 18; then player *A* went up to 16; and this was accepted by player *B*.

Table 7
Average length (rounds) of negotiations.

	IA-0	IA-5	BC-0	BC-5
$y = 15$	1.048 (0.218), {21}	1.5 (0.985), {18}	1.818 (1.79), {22}	1.55 (1.791), {20}
$y \neq 15$	1.667 (1.155), {3}	1.556 (1.014), {9}	4.5 (2.121), {2}	7 (6.52), {5}

Note: Standard deviations are in parentheses; number of observations are in braces; two cases in BC-5 where subjects failed to reach an agreement are excluded.

Table 8
Acceptance rates of initial proposals.

	IA-0	IA-5	BC-0	BC-5
Initial proposal = 15	20/20 100%	13/15 86.7%	14/15 93.3%	16/17 94.1%
Initial proposal ≠ 15	2/4 50%	6/12 50%	0/9 0%	1/10 10%

Why do we observe this bargaining delay with binding agreements but not with informal agreements? Our intuition is as follows: Agreements that are predicted by the theory incorporate a certain fairness standard, which makes them legitimate proposals. In BC-treatments an equal-split is the only legitimate agreement and hence the only legitimate proposal to make. Insisting on $y \neq 15$ should lead to conflict and bargaining delay. In IA-treatments, these offers could be justified based on subjects' different γ_B's. Therefore, the terms favoring one of the subjects might be more easily agreed upon.

There are multiple ways of evaluating this conjecture. One is in terms of the length of negotiations. Table 7 provides clear support for the argument suggested above. The average length of negotiations (number of rounds of offer-exchange) for proposals that cannot be easily justified (in the pooled data from BC-treatments when $y \neq 15$) is distinctly longer, 4.5–7 rounds, than for all other proposals (in the pooled data from the remaining treatments), 1.048–1.818.[34]

Another way of looking at the same issue is by comparing acceptance rates for the initial proposals.[35] Table 8 gives the summary of the data. The acceptance rate for initial proposals that are difficult to justify ($y \neq 15$ in BC-treatments) is

---

[34] The difference in distributions is statistically significant; $p = 0.000$ on an Epps–Singleton test.
[35] Only for the initial offers we are guaranteed to have independent observations.

significantly lower (0–10%) than for the rest of the opening proposals ($\geq$ 50%).[36] Equal split proposals clearly attract the higher acceptance rate ($\geq$ 86%). This suggests that departures from equal splits in the BC-treatments are typically viewed as unjustified, and so become hard bargains. The resulting deals are then likely driven by an imbalance in subjects' patience and obstinacy.

4. Discussion

Informal agreements have been given scant attention in economic theory. Are they unimportant? Couldn't agents rather rely on binding contracts? We do not think so for several reasons.

First, effective binding contracts may be infeasible. Consider two impatient fishermen in a developing country where neither courts nor enforcement are reliable. It may be close to impossible to legally enforce a contract which regulates access to a nearby lake. Does this doom the fishermen to excessive depletion of the fish stock? Even if the interaction is repeated, classical theory would say yes (because of the impatience). According to our theory, the answer may be no, if the fishermen rely on an informal agreement.

Dixit (2004) discusses countries and settings where contract enforcement is lacking,[37] and explores how informal agreements must be enforced by other means, including repetition, third-party enforcement, information exchange, and social norms of punishment. The notion that people may (to a degree) simply be honest complements his perspective.

Second, binding contracts may be illegal. Think of collusion in a one-shot government procurement auction in industrial countries. Courts exist, police can be relied on, yet bid rigging is illegal. Does that imply that the outcome will be as competitive as standard auction theory suggests? Perhaps not. Suppose firm representatives meet in a bar, have a pint, shake hands, and agree to collude. If they act as in our theory, their deal may stick.

Third, even if binding contracts are feasible in principle (as they perhaps usually are) they may be costly. A man meets a woman and they play the (one-shot, sequential) game of life with decisions on having kids, who works, divorce, alimony, etc. A binding contract may involve significant costs ranging from lawyers' fees to unforeseen contingencies to awkward feelings regarding legal chit-chat during courtship. Perhaps, instead, the couple shun the formalities, look one another in the eye, promise to be faithful forever?

