Economics Letters 254 (2025) 112437

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Rating rules in the helping game: An axiomatic approach

Andrea Marietta Leina [a,b,*], Amrish Patel [b], Robert Sugden [b], Theodore L. Turocy [b]

[a] University of Verona, Italy
[b] University of East Anglia, UK

ARTICLE INFO

JEL classification:
C7
D8

Keywords:
Helping game
Indirect reciprocity
Reputation
Axiomatic approach

ABSTRACT

We examine a stylised helping game in which players recurrently decide whether to help others at personal cost and are assigned binary ratings of "helpfulness" based on previous choices. We propose axioms requiring that ratings are responsive to players' decisions and change monotonically with respect to helping given or withheld. Only four rules satisfy these axioms: two standing rules and two versions of a form of binary image scoring. These results show how a single rating can encode both the "desert" perspective, linking worthiness to kindness and deservingness, and the "club" perspective, relating worthiness to a cooperative club's membership.

1. Introduction

In human and animal societies, there are various dyadic interactions that can be modelled as *helping games* — games where one player chooses whether to incur a cost to confer a benefit on the other. When helping games are played recurrently by pairs of individuals drawn randomly from large populations, cooperation may be sustained by *indirect reciprocity* (Alexander, 1987), i.e., helping coplayers who have previously helped others. For this to work, there must be some reputational mechanism encoding each player's past "helpfulness" in a form accessible to others, and individuals must be motivated to differentially help "helpful" coplayers.

A large literature in economics and biology assumes self-interested motivations and investigates the effectiveness of different reputational mechanisms in inducing cooperation. There has been particular interest in two classes of *rating rules* (i.e., rules for encoding helpfulness on a one-dimensional scale). *Image scoring* (Nowak and Sigmund, 1998) records the difference between the number of previous games in which a player chose "help" and the number in which they chose "not help". *Standing* (Sugden, 1986; Leimar and Hammerstein, 2001) records a binary rating of "good" or "bad" that is updated recursively and takes account of the ratings of the players who were helped or not helped. Ohtsuki and Iwasa (2004) identify evolutionarily stable strategies induced by each of the 256 possible binary rating rules and calculate measures of their efficiency for given parameters. Eight rules (the "leading eight"), including two versions of standing, perform particularly well.

A complementary approach in behavioural economics assumes that individuals may have non-self-interested preferences for reciprocity. Given this approach, a rating can be interpreted as a norm for judging a player's help-worthiness. Such norms may support cooperation even when helping is contrary to self-interest. Our objective is to formulate principles that express norms of worthiness and examine their implications for binary rating rules.

We distinguish between two perspectives on "worthiness", corresponding with different conceptions of reciprocity. In the social preference literature, preferences for reciprocity are often represented as preferences for rewarding kind intentions and punishing unkind intentions (Rabin, 1993; Levine, 1998). If a "good" rating signals a disposition towards kindness, preferences for reciprocity can motivate players to differentially help "good" coplayers (Engelmann and Fischbacher, 2009). This is the *desert* perspective on worthiness. In the contrasting *club* perspective, reciprocity is understood as a disposition to cooperate with others in the achievement of mutual benefit (Sugden, 2018; Isoni et al., 2023). A set of players helping one another can be interpreted as a cooperative club, and a "good" rating can signal club membership. Worthiness is internal to the club — an entitlement to help from other members.[1]

We propose a set of axioms that impose restrictions on binary rating rules. Each axiom can be justified in terms of either perspective on worthiness. We show that these axioms restrict the 256 possible rules to only four.

* Corresponding author at: University of Verona, Italy.
E-mail addresses: andrea.mariettaleina@gmail.com (A. Marietta Leina), Amrish.Patel@uea.ac.uk (A. Patel), R.Sugden@uea.ac.uk (R. Sugden), T.Turocy@uea.ac.uk (T.L. Turocy).

[1] Our definition of a rating rule does not allow non-active players' ratings to change according to whether they are helped. We bracket out a third perspective on worthiness, related to gift exchange, in which a player's rating represents a "credit balance" in their record of helping and being helped (Bigoni et al., 2020).

https://doi.org/10.1016/j.econlet.2025.112437
Received 8 May 2025; Received in revised form 4 June 2025; Accepted 5 June 2025
Available online 18 June 2025

2. Definition

Let N denote the set of players. The game unfolds over a sequence of discrete time periods T. There is a set of two ratings $R = \{0, 1\}$. In each period $t \in T$, each player i has a rating $r_i^t \in R$.

