

Conformity, reciprocity and the sense of justice. How social contract-based preferences and beliefs explain norm compliance: the experimental evidence

Lorenzo Sacconi · Marco Faillo

© Springer Science+Business Media, LLC 2009

Abstract Compliance with a social norm is a matter of *self-enforceability* and *endogenous motivation to conform* which is relevant not just to social norms but also to a wide array of institutions. Here we consider endogenous mechanisms that become effective once the game description has been enriched with pre-play communication allowing impartial agreements on a norm (even if they remain not binding in any sense). Behavioral models understand conformity as the maximization of some “enlarged” utility function properly defined to make room for the individual’s “desire” to comply with a norm reciprocally adhered to by other participants—whose conformity in turn depends on the expectation that the norm will be *in fact* reciprocally adhered to. In particular this paper presents an experimental study on the “conformity-with-the-ideal preference theory” (Grimalda and Sacconi in Const Polit Econ 16(3):249–276, 2005), based on a simple experimental three person game called the “exclusion game”. If the players participate in a “constitutional stage” (under a veil of ignorance) in which they decide the rule of division unanimously, the experimental data show a dramatic change in the participants’ behavior pattern. Most of them conform to the fair rule of division to which they have agreed in a pre-play communication stage, whereas in the absence of this agreement they behave more egoistically. The paper also argues that this behavior is largely consistent with what Rawls (A theory of justice, Oxford

Paper presented at the ISNIE 2008—12th Annual Conference in Toronto.

L. Sacconi (✉) · M. Faillo
Department of Economics, University of Trento, via Inama, 5, 38100 Trento, Italy
e-mail: lorenzo.sacconi@unitn.it

M. Faillo
e-mail: marco.faillo@unitn.it

L. Sacconi
EconomEtica, Interuniversity Center of Research, University Milano-Bicocca,
viale dell’Innovazione, 10, 20126 Milan, Italy

University Press, Oxford, 1971) called the “sense of justice”, a theory of norm compliance unfortunately overlooked by economists and which should be reconsidered after the behaviorist turn in economics.

Keywords Conformist preferences · Reciprocity · Veil of ignorance · Psychological games · Fairness · Experiments

JEL Classification C7 · C9 · D63 · D64

1 Introduction and motivations

In the past few years, a number of new theoretical models have been introduced in the field of behavioral game theory with the aim of explaining the systematic deviations from purely selfish behavior observed in experiments based on simple games (Ultimatum Game, Public Goods Game, Dictator Game, Trust Game etc.). Common to these models is the assumption that economic agents are characterized by complex systems of preferences in which there is room for motivations like altruism, inequity-aversion, spitefulness, desire to reciprocate other’s behaviors, and the like (Rabin 1993; Fehr and Schmidt 1999; Falk and Fischbacher 2006; Levine 1998).

Despite their contribution to enriching the traditional model of the *Homo Oeconomicus*, these theories fail to provide a thorough description of the motivations at the basis of the decision to comply with social norms. In brief, they are not able to provide an answer to this question because they incorporate a normative principle within individual preferences, without treating it as a separate object, and then they model the motivation (let it be a preference) for complying in interaction contexts with other agents. Hence, still missing is a behavioral explanation of how compliance is possible in situations where (1) norms prescribe choices which imply a cost in terms of material self-interest and (2) they cannot be fully enforced by (formal or informal) sanctions and rewards (Faillo and Sacconi 2007).

Compliance with a social norm is a matter of *self-enforceability* and *endogenous motivation to conform*, which are relevant not just to the study of social norms per se, but also to a wide array of institutions. The self-enforceability of norms is in fact also important for compliance with legal norms in all those fields where exogenous enforcement of the law is fragile or ineffective in providing incentives or sanctions, whilst self-interest is not aligned with the behavior prescribed by the law. Consider for example—at the micro level—the shared norms of business ethics which make incomplete contracts viable. Or—to jump to the level of macro institutions—conformity with the typically incomplete “constitutional contract” on a set of basic constitutional principles, which must elicit broad consensus and spontaneous compliance well before an effective mechanism of punishment of violators can be put into practice. Moreover, consider the extensive reliance on soft laws and self-regulations based on voluntary but largely accepted standards of behavior in many business fields (codes of ethics, quality standards, environmental standards, CSR and human rights international business standards etc.) typically prescribing that firms and business organizations should adopt what is *prima facie* counter-interested—at

least in the short run—economic conduct. The customary economic explanation of these phenomena usually resorts to the typical repeated games model analysis. But given that repeated game equilibria mainly rest on the quite *cognitively* fragile mechanism of reputation—how can one say that a commitment (or a type) has been carried out under unforeseen or unobservable contingencies?—it seems clear that they cannot provide a universal and self-contained explanation for these phenomena. Sometimes—most of time, we suspect—other motivational drives and cognitive mechanisms must already be at work in the one-shot game setting. These mechanisms become effective once the game description has been enriched with pre-play communication and unanimous pre-play agreements on a norm (which nevertheless remains not binding in any sense), so that the cognitive framing effects emerges that participants categorize the current situations as one where a social norm does in fact exist. A behavioral game theory explanation in these cases can view conformity as the maximization of some “enlarged” utility function, where “enlargement” is intended to make room for the presentation of an individual “desire” to comply with a norm reciprocally adhered to by other participants—which in turn is seen as depending on first- and higher-order beliefs consistent with the expectation that the norm will be *in fact* reciprocally adhered to. Contributions to the study of “norm conformity” along this line of argument are based on the models recently devised by Grimalda and Sacconi (2002, 2005, see also Sacconi and Grimalda 2007) and Bicchieri (2006).

The feature shared by these models is that the agent’s willingness to conform with a shared social norm depends on reciprocal expectations concerning what the other agents will do in a situation where a decision whether or not to conform is at stake, and at least a second-order prediction about what the second player is conjecturing about compliance by the first is involved. In some models, like Bicchieri’s, also normative expectations are important. In models of psychological games, like Grimalda and Sacconi’s, first- and second-order beliefs become parts of the argument of the utility functions—i.e., to what extent one player believes the other will conform operates not just as a probability weight that externally influences the expected value of conformity, but also as an *intrinsic motive to act expressed* by a preference for or against a given strategy. If the proper reciprocal expectations concerning reciprocity in conformity with the norm in a given state (strategy combination) are in place, the game payoffs may be accordingly influenced through the addition of a psychological utility parameter (call it “ideal conformity utility”). Grimalda and Sacconi’s model in particular sees compliance as the consequence both of agents’ participation in choosing the norm in a social contract setting ‘under a veil of ignorance’ and of the existence of expectations about reciprocal willingness to conform. Agents are characterized both by a standard *consequentialist* motivation, and by a *conditional willingness to conform with an ideal normative principle* of justice. The ideal is understood as a normative principle of justice rationalized as the agreed outcome of a social contract, modeled as the bargaining solution of a cooperative bargaining game carried out under a ‘veil of ignorance’ in a pre-play stage with respect to the play of the actual non cooperative game which is the focus of the analysis. Pre-play communication is cheap talk in standard non cooperative game theory terms. Even though principles

agreed upon by cheap talk are not effectively binding in any sense, the model explains how individual agents endowed with this two-tier motivational structure will conform with the ideal principle agreed, provided that they have unanimously agreed on it and that they also reciprocally expect mutual conformity.

In this paper, after a justification and a review of the conformist preference theory, we present an experimental study aimed at translating the “conformity-with-the-ideal preference model” into a properly defined experimental setting that makes it possible to explore the empirical validity of the theory. We devise a simple three person game (with two active players and one dummy player who can only receive the consequences of the active players’ decisions), called “exclusion game”, the purpose of which is to grasp some features of social situations where strong players decide and put into practice social institutions (for example rules on admittance into the distribution of a given social surplus) affecting not only their own well being but also that of weaker players. Standard economic behavior would naturally suggest that in this situation the active players will not share any part of the social surplus with the weak (dummy) player. If the players participate in a “constitutional stage” (under a veil of ignorance) in which they unanimously decide the rule of division by anonymous vote, even if this decision is effective only in admitting them to the play of the proper exclusion game—with respect to which it cannot bind them in any sense—the experimental data nevertheless show a dramatic change in the participants’ behavior pattern. Most of them comply with the fair rule of division that they have previously agreed upon (in the so to speak “cheap talk stage of impartial agreement on the rules”), even contradicting what they did in the first stage. This behavior is largely consistent with predictions derivable from the theory.

A similar approach to the compliance problem was suggested by John Rawls in *A Theory of Justice* (1971), where he proposed the “sense of justice” as a solution for the stability problem of a well-ordered society—i.e., a society whose institutions are arranged according to the principles of justice (norms in our sense) chosen under a ‘veil of ignorance’. This solution, however, was for long overlooked by economists and game theorists because it was at odds with the methodology of rational choice in that it resorted to socio-psychological assumptions common in theories on moral learning. Given the behaviorist turn in microeconomics, it is time to reconsider this neglected solution and to acknowledge that it suggests an illuminating explanation of why we comply with just institutions even if we have some direct material incentive not to do so.