Building on classical notions from the literature on binding contracts, we developed and tested four models of informal agreement formation for honest agents. All models fared well in that agreements formed and the degree of subsequent honesty was rather large. Only one theory, which we (in section 3.3) named the square deal solution, captured the following two key data features: the preponderance of agreements involve 50/50 splits of the total monetary gains and a fraction of struck deals deviate from 50/50 in a particular direction to favor the party who is most tempted to renege.

Our experiment is but an initial test. For example, one interesting further test (suggested by a referee) could relate to whether temptation costs ($\gamma_i(a)$) are strictly convex in stakes. If so, if one could increase the stakes, say by factor of 10, temptations would increase disproportionally. In lost wallet games, once player B gets to move, the temptation to renege (and take all $10 \times \$30 = \$300$) may be unbearable (i.e., incompatible with the strict gains condition). The testable implication would be that subjects should strike compensated deals (that favor B) more frequently.

The square deal solution, as we defined it, is applicable rather generally. We leave for the future further exploration of a variety of games, and broader economic modeling. In this connection, let us note several issues to keep in mind: First, our framework may be adequate for exploring "endogenously incomplete contracts" where parties elect to regulate only some choices through binding contracts. Second, we have restricted attention to settings where we could disregard moral hazard. The advantage was that we could highlight how for honest bargaining parties a key difference between informal agreements and binding contracts concerns the relative presence of temptation costs, and the way these may shape deals. However, settings with moral hazard are tremendously important and should be studied too. If this were done, one would be lead to consider combinations of binding contracts (for choices that can be monitored and enforced) and informal agreements (for choices that cannot be monitored or enforced) and temptation costs may then bear on the latter choices.[38] Third, our theories assume that agreements would be universally honored. In our data honesty was commonplace, but not universal. Addressing heterogeneity in honesty may be doable and worthwhile. Fourth, we limited attention to games with two players, but many situations involve multiple bargaining parties. Fifth, in many contexts material costs and revenues are not as readily observable as our above account (with given m_i functions) may suggest. For example, how should considerations of unobserved cost-of-effort or consumer surplus be dealt with? Sixth, even when dollar payoffs are given, 50/50 splits may not be focal in all settings and a refined theory may consider alternatives (recall our remarks in section 3.3, related to the important work of Isoni et al.).

---

[36] This difference is statistically significant – Fisher's exact test has p-value $= 0.000$.

[37] For example, he refers to Bearak (2000) who "reports that there are 25 million cases pending before the courts in India, and even if no new ones are filed, it will take 324 years to clear the backlog."

[38] This would seem relevant for example to the setting studied by Hart and Moore (2008), involving "consummate performance" which is the part of a party's effort that cannot be agreed on in a binding manner. It seems natural to imagine that parties would communicate and strike informal agreements about such choices.

We hope our paper will stimulate more work – theory & experiments – on the selection, shape, and impact of informal agreements.

Appendix

A.1. Additional information

Table 9
Treatments.

Name	Type of agreement	Outside options	Sessions	# of subjects
IA-0	Informal	(10, 0)	1	20
IA-0	Informal	(10, 0)	2	28
IA-5	Informal	(5, 5)	3	28
IA-5	Informal	(5, 5)	4	26
BC-0	Binding	(10, 0)	5 & 6	10
BC-0	Binding	(10, 0)	7	28
BC-5	Binding	(5, 5)	8	30
BC-5	Binding	(5, 5)	9	24
NN-0	No negotiations	(10, 0)	10 & 11	18
NN-0	No negotiations	(10, 0)	12	16
NN-5	No negotiations	(5, 5)	13	22
NN-5	No negotiations	(5, 5)	14	18
NN-5	No negotiations	(5, 5)	15	10

Table 10
Descriptive statistics.

