Kai A. Konrad
Florian Morath

Evolutionarily Stable In-group Favoritism and Out-group Spite in Intergroup Conflict

Max Planck Institute for Tax Law and Public Finance
Working Paper 2011 – 07

July 2011

Max Planck Institute for
Tax Law and Public Finance

Department of Public Economics

http://www.tax.mpg.de

Max Planck Institute for Tax Law and Public Finance
Marstallplatz 1
D-80539 Munich
Tel: +49 89 24246 – 0
Fax: +49 89 24246 – 501
E-mail: ssrn@tax.mpg.de
http://www.tax.mpg.de

Evolutionarily stable in-group favoritism and out-group spite in intergroup conflict

Kai A. Konrad and Florian Morath

Max Planck Institute for Tax Law and Public Finance

June 30, 2011

Abstract

We study conflict between two groups of individuals. Using Schaffer's (1988) concept of evolutionary stability we provide an evolutionary underpinning for in-group altruism combined with spiteful behavior towards members of the rival out-group. We characterize the set of evolutionarily stable combinations of in-group favoritism and out-group spite and find that an increase in in-group altruism can be balanced by a decrease in spiteful behavior towards the out-group.

Keywords: altruism, spite, in-group favoritism, conflict, evolutionary stability, indirect evolutionary approach

1 Introduction

Consider individuals who are members of one group which competes with another group. Social science has produced strong evidence of 'in-group favoritism', i.e., a more positive attitude or behavior towards members of the same group than towards members of a rival out-group (see, e.g., Brewer 1979 and Bernhard et al. 2006). This in-group favoritism emerges even if the 'groups' are not formed on the basis of common characteristics or intrinsically aligned interests, but are simply generated by a random process or by other ad hoc procedures (see, e.g., Tajfel and Turner, 1979).[1] Moreover, conflict with an out-group is likely to strengthen in-group favoritism.[2] In-group altruism and spiteful behavior towards members of an out-group is

[1] For this *minimal group paradigm* see, for instance, Pinter and Greenwald (2011) who compare different experimental methods for the induction of such minimal groups.

[2] Sherif et al. 1961; see also Stein (1976), Taifel (1982) and James (1987) for surveys of early seminal contributions.

an important dimension for 'in-group favoritism' (see Mifune et al., 2010, for altruism). We uncover a new reason why a whole set of combinations of in-group altruism and spiteful attitudes towards members of the competing out-group is evolutionarily stable and where in-group altruism and out-group spite are substitutes to each other.

We apply a modification of the equilibrium concept of evolutionary stability for finite populations, introduced and developed by Schaffer (1988). We use and extend this concept in the context of the "indirect approach", i.e., in an environment in which individuals are characterized by their preference types, rather than hard wired actions. Within a finite population, an individual can improve its own relative standing not only by behavior that increases the individual's own material payoff, but also by activities that reduce the material payoff of other players. Intuitively speaking, this observation is the driving force in Schaffer's (1988) framework, and it is also the driving force for our results.

Conflict environments in which single players fight with each other on an individual basis have been well studied both in biology and in economics. For the context with finite population size, the equilibrium in evolutionary stable strategies in conflict is often characterized by higher fighting effort than in the standard Nash equilibrium that emerges from simultaneous maximization of material payoffs (see, e.g., Leininger 2003). In contrast, we consider an environment in which individuals fight in groups against each other, and in which each member of a group makes his individual decision about own effort contribution to the own group's total fighting effort. Hence, conflict between two groups involves a collective good problem within each group: members of the group make contributions that improve the chances that the group wins, and each has potentially an incentive to free-ride on the contributions of other group members.[3]

From the perspective of evolutionary fitness, the own effort contributions of a player have several aspects. Own contributions to group effort have a direct effort cost to the player making this effort. They reduce the win probability for the competing group and increases the win probability for the own group. The decrease in the win probability of the competing group is beneficial for the player, as it decreases the expected monetary payoff of members of the competing group. This increases the player's material payoff relative to that of the members of the rival group. The increase in the win probability of his own group is a mixed blessing. It increases the

[3]The seminal paper that studied this problem is Olson and Zeckhauser (1966). It has generated a large literature, including contributions by Baik et al. (2001), Davis and Reilly (1999), Esteban and Ray (2001), Katz et al. (1990), Nitzan (1991), and Ursprung (1990).

own expected material gain of the player. However, the increase in win probability also increases the expected material payoff of all other members of the own group (who did not have to bear the additional cost of this effort). In isolation, this latter effect makes the individual less well-off compared to the other members of the own group.

The equilibrium in evolutionarily stable strategies that results from these partially countervailing effects is the starting point of our main analysis. The main analysis asks whether the combination of in-group altruism and out-group spite can be jointly established as evolutionarily stable preferences. Formally we adopt the "indirect approach" that considers the evolution of preferences rather than the evolution of actions introduced by Güth and Yaari (1992) in the context of evolutionary game theory. It suggests that -rather than actions being hard wired- individuals may be endowed with objective functions and make optimizing decisions based on their objective functions. These objective functions are genetically determined, but subject to possible mutations and evolutionary selection pressure. This line of reasoning has been used to provide an evolutionary foundation for a number of types of other-regarding preferences.[4] Different attitudes towards members of the own group than towards members of the out-group may be desirable from an evolutionary perspective. Indeed, our analysis confirms that a large set of combinations of in-group altruism and spiteful preferences with respect to the competing members in the out-group are evolutionarily stable preferences. This set has the property that in-group altruism and spite toward the out-group are substitutes.

We make use of a duality relationship that exists between evolutionarily stable strategies (in terms of effort choices) and evolutionarily stable preferences. A suitable mixture of in-group altruism and spiteful preferences towards members of the out-group can implement behavioral choices that are in line with the evolutionarily stable strategies in the direct approach. This duality exists in a framework where we assume that players cannot observe the preference type of their co-players.