The social contract, Rawls says, provides its own support to the stability of just institutions. In fact when institutions are just (here it is clear that we are taking the ex post perspective, i.e., once the constitutional decision from the ex ante ‘original position’ has already been taken and for some reason has been successful), those who take part in the arrangement develop a sense of justice that carries with it the desire to support and maintain that arrangement. The idea is that motives to act are now enriched with a new motivation able to overcome the counteracting tendency to injustice. Note that instability is clearly seen in term of a Prisoner’s Dilemma like situation: institutions may be unstable because complying with them may not result in the best response of each participant to other members’ behavior. However, the sense of justice, once developed, overcomes incentives to cheat and transforms fair behavior into each participant’s best response to the other individuals’ behaviors.

Consider the definition of ‘sense of justice’:

“Given that a person’s capacity for fellow feeling has been realized by forming attachment in accordance with the first two ... [levels] and given that a society’s institutions are just and are publicly known to be just, then this person acquires the correspondent sense of justice as he recognized that he and those for whom he cares are the beneficiaries of these arrangements” (Rawls 1971, p. 491).

It presupposes the development of lower-level moral sentiments of love and trust, understood as feelings of attachment to lower-level institutions (families and associations). But it is noticeable that the sense of justice is a desire to act upon general and abstract principles of justice as such, once they have been chosen under a veil of ignorance as the shaping principles of institutions—and hence have proved beneficial to ourselves in practice. It is not the case that we act upon the principles insofar as they are beneficial only to concrete persons with whom we have direct links and emotional involvements. Our desire to act upon the principles does not depend on other people’s approbation or on other contingent facts such as satisfaction of the interests of some particular concrete person. On the contrary, it is the system of principles of justice in itself that constitutes the object of the sense of justice.

The question to be answered thus becomes how it is possible that principles by themselves are capable of influencing our affections—that is, of generating the sense of justice as a relatively self-contained “desire to conform with the principles”. The answer is twofold.

First, the sense of justice is not independent of the *content* of principles. These are principles that we could have decided to agree upon under a veil of ignorance as expressions of our rationality as free and equal moral persons. These principles are mutually advantageous and hence impartially acceptable—even if endorsed from an impartial perspective—for they promote our interests and hence have some relation with our affections (preferences). Thus, in order for a sense of justice to develop, principles cannot be arbitrary. They must be those principles that would have been chosen by a rational impartial agreement.

Second, despite the intellectual effect of recognizing that principles are rationally acceptable, the basic fact about the sense of justice is that it is by nature a moral sentiment inherently connected to natural attitudes. Moral sentiments are systems of dispositions interlocked with the human capability to release natural attitudes. Moral liability for lacking moral sentiments has a direct counterpart in the lack of certain natural attitudes which results in affective responses like a sense of guilt, indignation or shame. Thus the sense of justice retains a motivational force on its own, which can be traced back to its nature as a moral sentiment or desire not entirely reducible to the experience of its intellectual justification.

Moreover, reciprocity is a basic element of the definition. In fact reciprocity is understood as a deep-lying psychological fact of human nature amounting to the tendency to “answer in kind”. The sense of justice “arises from the manifest intention of other persons to act for our good. Because they recognize they wish us well we care for their well being in return. Thus we acquire attachment to persons and institutions according to how we perceive our good to be affected by them. The basic idea is one of reciprocity, a tendency to answer in kind” (p. 494).

Important here are the following elements taken from Rawls's analysis and incorporated into the model of conformist preference explained in the next section. *First*, there is an exogenous disposition in our motivational system of drives to action—the capacity of a desire to act upon principles of justice. In an ideal theory, the weight of this motivation would be overwhelming. But in a realistic situation, where the conditions for its complete development are not fully satisfied, this exogenous motivational factor must *balance* with other motivational drives. *Second*, the foregoing element defines just a *capacity* for the sense of justice, but its effectiveness depends upon conditions relative only to the principles of justice and their compliance, as follows

1. agents construe and justify norms as the result of an impartial agreement under the 'veil of ignorance'; as a consequence, different states of affairs and compliance decisions are assessed in term of their consistency with fair principles—hence compliance is not arbitrary;
2. each agent knows that also others assess compliance decisions in a similar way;
3. each agent knows, or has the reasoned belief, that other agents are effectively playing their part in carrying out the principles, and this behavior, because of the content of the principles, expresses an intention to be beneficial to her/himself in impartial terms. Thus by playing her/his part in compliance s/he may be understood as reciprocating other agents' intentions;
4. owing to the same hypothesis of public knowledge, also other agents are predicted as having the reasoned belief that each agent does her/his part in benefiting them in an impartial manner by acting upon the principles, and thus they may be seen as reciprocating her/his intention expressed by compliance with the principles—hence her/his compliance is conditional on their reciprocity as well.
5. When these conditions are satisfied, our capacity to form a "sense of justice" becomes effective and translates into a motivational force able to counteract incentives to act unjustly in situation like the Prisoner's Dilemma game—i.e., a psychological preference for complying overcomes the preference for personal advantages gained by not complying and opportunistically exploiting other agents' cooperation.

An alternative interpretation could assume that simply because all individuals know that institutions are fair in terms of the principles, then any particular individual develops the desire to comply with them. But in this case it would be entirely unclear how an individual is able to understand that other agents' behaviors are expressing the intention to benefit her/him by following the principle of justice, which seems a necessary condition for saying that by complying with the principle s/he 'responds in kind'. If her/his response in kind does not simply amount to intellectual acceptance of the principles but also consists in complying with them, it is necessary that other agents do not simply accept or recognize intellectually that institutions are just; they must also be seen as acting upon the principles in practice. Only in this case can compliance be a response in kind—compliance in return for compliance. Thus the sense of justice not only depends on the direct assessment of any decision in terms of its coherence with principles but is also conditional on beliefs concerning the effective compliance by other agents given what they

themselves believe. Even if this seems to be the correct understanding of Rawls, we call it a weak version of Rawls's sense of justice.

2 Conformist preferences

The theory of conformist preference (Grimalda and Sacconi 2002, 2005; Sacconi and Grimalda 2007) was developed in order to explain nonprofit organizations, but it proves entirely consistent (but more precisely testable) with the general idea of norm compliance derived from Rawls.¹

Assume that two or more players are involved in a typical non cooperative game where Nash equilibria are suboptimal, or in a non cooperative division game such that the Nash equilibrium is so defined that at least some players are completely excluded from the surplus and hence equilibria are not mutually beneficial (on this game see the next section). Before such a game is played, it is assumed that a pre-play communication stage occurs wherein, by an impartial ('behind the veil of ignorance') decision, players may agree on a principle of distributive justice (a norm) assigning a solution to the ensuing game (even though this solution will not necessarily coincide with an equilibrium point, i.e., it can be incentive incompatible). In classical game theory terms, this pre-play communication phase is simply "cheap talk" in that agreements reached in this phase are not binding commitments and hence do not constrain or restrict the strategy space of the ensuing game in any way.

Nevertheless, at this stage, players put themselves in the hypothetical situation of an ex ante potential agreement. They perform the collective thought experiment of playing a bargaining game under a 'veil of ignorance'—each of them concealing from the others his/her identity and role as a player in the ensuing actual game. By this stage they can agree on a principle of justice able to determine a solution for the ensuing game from a normative point of view. The theory of conformist preference explains why this pre-play communication stage can result in effective decisions to comply with the agreed principle through an endogenous engendering of a preference favorable to compliance with the agreed principle that may counterbalance the material incentives represented in the initial description of the game.

The idea is that economic agents are motivated both by consequentialist (and mainly self-interested) and "conformist" preferences—that is, the intrinsic motivation to act according to an agreed principle if complied with reciprocally by other interacting agents as well. Thus, the utility maximization model of a rational economic man can be considerably revised, extending its explanatory and normative

¹ We do not say that the theory is entirely Rawlsian, since it assumes that in the ex ante decision a social contract is subscribed on the Nash bargaining solution of the relevant game, whereas Rawls would have suggested the maximin solution. What we simply say is that the solution given to the norm compliance problem through conformist preferences is strictly consistent with Rawls' idea of the sense of justice. However, consider that the Nash bargaining solution in a symmetric bargaining situation implies the egalitarian solution which is also consistent with Rawls' maximin. As Binmore shows, this consistency illustrates an essential feature of the decision under the veil of ignorance, when it is restricted to the payoff space resulting from the symmetric translation of the equilibrium set with respect to the Cartesian axes representing the players' payoffs (Binmore 2005).

power at a substantive level by representing these different kinds of preferences in the corresponding part of a *comprehensive* utility function.