Treatment	Observations	Agreement			Decisions	
		Agreed (%)	Pl. A: *In* (%)	Pl. B: *y* (st. dev.)	Pl. A: *In* (%)	Pl. B: *y* (st. dev.)
IA-0	24	24 (100)	24 (100)	14.08 (3.19)	23 (96)	16.13 (5.65)
IA-5	26	26 (100)	25 (96)	16.11 (2.321)	26 (100)	18.54 (4.99)
BC-0	24	24 (100)	24 (100)	15.17 (1.05)	24 (100)	15.17 (1.05)
BC-5	27	25 (93)	25 (100)	15.48 (1.45)	25 (100)	15.48 (1.45)
NN-0	25	–	–	–	17 (65)	19.71 (6.11)
NN-5	26	–	–	–	23 (92)	21.35 (5.26)

Note: In IA-0 one pair has agreed that player *A* chooses *Out*. Following this player *A* chose *In* and player *B* kept 15. In only two instances, both in BC-5, subjects have disagreed. In both cases player *A*'s chose *In*, player *B*'s kept 20 and 30 respectively.

A.2. Instructions

In what follows we present the universal version of the instructions in which {... *or* ...} always contains two different versions of the text that was used appropriately in different treatments.

Now that the experiment has begun, we ask that you do not talk with each other for the duration of the experiment. If you have a question after we finish reading the instructions, please raise your hand and the experimenter will approach you and answer your question in private.

You will receive $5 for participating in this experiment. You may also receive additional money, depending on the choices made (as described below). Your earnings will be paid to you in cash individually and privately.

During the session, you will be paired with another person. However, no participant will ever know the identity of the person he or she is paired with.

In the experiment, one person from each pair will be randomly selected to be Player A and the other to be Player B. The players will interact in two stages: 1. The Negotiation Stage and 2. The Game. In the negotiation stage the players can form an agreement about how to play the game. Any agreement reached in the negotiation stage {*will or will not*} be enforced and the players {*will have to play according to the agreement or be free to make any decisions*} in the game that follows. The decisions {*in the game or in the negotiation stage*} will determine how much each of the players earns in the experiment.

We next describe first the game and then the negotiation stage that precedes it.

The game

Player A moves first and chooses either IN or OUT by clicking a button labeled either "IN" or "OUT."
Player B moves second:

- If Player A chose OUT, then the game ends. *Player A receives {$5 or $10} and Player B receives {$5 or $0}*.
- If Player A chose IN, then *Player B splits $30 between the two of them*: Player B keeps x and gives $30 - x$ to Player A, choosing x such that $0 \le x \le \$30$.

The negotiation stage

Before the game is played the players can form an agreement about how to play the game. One player from each pair will be randomly selected to make the first proposal and the other player will be asked to respond to it. A proposal describes the choices of Player A and Player B in the game.
It could be:

> Player A chooses OUT

or it could be

> Player A chooses IN and
> Player B keeps x and gives $30 - x$ to Player A.

The proposal is sent to the other player by clicking on the "Submit" button. The responding player observes the proposal and chooses one of the following three options:

- *Agree with the proposal* by clicking on the button "Agree." In this case an agreement is formed and {*will or will not*} be enforced.
- *Make a counter-proposal* by clicking on the button "Make a counter-proposal." This reverses the roles of the players in the negotiation. Now, the player who clicked this button makes a new proposal and sends it to the other player. The other player will then have the chance to respond by either agreeing with the proposal, or making a counter-proposal, or disagreeing.
- *Disagree and quit negotiating* by clicking on the button "Disagree and quit negotiating." In this case no agreement is reached and negotiations terminate. Both players proceed to play the game.

Control questions (computerized, not part of paper instructions)

1. Select the correct answer. In the Game, after Player A chose IN,
   a) Player B chooses how to split $30 between himself/herself and Player A.
   b) the experiment ends.
2. Select the correct answer. In the Negotiation Stage, if one player makes a proposal and the other player agrees with it:
   a) An agreement is made and both Player A and Player B will have to play the game according to the agreement.
   b) An agreement is made but both Player A and Player B will be able to choose their actions in the game that follows.
4. Please type in the answer. In the game Player A chose IN and Player B kept $30.
   a) The payoff of Player A is: ....
   b) The payoff of Player B is: ....
5. Please type in the answer. In the game Player A chose IN and Player B kept $11.
   a) The payoff of Player A is: ....
   b) The payoff of Player B is: ....
6. Please type in the answer. In the game Player A chose IN and Player B kept $0.
   a) The payoff of Player A is: ....
   b) The payoff of Player B is: ....
7. Please type in the answer. In the game Player A chose OUT.
   a) The payoff of Player A is: ....
   b) The payoff of Player B is: ....