Previous research has analyzed several evolutionary explanations for altruism or spite.[5] Altruism or harming or spiteful behavior have been derived and explained by evolutionary arguments in the context of group selection,[6]

[4]Related to this approach, Frank (1987) highlighted the importance of strategic effects of other-regarding preference traits in strategic situations, and the implications for evolution of preferences.

[5]For recent reviews and contributions in evolutionary biology see Lehmann and Keller (2006), West and Gardner (2010) and Marshall (2011).

[6]Smirnov et al. (2007), for instance, survey the literature on the willingness to take major risks, including the risk to sacrifice one's own life. They offer an evolutionary expla-

and kin selection[7] focussing on relatedness and inclusive fitness. Our explanation does not use any argument that is related to kin selection.

Altruism, spite, and other types of other-regarding preferences have also been shown to be evolutionarily stable in a framework in which players can observe other players' preference types and where they can base their behavioral choices in an interaction with another player on the preference type of this other player. Evolutionary stability of other-regarding preferences has generated considerable interest inside economics.[8] An important paper that is closest in spirit to our research question and considers competition between groups is Eaton et al. (2011) who also aim at an explanation of in-group favoritism. They consider a framework which combines two distinct action choices: production effort and appropriation effort in a model of conflict. In-group altruism may develop along one activity dimension and out-group spite may develop along the second activity dimension in this context, and the combination of both types of other-regarding preferences in one model approach nicely addresses the well-documented phenomenon of in-group favoritism. They consider an infinitely large population and what is sometimes called the 'transparent disposition' approach, in the spirit of Bester and Güth (1998). The driving force for evolutionary stability of these two types of other-regarding preferences in their context is type observability. In our framework, the combination of in-group favoritism and spiteful behavior toward the out-group emerges from one single activity, and the evolutionarily stability of the combination of in-group altruism and out-group spite emerges, even though other players' types are not observable. In the context of evolutionary biology, a variant of these considerations can be found based on repeated interaction and/or non-additive fitness consequences (see Fletcher and Doebeli 2006, 2009). A related effect has been studied as the phenomenon of "greenbeards", where altruistic behavior is conditional on relatedness of the recipient (see, e.g., Gardner and West 2009 and West and Gardner 2010). This selective behavior requires that strategies are conditioned on co-players' types; hence, it requires that other players' types are (at least partially) observable. Our approach does not rely on this mechanism

nation, based on group selection. Bowles (2009) analyses a related argument, considering whether a group selection argument can be based on the structure and interaction of groups in ancestral hunter-gatherer societies. For a discussion of group selection see, e.g., Sober and Wilson (1998), and for more critical views Reeve (2000) and Maynard Smith (1998), and Salomonsson (2010) for a recent survey on the group selection controversy.

[7]For a survey on altruism and spite in the context of kin selection see West and Gardner (2010).

[8]See, e.g., Güth and Yaari (1992), Sethi (1996), Bester and Güth (1998), Huck and Oechssler (1999), Koçkesen, Ok and Sethi (2000), Dufwenberg and Güth (2000), Ok and Vega-Redondo (2001), Abreu and Sethi (2003), and Dekel, Ely and Yilankaya (2007).

and does not require observability of type or relatedness.

Our approach has two important key aspects: first, we consider finite population size. This makes the consequences of a player's actions on other players' monetary payoff relevant for his fitness. Second, we consider conflict as taking place between two groups that do not cooperate internally. The inter-group competition aspect distinguishes our framework from evolutionary models of conflict between single individuals.[9] It allows us to address the phenomenon of in-group favoritism and the role of out-group competition for this attitude. Also, inter-group competition generates scope for a richer type space, by which the 'type' describes the behavior or the preference towards members of the in-group that can differ from the behavior towards the out-group - a complexity that cannot emerge in individual players' contests.[10]

Social psychology explained in-group favoritism relating to concepts of social identity and social comparison (see Tajfel and Turner 1979 for an outline) and by the 'realistic group conflict theory' (Sherif et al. 1961) as a theory for describing the role of an out-group for in-group favoritism. Recent work traces physiological roots of in-group favoritism using twin studies (Lewis and Bates 2010) and neuroimaging (Mathur et al. 2010). Our analysis complements these theories, showing that a genetic underpinning for in-group altruism combined with spiteful attitudes towards the rival out-group proves to be evolutionarily stable.

In the next section 2 we characterize the state game with two rival groups. We then consider evolutionarily stable strategies in the inter-group conflict in section 3. In section 4 we make use of a duality property to show that these evolutionarily stable strategies can be induced by evolutionarily stable preferences that exhibit in-group altruism and spite towards members of the out-group. Then we conclude.

[9]For conflict between individuals, Konrad (2004) found an evolutionarily stable bimorphism in which some share of the population has altruistic preferences and the other share has spiteful preferences. He rules out narrowly selfish behavior as a possible preference type, however, which discounts his result. In the context of conflict, Eaton and Eswaran (2003) and Leininger (2009) studied evolutionary stability of preferences in contests between individuals and also found evolutionary foundations for spiteful preferences.

[10]In a late stage of this work we became aware of a paper by Eaton et al. (2011), who also consider at the framework of in-group favoritism with out-group competitors. Their analysis is based on observability of types - hence their results are essentially based on a combination of known results about evolutionary stability of altruism (as in Bester and Güth 1998) and spite (Eaton and Eswaran 2003).