The model assumes what we call a *description-relative* viewpoint of preferences. The same states of affairs generated by the players' strategic decisions can be described in different ways according to their relevant characteristics. A first description of states views them as consequences: what happens to any particular participant, or only to the decision maker, because of a given course of action. In general, if a player defines his/her preferences only on states *described as consequences*, then s/he has *consequentialist personal preferences*. These preferences are accounted for by the typical utility function of a player, U_i which for convenience will be called the material *component of the utility function*.

But secondly, states can be described only as sets of interdependent actions and then characterized in terms of whether or not they are consistent with a given abstract principle of distributive justice seen as resulting from a (possibly hypothetical) ex ante agreement between the players involved in the interaction. The utility function component representing these preferences will be called the 'conformist utility' of a player and it must be defined so as to give a consistent representation of the deontological motive to act that underlies this preference. In fact, intuitively speaking, a player will gain intrinsic utility from the simple fact of acting in accordance with a principle, if s/he expects that in this way s/he will be able to contribute to fulfilling the distributive principle, admitted that s/he expects the other players also to contribute to fulfilling the same principle, given their expectations.

A complete measure of conformist preferences consists in the combination of the following four elements through the conformist-psychological component of a player utility function (Grimalda and Sacconi 2005):

First, a principle T , which is a social welfare function that establishes a distributive criterion of material utilities. Players adopt T (the norm) by agreement in a pre-play phase, and employ it in the generation of a consistency ordering over the set of possible states σ , each seen as a combination of individual strategies. The highest value of T is reached in situations σ where material utilities are distributed in such a way that they are mostly consistent with the distributive principle T within the available alternatives. Note that what matters to T is not "who gets how much" material payoff (the principle T is neutral with respect to individual positions), but how utilities are distributed across players. Satisfaction of the distributional property is the basis for conformist preferences. Let us assume that T coincides with the Nash bargaining product (NBP).²

Second, a measure of the extent to which, given the other agents' expected actions, the first player by her/his strategy choice contributes to a fair distribution of material payoffs in terms of the principle T . This may also be put in terms of the extent to which the first player is *responsible* for a fair distribution, given what (s/he expects that) the other player will do. It reduces to a conformity index assuming values from 0 (no conformity at all, when the first player chooses a strategy that

² The Nash bargaining solution may be understood as a formal model for the "social contract" that players would agree in an ex ante (possibly hypothetical) collective decision on the rules that should constrain (at least as a matter of "ought") the allocation of surpluses arising from their interactions (see Sacconi 2000; Binmore 1998, 2005).

minimizes the value of T given his/her expectation about the other's strategy choice) to 1 (full conformity, when the first player chooses a strategy that maximizes the value of T given the other player's expected strategy choice).

Third, a measure of the extent to which the *other* player is expected to contribute to a fair distribution in terms of the principle T , given what s/he (is expected to) expects from the first player's behavior. This may also be put in terms of the (expected) *responsibility* of the other player for generating a fair allocation of the surplus, given what s/he (is believed to) believes. This reduces to a reciprocal conformity index assuming values from 0 (no conformity at all, when the other player is expected to choose a strategy that minimizes T given what s/he expects from the first player) to 1 (full conformity, when the other player is expected to maximize the value of T given what s/he expects from the first players) formally identical to the conditional conformity index of the first player.

Fourth, an exogenous parameter λ representing the motivational force of the agent's psychological disposition to act on the motive of reciprocal conformity with an agreed norm.³

Steps two and three coalesce in defining an overall index F of conditional and expected reciprocal conformity for each player in each state of the game. This index operates as a weight (again between 0 and 1) on the exogenous parameter λ determining whether or not λ will actually affect (and, if so, to what extent) the player's payoffs. Summing up the effect of the different components, if a player expects that the other player will be responsible for the maximal value of T , given what the other player expects about his/her behavior, and s/he is also responsible for a maximal value of T , given the other player's (expected) behavior, then the motivational weight of conformity λ will fully enter his/her utility function. That is, the player's preference system will show all the force of the disposition to conform with agreed norms, so that complying with the principle will yield utility (in the psychological sense) additional to the material payoff of the same strategy.

As a consequence, the overall utility function of player i with reference to the state σ (understood as a strategy combination of player's 'strategy σ_i and the other players' strategies σ_{-i}), is the following—see Grimalda and Sacconi (2005) for details.

$$V_i(\sigma) = U_i(\sigma) + \lambda_i F[T(\sigma)]$$

where

1. U_i is player i 's material utility for the state σ ;
2. λ_i is an exogenous parameter that may be any positive number and expresses the motivational force of the disposition to comply with an agreed principle or norm;
3. T is a fairness principle (assumed to be a *social welfare function* with the specific form of NBS), whose value is defined here for the state σ ;
4. F is a compounded index expressing both agent i 's conditional conformity and the other individuals' expected reciprocal conformity with principle T in state σ , given player i 's beliefs of first and second-order (i.e., beliefs about the other players' first-order beliefs) predicting that state σ is in fact the case.

³ This assumption corresponds to Rawls's assumption of a capacity to form a sense of justice derivable from the lower-level moral sentiments.

3 A reference situation: the exclusion game

3.1 Exclusion game

Our experimental test of conformist preferences will be carried out in a game of particular interest for understanding the role played by impartial agreements and the desire to conform with it, notwithstanding that private incentives are not conducive to such behavior. We call it the exclusion game. This game also provides us with an opportunity to provide a more technical description of the theory.

The intuition underlying this game is a social interaction between strong players and a weak player such that their mutual interaction makes a social surplus affordable, but only the strong players have decision influence over the allocation and distribution of the social surplus, whereas the weak party has neither a voice in this decision nor any retaliation threat at his/her disposal. Strong players can then decide to include the weak player in the fair sharing of the surplus, or alternatively to exclude the weak player and share out the surplus among themselves only by deciding to conform with a principle of fairness given that their personal incentives induce them to exclude the weak players.

In order to give a formal description of the exclusion game, we consider a situation (a non cooperative game) in which two individuals (player 1 and player 2) must decide how to allocate a sum of money R among themselves and a third individual (player 3), who does not have an active role in the allocation decision but whose payoff is determined by the choices made by the two other players (active players). In particular, the active players can choose between two alternative strategies: first, asking for a large share of R , i.e., high demand, $d_i^h = \frac{R}{2}$, which jointly amount to the whole surplus; or second, asking for a small share of R , compatible with a fair distribution of the surplus, i.e., the low demand $d_i^l = \frac{R}{3}$, with $i = \{1, 2\}$. The third player's payoff is the remaining share of R after the two demands of the active players have been met, i.e., $R - (d_1 + d_2)$. For example, in the case of low demand strategies by both the active players, the third player's payoff is $s = R - (d_1^l + d_2^l) = \frac{R}{3}$.

As shown in Fig. 1, if both the active players decide to ask for half of the total sum R (high demand strategies), the third player's payoff is zero; if one of the two active players decides to ask for only one-third of R , while the other one chooses to ask for half of R , the third player's payoff is $R/6$. An equal division of R among all the players—including the dummy third player—($R/3$ each) results when players 1 and 2 ask for only $R/3$.

Given the assumption that player 3 is *dummy*, in that we maintain that the players are motivated only by the intent to maximize their material payoffs, the only equilibrium in dominant strategies of this game is the one in which both active players choose to ask for $d^h = \frac{R}{2}$. Thus, the exclusion game played by self-interested players will induce the exclusion of the player with no influence on the allocation decision.

Now suppose that, before playing the game, the three players can agree on a fairness rule about how to play the exclusion game. Suppose also that they know that they will later play the exclusion game, but they are not aware of the roles that

Fig. 1 The exclusion game.
Payoff matrix

	d_2^l	d_2^h
d_1^l	$\frac{R}{3}, \frac{R}{3}, \frac{R}{3}$	$\frac{R}{3}, \frac{R}{2}, \frac{R}{6}$
d_1^h	$\frac{R}{2}, \frac{R}{3}, \frac{R}{6}$	$\frac{R}{2}, \frac{R}{2}, 0$

they will perform in it. Thus the three players have to choose, behind a veil of ignorance, what is the right way to play the exclusion game. Our key question is this: after having chosen a rule, would the players assigned an active role decide to implement the rule even if it dictated an action contrary to the maximization of their personal payoff (like the inclusion strategies in the exclusion game). And if we observe compliance with this norm, how can we explain it?

3.2 Exclusion game under conformist preferences and reciprocity

In order to apply the theory of conformist preferences to the exclusion game, we must first re-describe states of affairs resulting from the game in terms of their consistency with the ideal of fairness. Assuming that in a pre-play phase the *three* players have the opportunity to agree on a distributive principle *which involves all of them*, we can formalize the agreed principle of fairness T as the Nash social welfare function (or Nash product):

$$T(\sigma) = \prod_{i \in I} (U_i(\sigma) - c_i) \tag{1}$$

where c_i , which represents the reservation utility that players can obtain when the process of bargaining breaks down, is assumed to be equal to zero for each player.