References

Akerlof, G.A., Kranton, R.E., 2000. Economics and identity. Quart. J. Econ. 115, 715–753.
Anbarci, N., Feltovich, N., 2013. How sensitive are bargaining outcomes to changes in disagreement payoffs? Exper. Econ. 16, 560–596.
Andreoni, J., Bernheim, B.D., 2009. Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. Econometrica 77, 1607–1636.
Battigalli, P., Dufwenberg, M., 2007. Guilt in games. Am. Econ. Rev. Pap. Proc. 97, 170–176.
Bearak, B., 2000. In India, the wheels of justice hardly move. New York Times.
Benabou, R., Tirole, J., 2002. Self-confidence and personal motivation. Quart. J. Econ. 117, 871–915.
Benartzi, S., Thaler, R., 2004. Save more tomorrow: using behavioral economics to increase employee savings. J. Polit. Economy 112, 164–187.
Ben-Ner, A., Putterman, L., 2009. Trust, communication and contracts: an experiment. J. Econ. Behav. Organ. 70, 106–121.
Binmore, K., Shaked, A., Sutton, J., 1989. An outside option experiment. Quart. J. Econ. 104, 753–770.
Binmore, K., Proulx, C., Samuelson, L., Swierzbinski, J., 1998. Hard bargains and lost opportunities. Econ. J. 111, 1279–1298.
Bolton, P., Dewatripont, M., 2005. Contract Theory. MIT Press.
Brown, A.L., Chua, Z.E., Camerer, C.F., 2009. Learning and visceral temptation in dynamic savings experiments. Quart. J. Econ. 124, 197–231.
Camerer, C., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press.
Charness, G., Dufwenberg, M., 2006. Promises & partnership. Econometrica 74, 1579–1601.
Charness, G., Dufwenberg, M., 2011. Participation. Amer. Econ. Rev. 101, 1213–1239.
Cox, J.C., Friedman, D., Sadiraj, V., 2008. Revealed altruism. Econometrica 76 (1), 31–69.
Crawford, V., Sobel, J., 1982. Strategic information transmission. Econometrica 50, 1431–1451.
Dekel, E., Lipman, B., 2012. Costly self control and random self indulgence. Econometrica 80, 1271–1302.
Demichelis, S., Weibull, J., 2008. Language, meaning and games – a model of communication, coordination and evolution. Amer. Econ. Rev. 98, 1292–1311.
Dixit, A., 2004. Lawlessness and Economics: Alternative Modes of Governance. Princeton University Press.
Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games Econ. Behav. 30, 163–182.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.
Eller, K., 2004. Integrity Is All You've Got. McGraw–Hill.
Ellingsen, T., Johannesson, M., 2004. Promises, threats and fairness. Econ. J. 114, 397–420.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54, 293–315.
Farrell, J., Rabin, M., 1996. Cheap talk. J. Econ. Perspect. 10, 103–118.
Fehr, E., Schmidt, K., 2002. Theories of fairness & reciprocity – evidence & economic applications. In: Dewatripont, M., Hansen, L., Turnovsky, S. (Eds.), Advances in Economics & Econometrics, 8th World Congress. In: Econometric Society Monographs. Cambridge Univ. Press.
Feltovich, N., Swierzbinski, J., 2011. The role of strategic uncertainty in games: an experimental study of cheap talk and contracts in the Nash demand game. Europ. Econ. Rev. 55, 554–574.
Forges, F., 1986. An approach to communication equilibria. Econometrica 54, 1375–1385.
Fudenberg, D., Levine, D.K., 1993. Self-confirming equilibrium. Econometrica 61, 523–545.
Fudenberg, D., Levine, D.K., 2006. A dual-self model of impulse control. Amer. Econ. Rev. 96, 1449–1476.
Fudenberg, D., Levine, D.K., 2011. Risk, delay, and convex self-control costs. Am. Econ. J. Microecon. 3, 34–68.
Fudenberg, D., Levine, D.K., 2012. Timing and self-control. Econometrica 80, 1–42.
Fudenberg, D., Tirole, J., 1991. Game Theory. MIT Press.
Gneezy, U., 2005. Deception: the role of consequences. Amer. Econ. Rev. 95, 384–394.