2 The state game

We consider an environment in which a finite set N of $2n$ of players i constitutes the set of players who participate in the following state game.[11] The players are partitioned in two alliance groups of equal size, denoted as group A and group B, each consisting of n players. The conflict between the two groups is described by a fight that is mapped by a Tullock (1980) lottery contest[12] as follows: all players i simultaneously and independently expend an effort $x_i \geq 0$, which is also equal to their material cost of expending this effort. Efforts of members of the same group sum up to the total group effort, $X_A = \Sigma_{i \in A} x_i$ and $X_B = \Sigma_{i \in B} x_i$, respectively. These total group efforts determine the win probabilities for group A and B, respectively. Group A wins the contest with a probability

$$p_A(X_A, X_B) = \frac{X_A}{X_A + X_B} \text{ if } X_A + X_B > 0 \tag{1}$$

and with a probability $p_A(X_A, X_B) = 1/2$ if $X_A + X_B = 0$. With the complementary probability $p_B = 1 - p_A$ group B wins the contest. This mapping (1) is often referred to as the Tullock lottery contest success function.

If group $K \in \{A, B\}$ wins the contest, each member of the alliance K receives an equal amount $Q(n)$. The members of the losing alliance receive nothing, but have to bear their cost of effort. In general, Q can, but need not be a trivial function of n. For instance, $Q = Q_0/n$ refers to the one extreme case case where the prize of winning of monetary size Q_0 is a private good that is evenly shared within the group; $Q(n) \equiv Q_0$ refers to the case in which the prize of winning is a group-specific public good and all members of the group value the public good symmetrically. Values of $Q < Q_0/n$ are also feasible and emerge, for instance, if there is some fighting inside the winner group about how to allocate the prize between them.[13] This setup determines the material payoff of a player as a function of his own effort, the effort choices of the co-players with whom he is matched in the given state game, and the outcome of the respective lottery according to (1). For given

[11] The state game here is essentially the static game that is analysed by the literature on inter-group contests, following the tradition of Olson and Zeckhauser (1966).

[12] See chapter 2.3 in Konrad (2009) for a survey on the different fields (rent-seeking, military conflict, marketing, sports) in which this contest has been developed independently as a tool to describe conflict, for axiomatic foundations such as Skaperdas (1996), and for a survey on existing microeconomic foundations for this decision rule.

[13] See, for instance, Katz and Tokatlidu (1996) and Wärneryd (1998) for the analysis of subgame perfect Nash equilibrium in this context if players care only about their own absolute material payoff.

choices $(x_1, ..., x_{2n}) \equiv \mathbf{x} \in \mathbb{R}^{2n}_+$ the material payoff of a player i in group $K \in \{A, B\}$ is

$$\pi_i = p_K(\mathbf{x})Q - x_i. \tag{2}$$

This material payoff consists of the material benefit which i enjoys if i's group wins times the probability that i's group wins, minus the actual fighting effort which i contributed to the fighting effort of his group. This material payoff can be interpreted as an expected value if p_K is a probability. Alternatively, p_K need not be interpreted as a probability of winning, but as a share in the total prize, with the total prize being split between the two groups according to shares p_K and $(1 - p_K)$. These two interpretations are used equivalently in the rent-seeking literature, if players are risk-neutral. Focussing simply on expected material payoff, we disregard this possible distinction in what follows.

3 Evolutionarily stable strategies

We now search for an equilibrium in evolutionarily stable strategies and ask which $x_i = x^E$ for all $i \in N$ is an evolutionarily stable strategy. This implies that we concentrate on the case in which the equilibrium in evolutionarily stable strategies is a monomorphism, i.e., characterized by a single effort level x^E. We assume that mutations from such a monomorphism may happen, but we restrict the types of mutations that can emerge to one single mutant type at a time. That is, starting from a homogenous population in which all players follow the strategy x^E, a mutant player may appear who chooses a different effort x_M.

We can now provide a definition for a monomorphic equilibrium in evolutionarily stable strategies that rests on the definition of stability introduced by Schaffer (1988). In this definition, the set of material payoffs of players $i \in N$ is the determinant of evolutionary success as follows:

Definition 1: *The strategy $x^E > 0$ is an evolutionarily stable strategy if x^E is a solution to*

$$\max_{x_i}[\pi_i(x_i, \mathbf{x}^E_{-i}) - \pi_{-i}(x_i, \mathbf{x}^E_{-i})], \tag{3}$$

where

$$\pi_i(x_i, \mathbf{x}^E_{-i}) = \frac{(n-1)x^E + x_i}{(2n-1)x^E + x_i}Q - x_i, \tag{4}$$

$$\pi_{-i}(x_i, \mathbf{x}^E_{-i}) = \frac{n-1}{2n-1}\left[\frac{(n-1)x^E + x_i}{(2n-1)x^E + x_i}Q - x^E\right] + \frac{n}{2n-1}\left[\frac{nx^E}{(2n-1)x^E + x_i}Q - x^E\right] \quad (5)$$

and $\mathbf{x}^E_{-i}$ *is the vector of the efforts of all* $2n-1$ *players other than* i*, who all choose* x^E*.*

In words, $\pi_i(x_i, \mathbf{x}^E_{-i})$ as in (4) is the expected material payoff of a player i who chooses effort x_i given that all other players choose effort x^E, and $\pi_{-i}(x_i, \mathbf{x}^E_{-i})$ as in (5) is the expected material payoff of a player who chooses x^E if all but one other players also choose x^E, and this one other player chooses effort x_i. This one other player may belong to the same group as player i, which happens with a probability $\frac{n-1}{2n-1}$ and to the rival group with a probability $\frac{n}{2n-1}$.

Definition 1 is taken directly from Schaffer (1988). It formalizes the standard idea of evolutionary stability, but for a finite population. A population which consists of players who all follow the strategy x^E is evolutionarily stable if it cannot be successfully invaded by a mutant who chooses a mutant strategy $x_M = x_i$. According to this definition, suppose there is a mutant playing x_i. If this mutant has a strictly higher payoff than the average payoff of the non-mutants who all choose x^E, then this violates the property that x^E is a solution to (3), and this mutant does better than the average player in the group of non-mutants. Due to the association in groups, the mutant belongs to one of the groups and plays against a homogeneous group consisting of players who all choose x^E. This in turn yields different material payoffs to the non-mutants, depending on whether they are in the same group as the mutant, or in the rival group.