On applying the Nash product to the states resulting from the possible plays of our game, we obtain the following fairness ordering of the four strategy combinations

$$d_1^l d_2^l > d_1^h d_2^l = d_1^l d_2^h > d_1^h d_2^h \tag{2}$$

based on

$$\begin{aligned}
 T^{\text{MAX}}(d_1^l, d_2^l) &= \left(\frac{R}{3} \cdot \frac{R}{3} \cdot \frac{R}{3}\right) = \frac{R^3}{27} \\
 T(d^h, d^l) = T(d^l, d^h) &= \frac{R}{3} \cdot \frac{R}{2} \cdot \frac{R}{6} = \frac{R^3}{36} \\
 T^{\text{MIN}}(d^h, d^h) = N(d_1^h, d_2^h) &= \frac{R}{3} \cdot \frac{R}{3} \cdot 0 = 0
 \end{aligned}
 \tag{3}$$

We can use these values to compute the overall utility values on the basis of the conformity indexes for each pair of actions and for the relative beliefs (Fig. 2). In this new context the appropriate notion of equilibrium is that of *Psychological Nash*

Fig. 2 Application of the conformist preferences model to the exclusion game

	d_2^l	d_2^h
d_1^l	$\frac{R}{3} + \lambda_1, \frac{R}{3} + \lambda_2$	$\frac{R}{3}, \frac{R}{2}$
d_1^h	$\frac{R}{2}, \frac{R}{3}$	$\frac{R}{2}, \frac{R}{2}$

Equilibrium (Geanakoplos et al. 1989⁴), which is an extension of the Nash equilibrium for situations in which expectations enter the player's utility function.

Accordingly, given the players' utilities defined as functions of their beliefs, we can easily compute the psychological equilibria of the game played by agents with conformist preferences. Strategy combinations that were not Nash equilibria in the basic game can now be defined as psychological equilibria. In particular, (d_1^l, d_2^l) is a psychological equilibrium *once it is granted that the weights λ_i are sufficiently high*:

$$V_1(d_1^l, b_1^l = d_2^l, b_1^2 = d_1^l) > V_1(d_1^h, b_1^1 = d_2^l, b_1^2 = d_1^h) \Leftrightarrow \lambda_1 > \frac{R}{2} - \frac{R}{3} = \frac{R}{6} \quad (4)$$

$$V_2(d_2^l, b_2^1 = d_1^l, b_2^2 = d_1^l) > V_2(d_2^h, b_2^1 = d_1^l, b_2^2 = d_1^h) \Leftrightarrow \lambda_2 > \frac{R}{2} - \frac{R}{3} = \frac{R}{6} \quad (5)$$

Put otherwise: given the vector of strategies (d_1^l, d_2^l) , every player i 's overall utility from strategy d_i^l —assuming a system of mutually consistent beliefs according to which each player predicts with probability 1 the symmetric strategy d^l by the opponent—is greater than the overall utility gained by deviating to the alternative strategy d_i^h , and this holds simultaneously true for both the active players. In our example this condition is satisfied when the weight of conformist preferences λ_i compensates for the loss of material utility deriving from the decision to comply with the ideal. Under these conditions there exists a psychological equilibrium of the game such that players 1 and 2 choose to ask for the lowest share of the total sum that guarantees an equal distribution of R among all three players. Thus, one of the equilibrium solutions of the psychological exclusion game is the effective inclusion of the third inactive party in the sharing of the surplus. In the event that the players have strong preferences for reciprocal conformity with the hypothetical social contract ideal of fairness, and if they have consistent reciprocal beliefs in that regard, a solution may be inclusion, not exclusion.

Note, however, that this strategy combination is not the only equilibrium of the game: also (d_1^h, d_2^h) is a psychological equilibrium when a system of beliefs exists such that both player 1 and 2 predict with probability 1 that nobody will conform with the principle, in that they have higher-order beliefs coherent with this expectation. In particular, if player 1 believes that the opponent will choose the worst action with regard to the moral principle (first-order belief), and if s/he also believes that player 2 believes that 1 will choose the same action (second-order belief), neither the opponent nor player 1 have incentives to respect the moral principle by acting against their material self-interest.

⁴ See also Grimalda and Sacconi (2005) for details.

In the following sections, after having introduced the experimental design and procedure (Sect. 4), we will show (Sect. 5) how this model can be used to formulate predictions about the choices made by subjects involved in the experiment.

4 Experimental design and procedures

The experiment took place at the *Computable and Experimental Economics Laboratory (CEEL)* of the University of Trento and it consisted of 10 sessions of 15 subjects, for a total of 150 participants.⁵ Each subject received a show-up fee of € 5 for participation.

We adopted a within-subject design, observing the behavior of the same subjects playing the exclusion game under two different conditions: before an agreement on a division rule and after the agreement on a division rule.

In particular, each experimental session was divided into three phases and lasted 1 h on average. At the beginning of each phase, one of the experimenters read out the instructions for that specific phase.

In phase 1 the subjects played a version of the exclusion game. They were assigned to groups composed of three members. Within each group, subjects were randomly attributed the roles of G1, G2 and G3. G1 and G2 were invited to play a game in which they had to decide how to allocate a sum of money ($R = €12$) between themselves and the third player, who did not have any active role in the game. In particular, active players were able to decide how much of the sum to ask for themselves ($d1, d2$), selecting one of three possible strategies: 25, 33 or 50% of R . Active players' payoffs corresponded to $d1$ and $d2$, while the third player's payoff was $R - (d1 + d2)$ (Fig. 3).

The subjects played the game three times, in three different rounds. At the beginning of each round, the three roles were randomly assigned to the members of the group. The selection mechanism was designed so that each player was able to take each of the three roles G1, G2 and G3 in turn. The subjects were told that at the end of the experiment the software would extract one of these three rounds at random, and the player's earning for phase 1 would be determined according to the outcome of that round.⁶ The game was played anonymously and subjects were not aware of the previous rounds' outcome. This procedure produced two observations for each player in this phase: his/her choice in the G1 role and his/her choice in the G2 role.

In phase 2 subjects were assigned to new groups consisting of three anonymous members. Without definition of roles, they were invited to agree, by means of a voting procedure, upon an hypothetical rule for the allocation of a sum between two

⁵ Participants were all students at the University of Trento (mainly from economics, law and sociology courses), recruited by responding to ads posted in the various departments.

⁶ This is an application of the procedure known as the *random lottery incentive system* (Starmer and Sugden (1991) and Cubitt et al. (1998)). On adopting this procedure, the round in which the subject has played occupying the G1 or G2 role is selected with a probability of 2/3, which is the same probability of being extracted as G1 or G2 in a one-shot version of the game. Note that, if we look at the third phase of the experiment, using this mechanism we can always compare the choice of each of the players who in that phase have an active role with his/her choice in the first phase.

		G2		
		3 (25%)	4 (33%)	6 (50%)
G1	3(25%)	3,3, 6	3, 4, 5	3, 6, 3
	4(33%)	4, 3, 5	4, 4, 4	4, 6, 2
	6(50%)	6, 3, 3	6, 4, 2	6, 6, 0

Fig. 3 The experimental exclusion game. Payoff matrix

active players and one non-active player. The agreement was to be reached by repeatedly playing the voting procedure until unanimity was reached, within a given limit of trials. No explicit communication, or mutual identification, was allowed among the players of any given group. In particular, after they had been informed that in the following phase they would play a game like the one played in the first phase, they were requested to vote for one of two general principles and one among some more specific rules deduced from the selected general principle (Fig. 4). Subjects were told that groups which reached unanimous agreement by voting for the same principle within five trials would pass to the voting on the specific allocation rule, upon which the groups had to agree within ten trials. A lack of unanimity after the last of the trials would prevent subjects from entering the third phase.

At the beginning of this phase, the experimenters informed the subjects about the voting procedure, stressing the correspondence between the specific rules and the game strategies of phases one and three. Absolute anonymity was guaranteed and the subjects were not allowed to communicate throughout the procedure.⁷

In phase 3, with the composition of the group unchanged, G1, G2 and G3 roles were randomly assigned to the members of each group that had agreed upon a given principle.

The subjects were involved in the same game as in the first phase, but now active players had the additional option of choosing between implementing the rule that they had agreed in the second phase or choosing one of the alternative strategies. If a player decided to implement the rule, then the corresponding strategy would be automatically selected; otherwise the strategy would be removed from his/her strategy set. Thus, for example, if player *i* was part of a group that in the second phase had reached agreement on rule 1.2 in Fig. 4 and if in phase three, when playing the role of G1, s/he decided to implement that rule, then strategy '4' was automatically selected.

Immediately after their choice, active players were asked to express their expectations about the opponent's willingness to implement the rule by guessing the outcome of the game.⁸ A scheme of the experimental procedure is given in Fig. 5.

⁷ See Appendix 1 for a detailed description of the voting procedure.

⁸ We asked the player to indicate the cell of the payoff matrix in which s/he thought the game would end. In this way we avoided explicitly asking for his/her opinion about the opponent's willingness to conform with the rule.