In each period t, some (non-overlapping) pairs of players are randomly matched. Within each match, one player is assigned the *active* role, the other player is *non-active*. Each player in N has some positive probability of being active in some matched pair, and some positive probability of being non-active in some matched pair. In each pair, the active player chooses an *action* from $A = \{h, n\}$, interpreted as *help* and *not help*, respectively. Relative to n, h benefits the non-active player but is costly to the active player. The non-active player does not make a choice.

Consider a given match at period t, between an active player i and a non-active player $j \neq i$. Let $a_i^t \in A$ denote the active player's action in that period. The evolution of the active player's rating is governed by a *rating rule* γ: $R \times R \times A \to R$. This function is common knowledge among all players and determines the active player's rating in the next period,

$$r_i^{t+1} = \gamma(r_i^t, r_j^t, a_i^t) \tag{1}$$

The non-active player's rating does not change, $r_j^{t+1} = r_j^t$.

Table 1 presents a compact representation of the rating rule as a 2×2 matrix, where the rows correspond to the active player's current rating r_i, and the columns correspond to the non-active player's current rating r_j. Each cell corresponds to a *scenario* (r_i, r_j), and contains a pair of values expressing the active player's next-period rating as a partial function of a in that scenario.

Table 1
Representation of the rating rule.

		Non-active player rating	
		$r_j = 1$	$r_j = 0$
Active player rating	$r_i = 1$	$\gamma(1,1,h)$, $\gamma(1,1,n)$	$\gamma(1,0,h)$, $\gamma(1,0,n)$
	$r_i = 0$	$\gamma(0,1,h)$, $\gamma(0,1,n)$	$\gamma(0,0,h)$, $\gamma(0,0,n)$

There are $2^3 = 8$ possible combinations of (r_i, r_j, a). Each combination can be mapped to either 0 or 1 in R, leading to $2^8 = 256$ possible rating rules. Let Γ denote the set of all these rules.

3. An axiomatic framework

We propose six axioms that capture desirable features of binary rating rules. The following two *generic* axioms treat h and n symmetrically.

Axiom G1 (*Own Rating Preservation*).
$\forall r_i, r_j \in R, \exists a \in A : \gamma(r_i, r_j, a) = r_i$
In every scenario, there is at least one action that preserves the active player's current rating.

Axiom G2 (*Action Consequentiality*).
$\forall r_i \in R, \exists r_j \in R : \gamma(r_i, r_j, h) \neq \gamma(r_i, r_j, n)$
For each rating of the active player, there is at least one scenario in which that player can change their rating by an act of choice.

In the desert perspective, G1 and G2 ensure that a person's moral status can change only as a result of their own chosen action, and that such changes, in either direction, always remain possible. In the club perspective, these axioms ensure that club membership is voluntary and open to all.

The following four *monotonicity* axioms relate a player's ratings to their decisions, encoding the 0 and 1 ratings respectively as "bad" and "good". In the desert perspective, "good" indicates the moral worthiness associated with reciprocal helping; in the club perspective, it indicates membership of a reciprocal helpers' club.

Axiom M1 (*Action Monotonicity*).
(a) $\forall r_i \in R, \gamma(r_i, 0, n) = 1 \implies \gamma(r_i, 0, h) = 1$
(b) $\forall r_i \in R, \gamma(r_i, 1, n) = 1 \implies \gamma(r_i, 1, h) = 1$
Helping the non-active player is always rated at least as highly as not helping them. This is the case whether the non-active player is "bad" (M1a) *or "good"* (M1b).

We note that M1a is inconsistent with the thought that not helping a "bad" player is a morally appropriate act of punishment, and that failure to punish in such a case is thereby deserving of second-order punishment. Clearly, such second-order punishment can help to stabilise an equilibrium in which self-interested players almost always cooperate (Ohtsuki and Iwasa, 2004). But if indirect reciprocity is based on a reliable reputation mechanism, the moral concept of "punishment" is arguably out of place. "Bad" players are not free-riders: they have merely chosen to opt out of a practice of reciprocal help.

Axiom M2 (*Coplayer Rating Monotonicity*).
(a) $\forall r_i \in R, \gamma(r_i, 0, h) = 1 \implies \gamma(r_i, 1, h) = 1$
(b) $\forall r_i \in R, \gamma(r_i, 1, n) = 1 \implies \gamma(r_i, 0, n) = 1$
Helping a "good" player is always rated at least as highly as helping a "bad" player (M2a), *and not helping a "good" player is always rated at least as lowly as not helping a "bad" player* (M2b).