The first-order condition[14] for a maximum of (3) evaluated at $x_i = x^E$ yields

$$x^E = \frac{1}{2}\frac{Q}{2n-1}, \quad (6)$$

and total effort per group is equal to

$$X^E = \frac{1}{2}\frac{Qn}{2n-1}.$$

We summarize this as

[14]Note that the second derivative of $\left[\pi_i(x_i, \mathbf{x}^E_{-i}) - \pi^E_{-i}(x_i, \mathbf{x}^E_{-i})\right]$ is strictly negative for all possible values of $x^E > 0$, and that the first marginal unit of effort x at $x = x^E = 0$ has a positive impact on the value of $\left[\pi_i(x_i, \mathbf{x}^E_{-i}) - \pi^E_{-i}(x_i, \mathbf{x}^E_{-i})\right]$.

Theorem 1 *A symmetric equilibrium in evolutionarily stable strategies of the inter-group contest is described by effort choices* $x^E = \frac{1}{2}\frac{Q}{2n-1}$.

This result shows that the evolutionarily stable strategy is increasing in Q and decreasing in the size of the total population. For instance, if $Q(n) = Q_0$ (the pure public goods case), individual contributions converge towards zero as the population size becomes very large. However, total contributions of each group converge towards $1/4$.[15] If $Q = Q_0/n$ (the pure private goods case), individual contributions and total group efforts converge to zero as $n \to \infty$.

4 Evolutionarily stable preferences

We can now consider evolutionary stability in the context of the evolution of preferences, in line with the indirect approach introduced by Güth and Yaari (1992). A player's material payoff relative to other players' material payoffs determines evolutionary success. However, rather than considering evolutionary strategies (fixed effort choices in the context here) and their evolutionary success, we allow players to differ in their subjective utilities, assuming that the players consciously maximize their subjective utility by their choices of actions, as in a strategic game. Players' types will be defined based on their subjective preferences and their beliefs about the preferences of others. Mutations and evolutionary selection then operates on the set of possible preference/belief types.

To be more specific, in each state game there is, again, a set N of players i. The players are randomly partitioned into the two groups A and B. Each player has a set of possible strategies $x_i \in [0, \infty)$ and chooses freely from this set. All players' choices are made simultaneously, and the two aggregate group efforts enter into the contest function and determine which of the groups wins with a prize Q. This prize is allocated to each member of the winning group, just as described in section 2. The definition of material payoff of players also remains as in (2). Players do not necessarily maximize this material payoff, however. Instead, each player has a 'subjective utility', where this utility function characterizes an individual's 'type'. We later ask which subjective utility is evolutionarily stable - with a definition of evolutionarily stable utility given further below. We are interested in the possible role of in-group altruism and spiteful preferences towards the members of

[15] Not surprisingly, for $n \to \infty$, the equilibrium in evolutionarily stable strategies coincides with the Nash equilibrium of intergroup contests, which is well studied and well known.

the out-group. Therefore, we consider the following parametric version of subjective utility of player i in group A as a function of material payoffs of all players,

$$U_i(a_i, s_i) = \pi_i + a_i \sum_{j \in A \backslash \{i\}} \pi_j - s_i \sum_{j \in B} \pi_j, \tag{7}$$

where π_i is i's own material payoff, the second term is the sum of the material payoffs of all players who are in the same group as i, and the third term is the sum of material payoffs of all players who are in the other group. Further $a_i \geq 0$ and $s_i \geq 0$ are the utility weights given to the monetary payoffs of other in-group players $j \in A \backslash \{i\}$ and out-group players in set B. A strictly positive value of a_i measures i's in-group altruism, a strictly positive s_i measures spiteful feelings vis-à-vis members of the rival, out-group B. The space of possible subjective utilities for members of group B is defined analogously. Accordingly, the preference type of a player i is determined by a pair (a_i, s_i).

To complete the description of the state game, we have to specify the information assumptions and players' beliefs that apply. Players' preference types are private information: each player knows his own type, but not that of others. Players cannot observe the preference type of other players, but nevertheless need to have or form beliefs about other players' types. Together with a player's preference parameters, the player's beliefs are part of the characterization of the player's type. We first define *robust beliefs*.

Definition 2: *Suppose player $i \in N$ has preference parameters (a_i, s_i). This player's belief about his co-players' types is a robust belief if i beliefs that all other players $j \in N \backslash \{i\}$ are also of preference type (a_i, s_i) with probability* 1.

In what follows we assume that all players have robust beliefs as defined in Definition 2. Hence, a player's type is fully characterized by a pair of preference parameters (a_i, s_i), and beliefs about other players that are identical with the player's own preference type.

We note several properties of robust beliefs. First, in a monomorphism of evolutionarily stable preferences all players have the same preference parameters. It follows directly that players' beliefs are consistent with the true distribution of types in each evolutionarily stable equilibrium. Beliefs are incorrect whenever the population N consists of individuals with different preferences, i.e., outside a monomorphism of evolutionarily stable preferences. These two properties are important and nice features of robust beliefs.

They can be seen as the evolutionary-equilibrium analogon to the requirement in Bayesian Nash equilibrium that beliefs must be correct along the equilibrium path - but not for out-of-equilibrium outcomes.