	PRINCIPLE 1.	PRINCIPLE 2.
Principles	“Every player should share the benefits; in particular, the player who has not been able to choose should not receive less than others”.	“Players who play under a decisional role could claim a higher share of benefits”.
		
	G1 G2 G3	G1 G2 G3
Rules	1.1) 33% 25% 42% 1.2) 25% 33% 42% 1.3) 33% 33% 33%	2.1) 50% 25% 17% 2.2) 33% 50% 17% 2.3) 50% 50% 0

Fig. 4 Second phase. Principles and rules

5 Empirical predictions

What should we expect the results of the experiment to be if we assume that the players had conformist preferences?

The answer to this question is provided by direct application to our experimental game of the model presented in Sect. 2.⁹ If $\lambda_1 > 8/3$ then player 1 will prefer strategy ‘4’ to strategy ‘6’, and the same holds for player 2. Thus the strategy combination

⁹ When subjects access the third phase they have already played the exclusion game in the first phase. The payoff from that phase (as far as they know at this stage) consists of the equi-probabilistic expectation of the two rounds they have played in the roles of the active players. This expected payoff hence is the status quo from which the gain of playing the third stage must be assessed. The total payoff a participant may earn from the first and third stages will be the sum of how much he expects from the division of the pie at the first stage ($E(\pi_1)$) and what he will get (π_2) from the pie division at the third stage. What actually is allocated to each player at the third stage hence is a *surplus*, or—to say it differently—the difference between the player total payoff and the expected payoff of the first stage, understood as the player status quo or reservation utility. In order to calculate the Nash product associated to the pie division at the third stage, thus we must account for the surplus accrued to each player for any given outcome by *adding* to the first stage payoff the amount earned at the third stage specifically *less* the status quo (i.e., the expected payoff earned at the first stage that cannot be changed at this point in time). The Nash products then become:

$$\prod_{i \in I} (U_i(\sigma) - c_i)$$

where $U_i(\sigma) = E(\pi_1) + \pi_2$ and $c_i = E(\pi_1)$ then $\prod_{i \in I} (E(\pi_1) + \pi_1 - E(\pi_1)) = \prod_{i \in I} (\pi_1 - 0)$

By applying this definition we can compute the ‘fairness values’ given by the function corresponding to the various states resulting from playing each strategy combination:

$$T(4,4) = T^{\text{MAX}} = 64$$

$$T(3,4) = T(4,3) = 60$$

$$T(3,3) = T(3,6) = T(6,3) = 54$$

$$T(4,6) = T(6,4) = 48$$

$$T(6,6) = T^{\text{MIN}} = 0$$

Hence, from the conformity indexes attached to each outcome of the game, we can compute the individual comprehensive utility values, assuming that in each state the player’s beliefs reciprocally predict exactly the strategy chosen by the opponent. These values are reported in the following matrix:

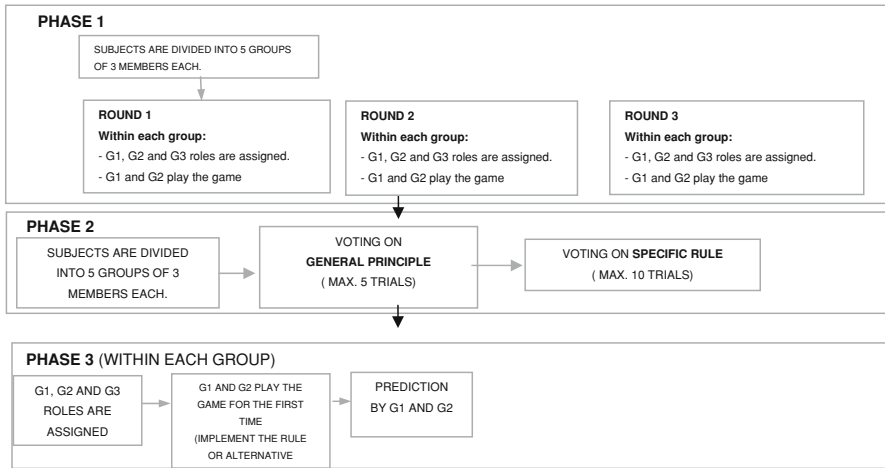


Fig. 5 The three phases of the experiment

(4, 4) is a psychological equilibrium if $\lambda_i > 8/3$ for $i = (1, 2)$ and if the players' reciprocal beliefs are coherent with these strategies. But if player 1 believes that player 2 will not choose the strategy that produces the outcome closest to the ideal one, s/he will do the same, choosing strategy '6'. Because the same holds for player 2, the strategy combination (6, 6) is a psychological equilibrium as well. As in the general case illustrated in Sect. 3, the empirical predictions about the solution of the game will depend upon what we can say about the combination of the players' reciprocal beliefs and the absolute weight of the conformist disposition. In phase 1, players have no information about the type of their opponents, nor they can refer to any pre-existing agreement about the way in which the game should be played. Thus, there is no basis for conformist preferences (there is no agreed principle to comply with nor any reason to expect compliance by the counterpart). Even though the players could in principle have a high level of λ_i , this weight simply remains inactive. We should therefore expect players, even those with a high conformist disposition, to believe that their opponents will choose the strategy that maximizes their own self-material interest, and consequently will ask for € 6 (50% of S).

Prediction 1 *In phase one, the choices of players motivated by conformist preferences will not be different from the choices of self-interested players. We will consequently find that they choose strategy '6'.*

Footnote 9 continued

	3	4	6
3	3, 3	$3 + (3/4)\lambda_1, 4 + (3/4)\lambda_2$	3, 6
4	$4 + (3/4)\lambda_2, 3 + (3/4)\lambda_1$	$4 + \lambda_1, 4 + \lambda_2$	4, 6
6	6, 3	6, 4	6, 6

In the second phase players must choose a rule on how to play an hypothetical exclusion game that may be played at a later moment. They know that if they are able to agree upon some principle of division, they will be able to play the exclusion game later, even though they do not yet know in what role they will play it again. This is a typically constitutional perspective. Such a perspective allows for the choice of general principles and rules on behavior, incorporating a view of fairness. According to a contractarian approach to the constitutional choice of principles, players will assume an impartial perspective: that is, they will judge the outcomes of the game from the point of view of each of the three roles in turn, and then choose a principle and a rule acceptable from whichever point of view. This implies a solution that must be invariant to the permutation of the individual points of view, that is, equal distribution of the surplus—if it is available within the payoff set—given what is claimed as baseline by every player in the constitutional choice. Note that within the “cooperative payoff space” defined by the exclusion game, rational bargaining according to the Nash bargaining solution would select the ‘equal division’ outcome (in coherence with the ‘invariance to symmetries’ and the Pareto postulates, granted that the *status quo* is zero).

In this setting, ‘equal distribution’ is also an intuitively obvious choice, i.e., one with high ‘salience’. Given that agreement in this phase is a necessary condition for accessing the third phase, players may vote for the most salient rule to co-ordinate their choices in a limited number of trials. Salience, of course, may depend on the simplicity of the symmetric distribution; on the other hand, cognitive simplicity may also be connected to the fairness of equal division. Whether the cognitive simplicity or the intuitive fairness of a symmetric distribution comes first is difficult to say. We are here tempted to say that the cognitive and ethical features of symmetry are quite interlocked.

However, we must point out that available to conformist players in this step is also an agreement on ‘the powerful players get all the pie’ principle which, its crudeness notwithstanding, is a possible second principle of division. Hence, we cannot uniquely predict that conformist players will choose the equal division principle. Conformity enters the picture only when a principle has been chosen in the constitutional phase, whereas the nature of the specific principle chosen depends on proper understanding of the contractarian nature of the constitutional phase, which is independent of conformity per se. (We might say that also a constitutional choice of the “egoist” principle, even though it may reflect a misunderstanding of the symmetry of the contractarian choice, once it had been made and conformed with, would be consistent with the model of conformist preferences). Thus, from our normative model of the constitutional nature of the decision phase on principles, we can only predict that the equal division principle will have some intuitive force. Note that this is also a methodological necessity if the experiment is to be able effectively to test the hypothesis of conformist preferences. In fact, if conformist players—who in the first step decided to act out of their simple self-interest—decide in the second step only to agree on “the powerful take all the pie” principle, and then decide to comply with this principle, then no evidence of change in the players’ behaviors may be observed through the experiment, since self-interest and conformity would dictate the same behavior. No falsification of the conformist hypothesis could thus be provided. On the other hand, if a significant share of the players who made a selfish choice in the first phase subscribe to an equal division principle in the second phase, then we have a clear

empirical benchmark against which conformist theory can be tested. We need simply determine whether the mere fact of a ‘constitutional’ agreement on rules—which gives the opportunity to play a beneficial division game again at phase two—is able to activate motivational forces that will induce players to conform with the agreed principle (granted that players have the appropriate beliefs), changing their conduct with respect to how they behaved in the first phase.¹⁰ Thus, our second prediction is crucial to the falsification power of our experiment.