Uniquely in the set of monotonicity axioms, M1b turns out to be redundant.

Lemma 1. *If $\gamma \in \Gamma$ satisfies M1a, M2a, and M2b then it satisfies M1b.*

Proof. Fix $\gamma \in \Gamma$ and $r_i \in R$.
Suppose $\gamma(r_i, 1, n) = 1$. $\gamma(r_i, 1, n) = 1 \overset{\text{M2b}}{\implies} \gamma(r_i, 0, n) = 1 \overset{\text{M1a}}{\implies} \gamma(r_i, 0, h) = 1 \overset{\text{M2a}}{\implies} \gamma(r_i, 1, h) = 1$. □

We defer the proof that none of M1a, M2a or M2b is implied by the other three monotonicity axioms.

Theorem 1. *Let $\Gamma^\star \subseteq \Gamma$ denote the set of rating rules that satisfy G1, G2, M1a, M2a and M2b. $\Gamma^\star$ contains exactly the four rating rules in Fig. 1.*

(Rule 1) Leimar-Hammerstein good standing

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 0	1, 1
$r_i = 0$	1, 0	1, 0

(Rule 2) Sugden good standing

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 0	1, 1
$r_i = 0$	1, 0	0, 0

(Rule 3) Binary image scoring

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 0	1, 0
$r_i = 0$	1, 0	1, 0

(Rule 4) Modified binary image scoring

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 0	1, 0
$r_i = 0$	1, 0	0, 0

Fig. 1. $\Gamma^\star$, the rating rules that satisfy Axioms G1, G2, M1a, M2a, and M2b.

(a) Violates only G1

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 1	0, 1
$r_i = 0$	1, 0	1, 0

(b) Violates only G2

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 1	1, 1
$r_i = 0$	1, 0	1, 0

(c) Violates only M1a

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 0	0, 1
$r_i = 0$	1, 0	1, 0

(d) Violates only M2a

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 0	1, 0
$r_i = 0$	0, 0	1, 0

(e) Violates only M2b

	$r_j = 1$	$r_j = 0$
$r_i = 1$	1, 1	1, 0
$r_i = 0$	1, 0	1, 0

Fig. 2. Necessity of each axiom in Theorem 1.

We state two Lemmas to organise the proof.

Lemma 2. *If $\gamma \in \Gamma$ satisfies M1a, M2a, M2b, and G2, then for all $r_i \in R$, $\gamma(r_i, 1, h) = 1$ and $\gamma(r_i, 1, n) = 0$.*

Proof. Fix $\gamma \in \Gamma$ and $r_i \in R$. We establish each part of the conjunction by contradiction.

Suppose $\gamma(r_i, 1, h) = 0$. $\gamma(r_i, 1, h) = 0 \overset{\text{M2a}}{\implies} \gamma(r_i, 0, h) = 0 \overset{\text{M1a}}{\implies} \gamma(r_i, 0, n) = 0 \overset{\text{M2b}}{\implies} \gamma(r_i, 1, n) = 0$. Therefore, $\gamma(r_i, r_j, a) = 0\ \forall r_j \in R, a \in A$, which contradicts G2.

Suppose $\gamma(r_i, 1, n) = 1$. $\gamma(r_i, 1, n) = 1 \overset{\text{M2b}}{\implies} \gamma(r_i, 0, n) = 1 \overset{\text{M1a}}{\implies} \gamma(r_i, 0, h) = 1 \overset{\text{M2a}}{\implies} \gamma(r_i, 1, h) = 1$. Therefore, $\gamma(r_i, r_j, a) = 1\ \forall r_j \in R, a \in A$, which contradicts G2. □

Lemma 3. *If $\gamma \in \Gamma$ satisfies M1a and G1, then $\gamma(1, 0, h) = 1$ and $\gamma(0, 0, n) = 0$.*

Proof. Fix $\gamma \in \Gamma$.