Second, robust beliefs allow to solve for what we call the symmetric robust-beliefs Nash equilibrium effort of each player type, for any given true distribution of types among all other players. In fact, for robust beliefs, the player has a dominant choice and can be characterized as follows. Let $\mathbf{x}_i^* = (x_i^*, x_i^*, ..., x_i^*)$ be the vector of efforts in the symmetric Nash equilibrium of mutually optimal replies if all players are of type (a_i, s_i) and maximize (7). Then the dominant choice that maximizes (7) for player i given his preferences (a_i, s_i) and his robust beliefs is $x_i = x_i^*$. This implies that, should all players are of the same type (a_i, s_i), they end up with effort choices $(x_i^*, x_i^*, ..., x_i^*)$. Should some players have different preference parameters, the effort choice of player i of type (a_i, s_i) is still uniquely determined and equal to $x_i = x_i^*$, for any possible type (a_i, s_i).

Note that, for a set of players with heterogenous preference parameters, this choice behavior will not be ex post optimal, as the players are surprised about the effort choices by others. However, such surprises occur only out of the equilibrium in evolutionarily stable preferences. Note also that the assumption of robust beliefs and the players' choice behavior that is implied is convenient for the formal analysis, but does not drive our results. Our main result does not rely on this choice of beliefs.[16]

The assumption of unobservability of types is a major departure from the evolutionary literature on altruism. Observability of preference types is frequently assumed in the context of the indirect approach, starting with Frank (1987). Type observability has strategic implications: a player's type may induce the equilibrium actions of co-players, and a change in a player's type can therefore cause a change in co-players equilibrium actions. As has been shown by Bester and Güth (1998), the direction of this strategic effect of a player's own type for co-players' actions is crucial for the evolutionary success of particular preference traits, including altruism. With observed altruism or spite parameters, these induce a strategic effect on other players: other players' optimal effort choices become a function of the preference type of player i, as they anticipate that player i's effort choice depends on i's own preference, and different choices x_i induce different optimal replies for other players. We depart from this observability assumption, because the assumption of observability is a strong and empirically less plausible

[16]Note that in an evolutionary context in which the distribution of types develops stochastically and is endogenous, the standard concept of Bayesian beliefs with all players' types being random draws from a commonly known distribution does not work.

assumption, and because the strategic effect of type observability is known and well understood. This also implies that our findings on evolutionarily stable altruism and spite are not based on this strategic effect.

We now ask which combination of (a^E, s^E) is an evolutionarily stable type in the following sense. Let all $2n-1$ players be of type (a^E, s^E) and let one individual i be of type (a_i, s_i). If $(a_i, s_i) \neq (a^E, s^E)$ we call this individual i a mutant. Evolutionary stability of preferences in the line of reasoning of Schaffer (1988) is a property about the relative advantages of this preference type given the uniform preferences of all other players. Suppose that $2n-1$ players follow subjective utility maximization according to a given type characterized by (a^E, s^E). Let there be a single mutant with (a_i, s_i) different from (a^E, s^E). Let $\pi_i((a_i, s_i), (\mathbf{a}, \mathbf{s})^E)$ be the material payoff obtained in the robust-belief Nash equilibrium by the mutant if the mutant and all other players maximize their own subjective utilities and have robust beliefs, where $(\mathbf{a}, \mathbf{s})^E$ denotes the vector $((a^E, s^E), ..., (a^E, s^E))$ of preference types of all individuals other than i. Further, let $\pi_{-i}((a, s), (\mathbf{a}, \mathbf{s})^E)$ be the material payoff obtained in this equilibrium by each of the other $2n-1$ players who maximize their own subjective utilities and are of type (a^E, s^E). An adaptation of Schaffer's stability criterion for evolutionarily stable preference types then is as follows:

Definition 3: *The preference* (a^E, s^E) *is an evolutionarily stable preference if* (a^E, s^E) *is a solution to*

$$\max_{(a_i, s_i)} [\pi_i((a_i, s_i), (\mathbf{a}, \mathbf{s})^E) - \pi_{-i}((a_i, s_i), (\mathbf{a}, \mathbf{s})^E)] \tag{8}$$

Using this definition we can now state our main result:

Theorem 2 *Let* $\mathcal{P} = \{(a, s) \,\big|\, a \geq 0$ *and* $s \geq 0$ *and* $a = \frac{1}{(2n-1)(n-1)} - \frac{n}{(n-1)} s\}$. *Then each* $(a, s) \in \mathcal{P}$ *constitutes an equilibrium in evolutionarily stable preferences with robust beliefs.*

Proof. Let us denote the $(2n-1)$-dimensional vector of $x_j = x^E$ for $j \neq i$ as $\mathbf{x}^E_{-i}$. We make the following three observations:

Observation 1: Let $(a^E, s^E) \in \mathcal{P}$. Suppose all individuals have these preference parameters and robust beliefs. In this case

$$x_j = \frac{1}{4} Q \frac{1 + a(n-1) + sn}{n} \equiv y \text{ for all } j = 1, ..., 2n \tag{9}$$

is a set of mutually optimal replies. We call this a Nash equilibrium with robust beliefs.

To confirm this observation, suppose that some player i anticipates that all other players choose this effort y, i.e. $\mathbf{x}_{-i}^E = (y, ..., y)$. In this case, player i with preference parameters (a_i, s_i) chooses an effort that maximizes

$$\begin{aligned} U_i((a_i, s_i); \mathbf{x}_{-i}^E) &= \frac{x_i + (n-1)y}{x_i + (2n-1)y} Q - x_i \\ &\quad + a_i(n-1)\left(\frac{x_i + (n-1)y}{x_i + (2n-1)y} Q - y\right) \\ &\quad - s_i n \left(\frac{ny}{x_i + (2n-1)y} Q - y\right) \end{aligned}$$

This maximization problem has a unique interior solution at $x_i = y$, as can be shown as follows. From

$$\begin{aligned} \frac{\partial U_i((a_i, s_i); \mathbf{x}_{-i}^E)}{\partial x_i} &= \frac{nyQ}{(x_i + (2n-1)y)^2} (1 + a_i (n-1) + sn) - 1 \\ &= \frac{(2ny)^2}{(x_i + (2n-1)y)^2} - 1 \end{aligned}$$

it follows that $\partial U_i((a, s); \mathbf{x}_{-i}^E)/\partial x_i$ is strictly positive at $x_i = 0$ and strictly decreasing in x_i. The non-zero symmetric solution is $x_i = y$ where y is given in (9).