Prediction 2 *In phase two, a rule that assigns equal payoffs to all the players will be chosen by a significant proportion of the participants.*

Assume that the players have now reached phase three. Hence they must have been able to agree on the same principle and rule. If this is enough for them to believe that a chosen principle and rule will also be played by the other players who have agreed on the same principle and rule, then their reciprocity-based conformist preferences will be activated (both deviations from conformity indexes are close to zero), granted that exogenous weights attached to non material motivations are significant. Hence, a conformist player will comply with the principle and the rule. If we hypothesize that the exogenous weight of conformist motivation is a psychological feature widespread in the population, and granted that we predict that a significant proportion of the players will have chosen an equal division rule, then we must expect a significant number of them to choose strategy ‘4’. What is most important, however, is that, if the players are conformist, we must expect the largest part of those who agreed on the equal division principle and rule to comply with the rule in the third phase, if we have some evidence that they believe that the rule will be followed by the others.

Prediction 3 *Players with conformist preferences who (having agreed on a rule) predict an outcome of the game compatible with a belief about the opponent’s willingness to implement the agreed rule, will implement the rule as well.*

Prediction 4 *Given predictions 2 and 3, a significant number of players (with conformist preferences) will request ‘4’. Moreover, players with conformist preferences who behaved according to prediction 2 and satisfy prediction 3, will request ‘4’ in the third phase.*

6 Results

6.1 Choices in phase 1 and phase 3

We begin with a general description of the subjects’ choices in phases 1 and phase 3. In order to be clear, let us recall that each subject plays two *rounds* of the game in

¹⁰ Note that this implies that only a subset of the observations consistent with the theory may have crucial discriminating force amongst different theoretical hypotheses on rational action, and we are mainly interested in producing exactly this kind of evidence. This would also have justified us in placing somewhat more stress on the ethical nature of the second phase decision in order to test the level of conformism in the third phase.

phase 1 (once in the G1 role and once in the G2 role), and then (if s/he is not selected to be a dummy player) s/he again plays the basic game in phase three.

Thus, for each subject we have data on: his/her choice in the G1 and G2 role in phase 1; the rule selected; (if active) his/her choice in phase 3; and his/her expectations about the opponent's willingness to implement the rule.

Considering all the 150 subjects involved in the ten sessions we observe that in the first phase a large majority of players (59.3%) choose to ask for € 6 when playing in both the G1 and G2 roles, leaving nothing for the third player (Table 1).

Less than one-third of the subjects made different choices in the two symmetrical roles G1 and G2, choosing to ask for € 6 as G1 (G2) and € 4 as G2 (G1).

Jumping to phase three, some players were assigned the role of the dummy player. As a consequence, because they never played during this phase, their choice cannot be compared with their choice in the first phase. Hence, if we want to compare the results of phase one and phase three, we must limit our analysis to the 100 subjects that played in G1 or G2 in this phase.

Let us focus on purely selfish and purely egalitarian choices. Looking at the Table 2 we see that in the first phase 61% of subjects choose to behave in a purely selfish way (choosing to ask for six twice) and 12% of the subjects behave in a purely egalitarian way (choosing to ask for four twice). Moving to phase three, we can see that now 56% of subjects choose to ask for six euros and 44% of the subjects ask for four euros.

To assess whether there is a significant difference between the choices made by subjects in the first phase and the choices that the same subjects make in the third phase we organize data as in Table 3, and by running the McNemar's Chi-Squared test for marginal homogeneity¹¹ we can reject the hypothesis that the marginal frequencies are the same (McNemar's Chi-squared = 15,041 $df = 1$, p -value = 0.0001).

This result seems to be compatible with the hypothesis that what happens in phase 2 (the agreement) has a significant impact on subjects behavior. In addition, if we focus on the 22 subjects who choose 6-6 in the first phase but in the third phase choose to ask only for four, we find that for all of them this choice is the consequence of the decision to comply with the 33-33-33 rule voted by their groups.¹²

Before testing the hypothesis of the significant impact of the agreement on subjects' choice, let us give a brief description of the choice and the implementation of alternative rules.

6.2 Choice and implementation of the division rules

Inspection of the data in phase 2, when subjects were requested to agree on a principle and a division rule, shows that, considering all the 50 groups:

- 32 groups (96 subjects) chose principle 1 and rule (33, 33, 33%)
- 15 groups (45 subjects) chose principle 2 and rule (50, 50, 0%)
- 2 groups (6 subjects) chose principle 2 and rule (50, 33, 17%)
- 1 group (3 subject) chose principle 2 and (33, 50, 17%).

¹¹ For the details on this test see for example Siegel and Castellan (1988).

¹² For a more detailed description of subjects' choice across the three phases see Appendix 2.

Table 1 Choice in phase 1
($n = 150$)

Choice ^a (€)	Obs.	%
3–3	0	0.00
3–4	1	0.67
4–3	4	2.67
4–4	15	10.00
4–6	20	13.33
6–4	20	13.33
6–6	89	59.33
3–6	1	0.67
6–3	0	0.00
	150	100.00

^a The two numbers represent the choices made in the G1 and G2 roles respectively

Table 2 Active players' choices in phase 1 and phase 3: a more detailed view ($n = 100$)

	Phase 1 ^a		Phase 3	
			Choose to ask for € 4	Choose to ask for € 6
3–4	1		0	1
3–6	1		0	1
4–3	1		1	2
4–4	10		2	12
4–6	4		7	11
6–4	5		7	12
6–6	22		39	61
Total	44		56	100

^a The two numbers represent the choices made in the G1 and G2 roles respectively

Unanimity on a general principle for each group was reached within a maximum of four trials (one case), with a large majority of groups reaching agreement in the first trial. The maximum number of trials required to reach unanimous agreement on a specific division rule was nine (one case), but most groups did not go beyond the second trial.

With regard to active players,

- 43 of the 64 who chose the (33%, 33%, 33%) rule decided to implement it in phase 3.
- Rule (50%, 50%, 0%) was implemented by 29 of the 30 active players belonging to the groups that selected that rule (see Table 4).

In phase 3, almost all the players that implemented the rule predicted an outcome of the game compatible with reciprocal conformity. On the other hand, many of the players who decided not to conform with the rule predicted that their opponents would do the same. 74% of subjects made correct predictions about the willingness of the opponent to conform with the rule. For a detailed description of the dynamics of subjects' choice see Appendix 2.

Table 3 Purely selfish and purely egalitarian subjects in phase 1 and choice in phase 3. (freq.)

Phase 1	Phase 3		
	Ask for € 6	Ask for € 4	
6–6	39	22	61
4–4	2	10	12
Total	41	32	73

Table 4 Active payers. Rule implementation ($n = 100$)

Rule	First time	
	Choice (obs.)	Expect the same choice by the opponent (obs.)
33-33-33		
Implement	43	42
Do not implement	21	16
Total	64	
50-50-0		
Implement	29	28
Do not implement	1	1
Total	30	
33-50-17		
Implement	1	0
Do not implement	1	0
Total	2	
50-33-17		
Implement	2	2
Do not implement	2	1
Total	4	

6.3 The effect of the agreed rule on subjects’ choice

At this point, we can assess the relative effect of the agreed rule and of subjects’ beliefs about their opponents’ likelihood of asking for € 4 in phase three this is done by estimating the following probit regression model:

$$\Pr(\text{Choice3} = 1) = \Phi(\alpha_1 \text{Rule} + \alpha_2 \text{Choice1} + \alpha_3 \text{Belief})$$

in which the dependent variable *Choice3* is equal to one if the choice in phase 3 is “ask for € 4” and equal to zero if the choice in phase 3 is “ask for € 6”; *Rule* is a dummy variable which assumes value zero when the rule dictates “ask for € 6” and one when the rule dictates “ask for € 4”; *Choice1* is the average of the two choices in phase 1, and *Belief* is a dummy variable which assumes value one if the player believes that his/her opponent has asked for € 4 and zero if she expects that the opponent has asked for €6.

Table 5 The determinants of choice in phase 3. (Probit estimate)

Rule	1.74 (0.757) ^a
Choice1	-0.27 (0.295)
Belief	2.75 (0.534) ^b
Constant	-1.77 (1.81)
Pseudo R ²	0.690
Prob > χ^2	0.0000
Number of obs.	100

Dependent variables: *Choice3*: dummy taking the value of one if the subject chooses to ask for €4 in phase 3 and zero otherwise. *Choice1*: average amount asked in the two choices of phase 1. *Rule*: dummy which takes value one if the subject has agreed on a rule dictating ‘ask for €4’ and 0 otherwise. *Belief*: dummy which takes value one if the subject believes that his/her opponent has asked for € 4 in phase 3 and zero otherwise

^a Significant at 5%

^b Significant at 1%; Standard errors in brackets

Both the variables *Rule* and *Belief* have a positive and significant effect on the likelihood of choosing €4. In particular, having agreed on the ‘ask only for € 4’ rule and expecting the opponent to choose to implement the rule increases the likelihood of choosing to ask for € 4 in phase 3. The effect of the choice in phase 1 is not significant (Table 5).