- $M1a \wedge G1 \implies (\gamma(1, 0, n) = 0 \vee \gamma(1, 0, h) = 1) \wedge (\gamma(1, 0, h) = 1 \vee \gamma(1, 0, n) = 1) \implies \gamma(1, 0, h) = 1$.
- $M1a \wedge G1 \implies (\gamma(0, 0, n) = 0 \vee \gamma(0, 0, h) = 1) \wedge (\gamma(0, 0, h) = 0 \vee \gamma(0, 0, n) = 0) \implies \gamma(0, 0, n) = 0$. □

Proof of Theorem 1. Lemmas 2 and 3 exclude all rules except the four shown in Fig. 1. It is straightforward to verify that each of these rules satisfy all the axioms. □

Rules 1 and 2 are the standing rules proposed respectively by Leimar and Hammerstein (2001) and Sugden (1986). Rule 3 is the binary form of the image scoring rule proposed by Nowak and Sigmund (1998), and Rule 4 is a variant of that rule. We interpret Theorem 1 as demonstrating a close normative relationship between the two types of rule, a relationship that holds whether reciprocity is understood in the "desert" or "club" sense.

Notice that the axioms fix the value of γ for all cases except $\gamma(1, 0, n)$ and $\gamma(0, 0, h)$. The two image scoring rules differ from the two standing rules with respect to $\gamma(1, 0, n)$, i.e., whether "good" players retain that status if they fail to help "bad" players (the standing rules) or lose it (the image scoring rules). Rules 1 and 2 are members of Ohtsuki and Iwasa's "leading eight" but, because of this difference, Rules 3 and 4 are not. In the club perspective, $\gamma(1, 0, n) = 1$ (as in standing rules) has an obvious normative justification: the obligation of club members is to help one another. In the desert perspective, however, failure to help a "bad" player can be viewed as morally unworthy.

Rules 1 and 3 differ respectively from Rules 2 and 4 with respect to $\gamma(0, 0, h)$, i.e., whether a "bad" player who helps a "bad" player retains their "bad" rating (Rules 2 and 4) or achieves a "good" rating (Rules 1 and 3). In this respect, Rules 2 and 4 give players stronger incentives to achieve and maintain "good" ratings, but Rules 1 and 3 make it easier for players to regain "good" ratings after mistakes.

Fig. 2 demonstrates the necessity of each axiom in the antecedent of Theorem 1 by exhibiting a rating rule that contravenes only that axiom. Each of these rules also satisfies M1b, establishing that none of M1a, M2a or M2b is implied by the other three monotonicity axioms.

Acknowledgements

This work was supported by the University of East Anglia, School of Economics, United Kingdom and the Centre for Behavioural and Experimental Social Science (CBESS). We thank the reviewer for their helpful suggestions. Additionally, Andrea Marietta Leina thanks Héctor Hermida-Rivera and Rui Silva for useful discussions.

Data availability

No data was used for the research described in the article.

References

Alexander, R.D., 1987. The Biology of Moral Systems. Aldine de Gruyter, New York, URL: https://doi.org/10.4324/9780203700976.

Bigoni, M., Camera, G., Casari, M., 2020. Money is more than memory. J. Monet. Econ. 110, 99–115, URL: https://www.sciencedirect.com/science/article/pii/S0304393219300054.

Engelmann, D., Fischbacher, U., 2009. Indirect reciprocity and strategic reputation building in an experimental helping game. Games Econom. Behav. 67 (2), 399–407, URL: https://www.sciencedirect.com/science/article/pii/S0899825609000025.

Isoni, A., Sugden, R., Zheng, J., 2023. Voluntary interaction and the principle of mutual benefit. J. Political Econ. 131 (6), 1576–1616, URL: https://www.journals.uchicago.edu/doi/abs/10.1086/722930.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. Proc. R. Soc. Lond. B 268, 745–753, URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2000.1573.

Levine, D., 1998. Modeling altruism and spitefulness in experiment. Rev. Econ. Dyn. 1 (3), 593–622, URL: https://www.sciencedirect.com/science/article/pii/S1094202598900230.

Nowak, M., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature 393, 573–577, URL: https://www.nature.com/articles/31225.

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. J. Theoret. Biol. 231 (1), 107–120, URL: https://www.sciencedirect.com/science/article/pii/S0022519304002772.

Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83 (5), 1281–1302, URL: http://www.jstor.org/stable/2117561.

Sugden, R., 1986. The Economics of Rights, Co-operation and Welfare. Basil Blackwell, Oxford, UK, URL: https://link.springer.com/book/10.1057/9780230536791.

Sugden, R., 2018. The Community of Advantage: A Behavioural Economist's Defence of the Market. Oxford University Press, URL: https://doi.org/10.1093/oso/9780198825142.001.0001.