Observation 2: The set of (a_i, s_i) that implements effort choices $y = x^E$ $(= \frac{1}{2}\frac{Q}{2n-1})$ in the symmetric Nash equilibrium with robust beliefs is characterized by $a_i = \frac{1}{(2n-1)(n-1)} - \frac{n}{(n-1)} s_i$.

For Observation 2 to be true it must hold that $x^E = y$, or $\frac{1}{2}\frac{Q}{2n-1} = \frac{1}{4} Q \frac{1 + a_i n - a_i + s_i n}{n}$. Solving this for a_i as a function of s_i yields

$$a_i = \frac{1}{(2n-1)(n-1)} - \frac{n}{(n-1)} s_i.$$

Observation 3: Consider a possible mutant player i in a population in which all other players $j \neq i$ have preference parameters $(a^E, s^E) \in \mathcal{P}$. By Observation 1, choices of players $j \neq i$ are given by $\mathbf{x}_{-i} = \mathbf{x}_{-i}^E$, independent of the actual preference parameters of player i. Given this behavior of other players, consider player i and his material payoff. Theorem 1 revealed that the effort choice that maximizes i's evolutionary fitness in this case is $x_i = x^E$. Accordingly, the set of preferences (a_i, s_i) that maximizes i's material payoff in the preference domain is equal to the set of preferences (a_i, s_i) that induce $x_i = x^E$ given $x_j = x^E$ for all other j. Observations 1 and 2 revealed that all $(a_i, s_i) \in \mathcal{P}$ induce this effort choice. This concludes the proof. ■

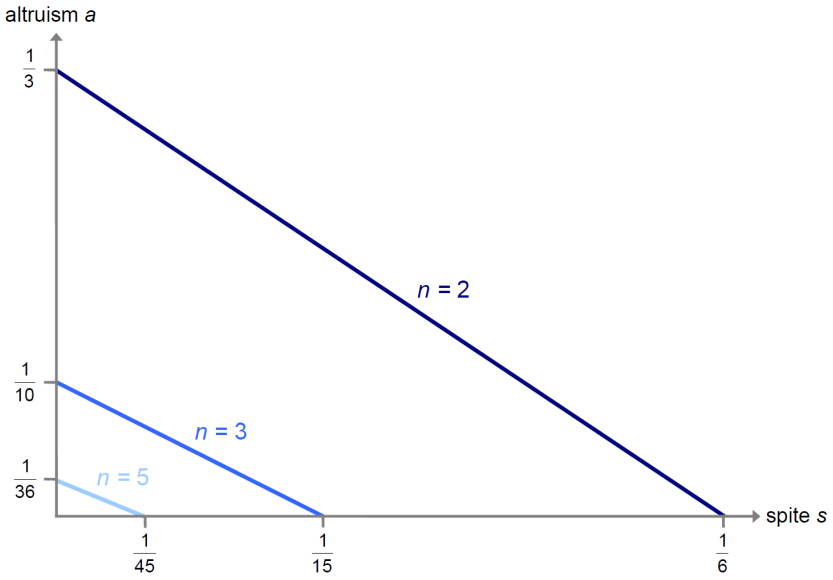

Figure 1: In-group altruism and out-group spite are substitutes.

Theorem 2 is our main result. It shows that there is a whole set of combinations of preference parameters that, if all individuals have these preferences, the symmetric Nash equilibrium with robust beliefs induced by these subjective utilities has effort choices that are the same as the evolutionarily stable equilibrium strategies that were characterized in Theorem 1. The function

$$a^E(s^E) = \frac{1}{(2n-1)(n-1)} - \frac{n}{(n-1)}s^E \tag{10}$$

describes the set of pairs (a^E, s^E) that constitute evolutionarily stable preferences. Evolutionary stability therefore allows for a number of combinations of in-group altruism and spiteful attitudes towards the members of the out-group. The function (10) also shows that more in-group altruism comes together with less spiteful behavior towards the out-group: in-group altruism and spite towards members of the out-group are substitutes as regards $a^E(s^E)$. Furthermore, as the group size n becomes large, in-group altruism and out-group spite become less pronounced. Figure 1 illustrates this relationship for $n = 2$, $n = 3$, and $n = 5$, with the weight for altruism a^E in (7) on the $y-$axis, and the weight s^E measuring spite in (7) on the $x-$axis.

The intuition for this negative relationship is as follows. Starting from a situation along the frontier $a^E(s^E)$, if a player has higher altruism towards the members of his group, in isolation this induces the player to make higher contributions to group effort. In comparison to the evolutionarily stable

effort choices in Theorem 1, this effort choice is "too high". In order to bring the effort that maximizes the player's subjective utility back in line with the effort level that is characterized in Theorem 1, an appropriate reduction in spiteful behavior towards the members of the rival group is desirable, as a reduction in s will reduce the effort that maximizes the player's subjective utility.

It is important to note that the evolutionary stability of in-group altruism and out-group spite in Theorem 2 has a different, and new reason, compared to the approach taken by Frank (1987), Bester and Güth (1998) and others. That a player is an altruist in this framework does not induce a behavioral reaction by other players that benefit the altruist. This 'strategic effect' channel is the basis for evolutionarily stable behavioral attitudes in many other analyses, but this channel is closed in our framework by the assumption of robust beliefs. Whether a player is an altruist or not does not affect the effort choices of all other players if they have robust beliefs. Given their beliefs, all other players are essentially guided only by their own preference parameters. Nevertheless, in the evolutionary equilibrium, the beliefs about other players' types and about their effort choices are perfectly consistent.