Greenberg, J., 2000. The right to remain silent. Theory Decis. 48, 193–204.
Gul, F., Pesendorfer, W., 2001. Temptation and self-control. Econometrica 69, 1403–1435.
Hart, O., Moore, J., 2008. Contract as reference points. Quart. J. Econ. 123, 1–48.
Hoffman, E., Spitzer, M.L., 1982. The coase theorem: some experimental tests. J. Law Econ. 25, 73–98.
Irlenbusch, B., 2004. Relying on a man's word? An experimental study on non-binding contracts. Int. Rev. Law Econ. 24, 299–332.
Isoni, A., Poulsen, A., Sugden, R., Tsutsui, K., 2013. Focal points in tacit bargaining problems: experimental evidence. Europ. Econ. Rev. 59, 167–188.
Isoni, A., Poulsen, A., Sugden, R., Tsutsui, K., 2014. Efficiency, equality, and labeling: an experimental investigation of focal points in explicit bargaining: dataset. Amer. Econ. Rev. 104, 3256–3287.
Kalai, E., 1977. Proportional solutions to bargaining situations: inter-personal utility comparisons. Econometrica 45, 1623–1630.
Kalai, E., Smorodinsky, M., 1975. Other solutions to Nash's bargaining problem. Econometrica 43, 513–518.
Kalai, E., Tauman Kalai, E., 2010. A Cooperative Value for Bayesian Games. Working Paper 1512. Northwestern University.
Kartik, N., 2009. Strategic communication with lying costs. Rev. Econ. Stud. 76, 1359–1395.
Kerr, N.L., Kauffman-Gilliland, C.M., 1994. Communication, commitment, and cooperation in social dilemmas. J. Pers. Soc. Psychol. 66, 513–529.
Kessler, J., Leider, S., 2012. Norms & contracting. Manage. Sci. 58, 62–77.
Levin, J., 2003. Relational incentive contracts. Amer. Econ. Rev. 93, 835–857.
Lipman, B., Pesendorfer, W., 2013. Temptation. In: Acemoglu, D., Arellano, M., Dekel, E. (Eds.), Advances in Economics and Econometrics: Tenth World Congress, vol. 1. Cambridge University Press.
Loewenstein, G., 1996. Out of control: visceral influences on behavior. Organ. Behav. Hum. Decis. Process. 65, 272–292.
Loewenstein, G., O'Donoghue, T., 2005. Animal spirits: affective and deliberative processes in economic behavior. Unpublished manuscript.
MacLeod, B., Malcolmson, J., 1989. Implicit contracts, incentive compatibility and involuntary unemployment. Econometrica 56, 447–480.
Malhotra, D., Murnighan, J., 2002. The effects of contracts on interpersonal trust. Adm. Sci. Q. 47, 534–559.
Martinsson, P., Myrseth, K., Wollbrant, C., 2012. Reconciling pro-social vs. selfish behavior: on the role of self-control. Judgm. Decis. Mak. 7, 304–315.
McCutcheon, B., 1997. Do meetings in smoke-filled rooms facilitate collusion? J. Polit. Economy 105, 330–350.
Miettinen, T., 2006. Promises and Conventions – A Theory of Pre-Play Agreements. HECER Discussion Paper #97.
Miettinen, T., 2013. Promises and conventions – an approach to pre-play agreements. Games Econ. Behav. 80, 68–84.
Myerson, R., 1986. Multi-stage games with communication. Econometrica 54, 323–358.
Nash, J., 1950. The bargaining problem. Econometrica 18, 155–162.
Nash, J., 1953. Two-person cooperative games. Econometrica 21, 128–140.
Ozdenoren, E., Salant, S.W., Silverman, D., 2012. Willpower and the optimal control of visceral urges. J. Eur. Econ. Assoc. 10, 342–368.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Amer. Econ. Rev. 83, 1281–1302.
Schelling, T., 1960. The Strategy of Conflict. Harvard University Press.
Servátka, M., Tucker, S., Vadovič, R., 2011. Words speak louder than money. J. Econ. Psych. 32, 700–709.
Thaler, R., Shefrin, H., 1981. An economic theory of self-control. J. Polit. Economy 89, 392–410.
Vanberg, C., 2008. Why do people keep their promises? An experimental test of two explanations. Econometrica 76, 1467–1480.
Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press.