7 Discussion and conclusions

To summarize our findings:

1. The observation that most players chose strategy ‘6’ in both the rounds of the first phase is consistent with prediction 1.
2. The fact that, in the second phase, a large number of players agreed on the (33, 33, 33%) rule is consistent with prediction 2.
3. For a significant number of subjects, having agreed on a rule seems to have been sufficient reason to generate expectations about reciprocal conformity.¹³ This is consistent with our consideration concerning the “salience” of the fair solution and knowledge of its salience, so that any rational player, in the absence of evidence to the contrary, will predict that a similar opponent will be induced to act upon the same solution that s/he has chosen.
4. There is a close correlation between a player’s belief about the opponent’s willingness to conform with the rule and his/her decision to implement it; in particular, those who predict compliance with the rule (33, 33, 33%) will in fact choose strategy ‘4’ (prediction 3).
5. A significant number of those players who egoistically chose strategy ‘6’ in phase one and who agreed on the rule of equal division in phase two, decided to implement the rule in phase three; and these are definitely *most* of those who,

¹³ This hypothesis is supported by the replies to the debriefing questionnaire.

having acted as just described, believed that the agreed rule would have been played by their counterpart. All this is in accordance with prediction 4.

The significant shift in the players' behavior in the transition from phase one to phase three is consistent with the hypothesis that, because these subjects realized that the constitutional nature of the choice in phase two asked for a fairness principle of conduct incompatible with their behavior in phase 1, the mere fact of having agreed on that principle activated their conformist motivation to conform with it, granted that they believed that the same principle was being complied with by the counterparties. Notice that, whereas the players who changed their behavior crucially corroborate the theory, also most of players that exhibited the same behavior in phases one and three (given their choice in phase two and their beliefs) are consistent with the theory (i.e., do not provide any anomaly to the model). This cannot be said for concurrent models, which cannot explain our data. With regard to these alternative theories, in fact, we can conclude that:

- a) Models of inequity-aversion (Fehr and Schmidt 1999) fail to explain the observation of different behaviors in phase one and in phase three by the same subjects. Why should the subjects be inequity averse in the third phase if they were not so motivated in the first? This finding may be explained within the "inequity-aversion" framework by saying that the introduction of phase two, modeled as a constitutional choice, induced a change in the definition of the reference group whereby subjects that in the first phase did not consider the payoff of the dummy as relevant instead included it in the third phase. However, this explanation would make the inequity-aversion approach closely akin to the conformist preference model, where players that consider the advantage connected to the constitution of a social union (that is, the advantage of being allowed to play the game because of agreement on a rule) develop the motivational basis for a change in their behavior (but consider that we found that players' actions also depended on their beliefs concerning the reciprocity of the counterpart, an aspect that does not have any significant role in the inequity-aversion model).
- b) 'Direct reciprocity' models, or reciprocity models based on 'direct kindness' (Rabin 1993; Falk and Fischbacher 2006) fail to predict the dramatic change in the behavior pattern shown by subjects between the first phase, when it in fact accords with the direct kindness prediction, to the third one, when it diverges substantially from it. Note that dummy players did not change their status during the process that took the player from phase one to phase three (i.e., in the last step there were still dummy players), and they could not manifest a direct attitude or intention toward the active players since they simply did not make decisions. Nevertheless, having a rule in mind—one that has been agreed even if it is not binding or exogenously enforced—seems to be enough to change the players' behavior significantly. This suggests that some sort of commitment to the principle itself, and beliefs concerning reciprocal conformity with it, has motivational effects. Quite paradoxically, in the situation under experimentation 'fairness', understood according to Rabin's model as direct reciprocity between the two active players, would imply discrimination against the weak

player, and would result in a behavior completely indistinguishable from the conduct that shares all the pie amongst the strong players. Players who are fair according to our model act against ‘fairness’ in Rabin’s sense, and would elicit a punitive response from the other players if Rabin’s model were true.

If social preferences and reciprocity-based theories fail in explaining the evidence of our experiment, in the recent behavioral and experimental economics literature one can find a few theories that could account for the subjects’ willingness to comply with the agreed rule. The already mentioned theory of social norm devised by Bicchieri (2006), for example, would explain compliance with the rule in terms of emergence of empirical (about others’ choices) and normative (about what they think one ought to do in a particular context) expectations induced by the agreement. According to Bicchieri’s theory, subjects would comply with the rule because the agreement would make the rule salient, each member of the group would expect that the other members will comply and she also expects that the other members of her group think that she ought to comply.

A different class of theories, aimed at explaining promise keeping, would interpret the agreement of our experiment as an exchange of reciprocal promises to choose a particular strategy in phase 3. In particular, theories that explain promise keeping in terms of guilt aversion (Charness and Dufwenberg 2006; Dufwenberg 2008; Battigalli and Dufwenberg 2007¹⁴) would look at subjects’ compliance with the agreement as the consequence of their desire to avoid the feeling of guilt for not fulfilling what they think are others’ legitimate expectations.

Further research on the nature of subjects’ reciprocal expectation is needed to discriminate among these alternative explanations. And this could be one of the extensions of the present work.

- c) Let us conclude by explaining how our experiment is consistent with a Rawlsian perspective on norm compliance. The participants in the experiment were young people (students at the University of Trento), who had grown up in the context of a nearly well-ordered society, wherein they had experienced the functioning of at least some lower-level nearly just institutions like the family, associations and the like. Through these experiences they have developed—to some extent—the capacity to form a sense of justice (in our model represented by the extent of the parameter λ). Despite this fact, when confronted with the exclusion game for the first time, they still did not perform the mental experiment of an impartial justification of whatever solution, and hence they acted according to their self-interest. When the second stage was reached, however, the subjects performed the experiment of being put behind a veil of ignorance in order to agree on a principle and rule of distributive justice, with the simple cooperative intent of giving each of them the chance to participate again in the same non cooperative game. Agreeing on a principle/rule is similar to taking part in a constitutional decision. The constitution mattered to the participants not so much because it provided binding commitments but more

¹⁴ For alternative explanation of promise keeping see also Ellingsen and Johannesson (2004), Vanberg (2008).

simply because it gave a chance to participate in a potentially mutually beneficial game only by reaching an anonymous agreement on the principle/rule. Once the agreement had been stipulated, they were fully free to violate it, but they were also in the position to benefit from each another in so far as the agreement gave an opportunity to play the exclusion game again.

At the third stage in the experiment, the subjects entered a situation that Rawls would have recognized as a ‘well-ordered society’ decision case. Because they had developed the capacity to form a sense of justice, they hence should have been able to comply with principles of justice agreed under the veil of ignorance. We then can check whether the experiment satisfied Rawls’s hypothesis (see Sect. 1) and engendered behavior consistent with the idea of the emergence of a sense of justice.

Consider the players who agreed on the egalitarian principle/rule (33, 33, 33%). All of them, within their own group, knew that the others had made the same choice, i.e., under a Rawlsian interpretation that they all justified the same course of action under a veil of ignorance. In fact, they took no more than a very few vote rounds to agree on that principle/rule in stage two.

What about the shared knowledge that the opponents’ effective behavior is beneficial to each single member of the group which is necessary in order to elicit reciprocity? At this point players could not rely on the evidence of a long past history of norm compliance. Nevertheless, most of those who agreed on an egalitarian principle/rule (33, 33, 33%) also believed that their opponents would conform. This suggests that the agreement under a veil of ignorance may by itself have a strong causal effect on shaping reciprocal expectations.

This was not implicit in the conformist preference model, but is a natural consequence of the veil of ignorance reasoning format, which accords with the idea of default reasoning, and receives surprisingly strong evidence from the experiment. In order to make sense of this fact, it is important to realize that there is no logical necessity in the inference from the *ex ante* agreement to the expectation of *de facto* compliance by other participants in the stage two agreement. On the contrary, this involves a cognitive mechanism known as *default* reasoning (Reiter 1980; Bacharach 1994; Sacconi and Moretti 2008). The idea is simply that if each player has actually adopted an unanimous impartial agreement in the *ex ante* perspective, then s/he will acquire at least the *mental model* of a decision maker who acts in accordance with a plan coinciding with the terms of the agreed course of action. Agreeing on a set of actions to be carried out later implies having a mental representation of an agent carrying out a plan of action—which is simply the content of the statement of agreement.

A normally rational agent cannot fail to have this mental model because it is derived from introspection, and because the player him/herself is an exemplar of an agent who has planned to act in accordance with the content of the statement of agreement later on. But then consider that mental models are necessarily used in order to figure out possible situations and predict them. And hypothesize that at this point in time no mental model of a rational agent ‘comes to the players’ mind’ (Bacharach 2006) other than that of an agent who *will act according to the content of the agreement*. If no contrary evidence is forthcoming, the only way an agent can

simulate the other players' choice is to resort by default to his/her own mental model of a rational agent. By default, then, the same mental model is used to simulate every player's reasoning and behavior. This simulation may be recursive, so that a player uses his/her mental model not only to predict another player's behavior but also in order to simulate the other player's reasoning and beliefs, so that a *shared mental model* of all the rational agents is such that they are all expected to conform with the terms of agreement.