Out of the evolutionary equilibrium, with robust beliefs, players may have the wrong perceptions about the types of their co-players and need not anticipate their effort choices correctly. This emphasizes the absence of a strategic effect of the 'type'. A mutant who enters into an otherwise homogeneous population does not induce them to make different choices. An inconsistency of this concept seemingly is that the mutant has 'wrong' beliefs about the preferences of the population which he tries to invade and does not correctly anticipate other players' effort choices in this case. But as the state game is a single shot game, there is no way the mutant can learn, update beliefs and adjust behavior. Moreover, for the evolutionary stability of elements of the preference set $\mathcal{P}$, the specific belief of the mutant about the other players' types is not important. This can be confirmed as follows. Suppose that, unlike in the proof of Theorem 2, the mutant with preference parameters (a, s) assumes that all other players are of preference types (a^E, s^E), and as mutations are extremely rare, these other players believe that all players in N are of preference type (a^E, s^E). In this case it turns out that again the set of preference parameters in the set $\mathcal{P}$ constitute the set of preferences that fulfill the criterion of evolutionarily stable preferences in Definition 3.

5 Conclusions

In this paper we show that in-group altruism together with out-group spite can be explained as being the preference parameters of evolutionarily stable subjective utility in a framework with two groups which fight with each other. This result provides an evolutionary explanation for the strong in-group favoritism that is empirically well established for groups that are in conflict with other groups. We have also seen that spite and altruism are substitutes in the functional relationship that describes the full set of evolutionarily stable preference types, and that the role of altruism and spite is more important the smaller the groups are. For very large groups the amount of spite and altruism that is evolutionarily stable converges to zero. These comparative static properties about the role of group size yield an empirically testable hypothesis about in-group favoritism.

In order to address unobservability of types in an evolutionary context we introduced a new concept of belief types: robust beliefs. This concept is compatible with a stochastic and unobserved change in the distribution of types in the population, and still allows for the beliefs to be consistent in the equilibrium of evolutionarily stable strategies.

We note, however, that our results on evolutionary stability are not driven by this framework of beliefs. It is also important to note that the evolutionary argument that supports in-group altruism and out-group spite here is different from some of the arguments that have been used to provide an evolutionary foundation for altruism between individuals. First, the foundation here does not depend on considerations of kin-selection, or even of group-selection. To demonstrate this, we may allow for a complete re-grouping of the members (or their descendants) of the two groups between one state game and the next, without altering the analysis or the results at all. Second the argument does not build on strategic behavioral effects that might emerge if preference types are observable by others. This channel by which altruism and other types of other-regarding preferences have been established previously is strictly closed here by the unobservability of types.

References

[1] Abreu, Dilip, and Rajiv Sethi, 2003, Evolutionary stability in a reputational model of bargaining, *Games and Economic Behavior*, 44(2), 195-216.

[2] Baik, Kyung Hwan, In-Gyu Kim, and Sunghyun Na, 2001, Bidding for a group-specific public-good prize, *Journal of Public Economics*, 82(3),

415-429.

[3] Bernhard, Helen, Urs Fischbacher, and Ernst Fehr, 2006, Parochial altruism in humans, *Nature*, 442(7105), 912-915.

[4] Bester, Helmut, and Werner Güth, 1998, Is altruism evolutionarily stable?, *Journal of Economic Behavior & Organization*, 34(2), 193-209.

[5] Bowles, Samuel, 2009, Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? *Science*, 324(5932), 1293-1298.

[6] Brewer, Marilynn B., 1979, In-group bias in the minimal intergroup situation: A cognitive-motivational analysis, *Psychological Bulletin*, 86(2), 307-324.

[7] Davis, Douglas D., and Robert J. Reilly, 1999, Rent-seeking with non-identical sharing rules: An equilibrium rescued, *Public Choice*, 100(1-2), 31-38.

[8] Dekel, Eddie, Jeffrey C. Ely, and Okan Yilankaya, 2007, Evolution of preferences, *Review of Economic Studies*, 74(3), 685-704.

[9] Dufwenberg, Martin, and Werner Güth, 2000, Why do you hate me? On the survival of spite, *Economics Letters*, 67(2), 147-152.

[10] Eaton, B. Curtis, and Mukesh Eswaran, 2003, The evolution of preferences and competition: A rationalization of Veblen's theory of invidious comparisons, *Canadian Journal of Economics*, 36(4), 832-859.

[11] Eaton, B. Curtis, Mukesh Eswaran, and Robert J. Oxoby, 2011, Us and them: The origin of identity, and its economic implications, forthcoming in: *Canadian Journal of Economics*.

[12] Esteban, Joan M., and Debraj Ray, 2001, Collective action and the group size paradox, *American Political Science Review*, 95(3), 663-672.

[13] Fletcher, Jeffrey A., and Michael Doebeli, 2006, How altruism evolves: Assortment and synergy, *Journal of Evolutionary Biology*, 19(5), 1389-1393.

[14] Fletcher, Jeffrey A., and Michael Doebeli, 2009, A simple and general explanation for the evolution of altruism, *Proceedings of the Royal Society B - Biological Sciences*, 276 (1654), 13-19.

[15] Frank, Robert H., 1987, If homo economicus could choose his own utility function, would he want one with a conscience, *American Economic Review*, 77(4), 593-604.

[16] Gardner, Andy, and Stuart A. West, 2009, Greenbeards, *Evolution*, 64(1), 25-38.

[17] Güth, Werner, and Menahem E. Yaari, 1992, Explaining reciprocal behavior in simple strategic games: An evolutionary approach, in: Ulrich Witt, ed., Explaining Process and Change, Approaches to Evolutionary Economics, The University of Michigan Press, Ann Arbor, 23-34.