This explains—if not logically justifies—why the agent (as long as there is no proof to the contrary) may frame the case as a situation wherein agents conform with the norm. The *ex ante* agreement on a principle of fairness allows by default the formation of a prior belief that the propositional content of the mental model representing an agent discharging his/her commitments to an agreement is true. Just after the agreement there is no evidence that any player will not conform, whereas there is the intuitive evidence of the mental representation of an agent who agrees to a principle and hence expresses at least at that point in time the commitment to carry out a certain behavior later on. Although it would be excessive to say that this completely resolves the players' prior uncertainty, it explains how, after an agreement has been worked out—in so far as it is understood as being a constitutional, fair, initial (*ab origine*) agreement under the 'veil of ignorance'—the model of a compliant agent 'comes to their minds' with most *vividness*.¹⁵

The result is that also the fourth condition for a "sense of justice" is satisfied in the case of the groups choosing rule (33, 33, 33%): not only were their members capable of it, they agreed on a principle of justice under the veil of ignorance and had shared knowledge that they all agreed, but they also had the shared belief that they were behaving in such a way that they were impartially beneficial to each other. It follows that, if Rawls is right, those subjects who satisfied these assumptions—those who belonged to the group choosing rule (33, 33, 33%)—should show the formation of a sense of justice sufficiently strong to induce them to comply with the rule chosen. Which in fact was verified by our experimental evidence.

¹⁵ As suggested by an anonymous referee, our explanation of subjects' compliance seems to be compatible with the theory of the role of cheap talk in fostering coordination devised by Farrell and Rabin (1996), but also, we would add, with the Lewis's (1969) hypothesis that informal agreement can facilitate the convergence of players' expectations in coordination problems. Note, however, that in this article we do not deal with the problem of the selection among pre-existing multiple equilibria. According to conformist preferences theory, the agreement activates the deontological component of player's preferences, and this activation is a necessary condition for the existence of a particular equilibrium. In our experiment, without the agreement the {"ask for 4", "ask for 4"} equilibrium would not exist. At the same time, as it is typical in the context of psychological games, the existence of an equilibrium depends on the emergence of beliefs compatible with it. If players hold reciprocal beliefs compatible with the strategy combination {"ask for 4", "ask for 4"} then that strategy combination will be the psychological equilibrium of the exclusion game. But if their reciprocal beliefs are compatible with the play of the strategy combination {"ask for 6", "ask for 6"}, then the psychological equilibrium of the exclusion game will be {"ask for 6", "ask for 6"}. What the experimental evidence suggests is that the agreement seems to have an important role also in the emergence of these beliefs, and this would be in line with Farrell and Rabin (1996).

PRINCIPLE 1

"Every player should share the benefits, in particular, the player who has not the possibility to choose should not receive less than others."

N	G1 (ACTIVE)	G2 (ACTIVE)	G3 (NOT ACTIVE)
1	33%(4)	25% (3)	42% (5)
2	33% (4)	33% (4)	33% (4)
3	25% (3)	33% (4)	42% (5)

PLAYER'S ID:.....

GROUP NUMBER:

N	Player's own choice	Other players (please do not fill)			
		1	2	3	A NA
1.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
2.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
3.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
4.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
5.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
6.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
7.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
8.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
9.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA
10.	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	<input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3	A NA

OUTCOME

<input type="checkbox"/> 1	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4
----------------------------	----------------------------	----------------------------	----------------------------

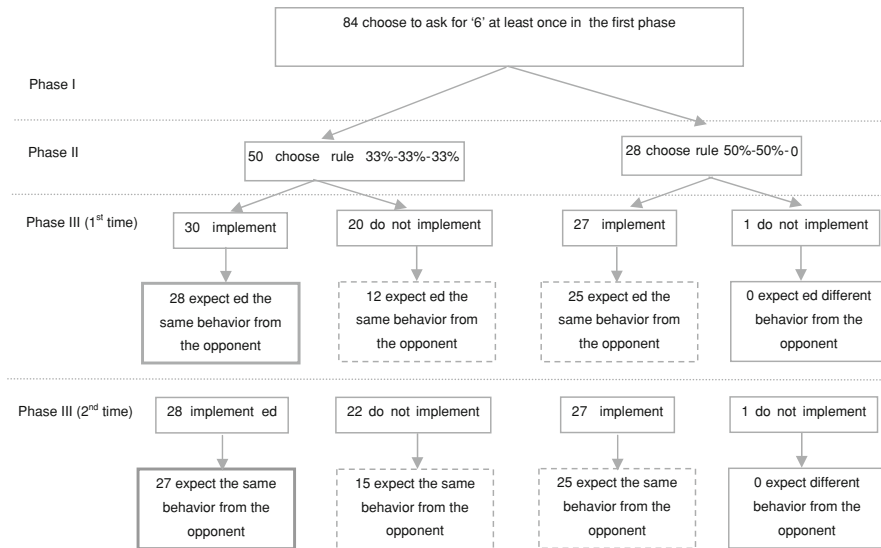
Fig. 7 Form for the selection of rule deduced from principle 1

The subjects were asked to fill in the "ID" and "Group's number" boxes and to select their preferred principle by ticking one of the two boxes in the "Player's choice" column. The experimenters collected the forms and checked the votes, writing on each player's form the choices of the other members of the group. If the members of some groups did not reach unanimous agreement, the experimenter again distributed the forms to all the subjects.¹⁶ Members of the groups that did not reach agreement were asked to vote again, while the others had to wait. The experimenter collected the forms, checked the votes and repeated the same

¹⁶ This made it impossible to identify the members of a particular group by exploiting the information about the outcome of the voting procedure.

Appendix 2

The dynamics of subjects' choice



References

- Bacharach, M. (1994). The epistemic structure of a game. *Theory and Decisions*, 37, 7–48.
- Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton: Princeton University Press.
- Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review, Papers and Proceedings*, 97, 170–176.
- Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.
- Binmore, K. (1998). *Game theory and the social contract, vol. 2: Just playing*. Cambridge, MA: MIT Press.
- Binmore, K. (2005). *Natural justice*. Oxford: Oxford University Press.
- Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74, 1579–1601.
- Cubitt, R., Starmer, C., & Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics*, 1, 115–131.
- Dufwenberg, M. (2008). Psychological games. In S. N. Durlauf & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). London: Palgrave Macmillan.
- Ellingsen, T., & Johannesson, M. (2004). Promises, threats and fairness. *The Economic Journal*, 114, 397–420.
- Faillo, M., & Sacconi, L. (2007). Norm compliance: The contribution of behavioral economics theories. In A. Innocenti & P. Sbriglia (Eds.), *Games, rationality and behaviour, essays in behavioural game theory and experiments*. London: Palgrave-MacMillan.
- Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2), 293–315.
- Farrel, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, 10(3), 103–108.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition and co-operation. *Quarterly Journal of Economics*, 114, 817–868.

- Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1, 60–79.
- Grimalda, G., & Sacconi, L. (2002). *The constitution of the no profit enterprise, ideals, conformism and reciprocity*. University Carlo Cattaneo—LIUC paper n. 155.
- Grimalda, G., & Sacconi, L. (2005). The constitution of the not-for-profit organization: Reciprocal conformity to morality. *Constitutional Political Economy*, 16(3), 249–276.
- Levine, D. K. (1998). Modeling altruism and spitefulness in experiment. *Review of Economic Dynamics*, 1(3), 593–622.
- Lewis, D. (1969). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83(5), 1281–1302.
- Rawls, J. (1971). *A theory of justice*. Oxford: Oxford University Press.
- Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, 13, 81–132.
- Sacconi, L. (2000). *The social contract of the firm: Economics, ethics and organisation*. Springer.
- Sacconi, L., & Faillo, M. (2005). *Conformity and reciprocity in the “Exclusion Game”: An experimental investigation*. Discussion Paper No. 12/05. Department of Economics University of Trento.
- Sacconi, L., & Grimalda, G. (2007). Ideals, conformism and reciprocity: A model of individual choice with conformist motivations, and an application to the not-for-profit case. In L. Bruni & P. L. Porta (Eds.), *Handbook of happiness in economics*. London: Edward Elgar.
- Sacconi, L., & Moretti, S. (2008). A fuzzy logic and default reasoning model of social norms and equilibrium selection in games under unforeseen contingencies. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 16(1), 59–81.
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Starmer, C., & Sugden, R. (1991). Does the random-lottery incentive system elicit true preferences? An experimental investigation. *American Economic Review*, 81, 971–978.
- Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. *Econometrica*, 76, 1467–1480.