[18] Huck, Steffen, and Joerg Oechssler, 1999, The indirect evolutionary approach to explaining fair allocations, *Games and Economic Behavior*, 28(1), 13-24.

[19] James, Patrick, 1987, Conflict and cohesion: A review of the literature and recommendations for future research, *Cooperation and Conflict*, 22(1), 21-33.

[20] Katz, Eliakim, Shmuel Nitzan, and Jacob Rosenberg, 1990, Rent-seeking for pure public goods, *Public Choice*, 65(1), 49-60.

[21] Katz, Eliakim, and Julia Tokatlidu, 1996, Group competition for rents, *European Journal of Political Economy*, 12(4), 599-607.

[22] Koçkesen, Levent, Efe A. Ok, and Rajiv Sethi, 2000, The strategic advantage of negatively independent preferences, *Journal of Economic Theory*, 92(2), 274-299.

[23] Konrad, Kai A., 2004, Altruism and envy in contests: An evolutionarily stable symbiosis, *Social Choice and Welfare*, 22(3), 479-490.

[24] Konrad, Kai A., 2009, Strategy and Dynamics in Contests, Oxford University Press 2009.

[25] Lehmann, L., and L. Keller, 2006, The evolution of cooperation and altruism - a general framework and a classification of models, *Journal of Evolutionary Biology*, 19(5), 1365-1376.

[26] Leininger, Wolfgang, 2003, On evolutionarily stable behavior in contests, *Economics of Governance*, 4(3), 177-186.

[27] Leininger, Wolfgang, 2009, Evolutionarily stable preferences in contests, *Public Choice*, 140(3-4), 341-356.

[28] Lewis, Gary J., and Timothy C. Bates, 2010, Genetic evidence for multiple biological mechanisms underlying in-group favoritism, *Psychological Science*, 21(11), 1623-1628.

[29] Marshall, James A.R., 2011, Ultimate causes and the evolution of altruism, *Behavioral Ecology and Sociobiology*, 65(3), 503-512.

[30] Mathur, Vani A., Tokiko Harada, Trixie Lipke, and Yoan Y. Chiao, 2010, Neural basis of extraordinary empathy and altruistic motivation, *Neuroimage*, 51(4), 1468-1475.

[31] Maynard Smith, John, 1998, The origin of altruism, *Nature*, 393(6686), 639-640.

[32] Mifune, Nobuhiro, Hirofumi Hashimoto, and Toshio Yamagishi, 2010, Altruism toward in-group members as a reputation mechanism, *Evolution and Human Behavior*, 31(2), 109-117.

[33] Nitzan, Shmuel, 1991, Collective rent dissipation, *Economic Journal*, 101(409), 1522-1534.

[34] Olson, Mancur Jr., and Richard J. Zeckhauser, 1966, An economic theory of alliances, *Review of Economics and Statistics*, 48(3), 266-279.

[35] Ok, Efe A., and Fernando Vega-Redondo, 2001, On the evolution of individualistic preferences: an incomplete information scenario, *Journal of Economic Theory*, 97(2), 231-254.

[36] Pinter, Brad, and Anthony G. Greenwald, 2011, A comparison of minimal group induction procedures, *Group Processes & Intergroup Relations*, 14(1), 81-98.

[37] Reeve, Hudson Kern, 2000, Unto others: The evolution and psychology of unselfish behavior, *Evolution and Human Behavior*, 21(1), 65-72.

[38] Salomonsson, Marcus, 2010, Group selection: The quest for social preferences, *Journal of Theoretical Biology*, 264(3), 737-746.

[39] Schaffer, Mark E., 1988, Evolutionary stable strategies for a finite population and a variable contest size, *Journal of Theoretical Biology*, 132(4), 469-478.

[40] Sethi, Rajiv, 1996, Evolutionary stability and social norms, *Journal of Economic Behavior and Organization*, 29(1), 113-140.

[41] Sherif, Muzafer, O.J. Harvey, B. Jack White, William R. Hood, and Carolyn W. Sherif, 1961, The Robbers Cave Experiment: Intergroup Conflict and Cooperation, Norman, Oklahoma, University Book Exchange.

[42] Skaperdas, Stergios, 1996, Contest success functions, *Economic Theory*, 7(2), 283-290.

[43] Smirnov, Oleg, Holly Arrow, Douglas Kennett, and John Orbell, 2007, Ancestral war and the evolutionary origins of "heroism", *Journal of Politics*, 69(4), 927-940.

[44] Sober, Elliott, and David Sloan Wilson, 1998, Unto Others: The Evolution and Psychology of Unselfish Behavior, Harvard University Press.

[45] Stein, Arthur A., 1976, Conflict and cohesion, a review of the literature, *Journal of Conflict Resolution*, 20(1), 143-172.

[46] Tajfel, Henri, 1982, Social psychology of intergroup relations, *Annual Review of Psychology*, 33, 1-39.

[47] Tajfel, Henri, and John Turner, 1979, An integrative theory of intergroup conflict, in: W. Austin and S. Worchel (ed.), The Social Psychology of Intergroup Relations, Monterey, CA., Brooks/Cole Publ., 33-48.

[48] Tullock, Gordon, 1980, Efficient rent seeking, in: James Buchanan, Roger Tollison, and Gordon Tullock (eds.), Toward a Theory of the Rent-Seeking Society, Texas A&M University Press, College Station, 97-112.

[49] Ursprung, Heinrich W., 1990, Public goods, rent dissipation, and candidate competition, *Economics & Politics*, 2, 115-132.

[50] Wärneryd, Karl, 1998, Distributional conflict and jurisdictional organization, *Journal of Public Economics*, 69(3), 435-450.

[51] West, Stuart A., and Andy Gardner, 2010, Altruism, spite and greenbeards, *Science*, 327, 1341-1344.