On the External Validity of Social-Preference Games: A Systematic Lab-Field Study

Matteo M. Galizzi[a], Daniel Navarro-Martinez[b]

[a]London School of Economics and Political Science (Email: matteo.m.galizzi@lse.ac.uk)
[b]Pompeu Fabra University and Barcelona Graduate School of Economics (Email: daniel.navarro@upf.edu)

January 1, 2015

Abstract: We present a lab-field experiment designed to assess systematically the external validity of social preferences elicited in a variety of experimental games. We do this by comparing behavior in the different games with a number of behaviors elicited in the field and with self-reported behaviors exhibited in the past, using the same sample of participants. Our results show that the experimental social-preference games do a poor job in explaining both social behaviors in the field and social behaviors from the past.

Keywords: Social preferences, experimental games, external validity, field behavior.

JEL classification: C92, C93, D03.

Acknowledgments: We are very grateful to Paul Dolan for his encouragement and invaluable support when we embarked on this project at the LSE. We thank Pablo Branas-Garza, Antonio Cabrales, Juan Camilo Cardenas, Susann Fiedler, Antonio Filippin, Urs Fischbacher, Andreas Glockner, Daniela Grieco, Benedikt Herrmann, Heather Kappes, John List, Sara Machado, Robert Metcalfe, Rosemarie Nagel, Chris Olivola, Giovanni Ponti, Wojtek Przepiorka, Jan Stoop, Roberto Weber, and the participants in the ESA conference in Prague, the SPUDM conference at IESE Business School, the BDRM Conference at London Business School, the FUR Conference at Erasmus University Rotterdam, and the London Experimental Workshop at Middesex University for helpful comments and suggestions. Matteo M. Galizzi gratefully acknowledges funding from the ESRC (Future Research Leader Fellowship ES/K001965/1); Daniel Navarro-Martinez acknowledges support from the Spanish Ministry of Economy and Competitiveness (MINECO-PSI2013-41909-P).

1. Introduction

The last couple of decades have seen a strong surge of interest in what is now widely known in economics as 'social preferences'. While the study of social behaviors has a long tradition in disciplines like economics (e.g., Smith, 1759), psychology (e.g., Triplett, 1898; Lewin, 1939) or sociology (e.g., Durkheim, 1893), in recent times, the term 'social preferences' has come to be associated with a more specific program of research originating mainly in experimental and behavioral economics (see, e.g., Guth et al., 1982; Andreoni, 1988; Forsythe et al., 1994; Camerer and Thaler, 1995; Fehr and Gachter, 2000, 2002; Charness and Rabin, 2002; Fischbacher and Gachter, 2010).

A key feature of this research program is that it has focused largely on the study of behavior in experimental games designed to target different aspects of social behavior, such as *altruism* (e.g., Forsythe et al., 1994; Andreoni and Miller, 2002), *reciprocity* (e.g., Berg et al., 1995; Cox, 2004) or *trust* (e.g., Berg et al., 1995; Ortmann et al., 2000). In typical studies, people play these games in laboratory settings, where special care is taken in stripping the games from contextual features that depart from the underlying game-theoretic structures on which they are based, and in providing real monetary incentives that are aligned with the payoffs of the games. This stylized approach has arguably become one of the building blocks of experimental and behavioral economics, with literally thousands of studies published on the topic, some of which are among the most widely cited papers in leading journals (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Fehr and Gachter, 2000, 2002; Charness and Rabin, 2002; Herrmann et al., 2008).[1]

Given such a major interest in the topic, it is surprising how little work has been done so far to investigate systematically the *external validity* of this experimental games approach to social preferences. This seems to us one of the most fundamental questions yet to be answered about the social-preference paradigm. Specifically, to what extent do the principles determining behavior in experimental social-preference games tap into the principles governing social behavior when it is put in context or taken outside the lab? Not addressing this question in a systematic way could put social-preference research at risk of becoming research on how people play certain games in the lab, instead of research on how people behave in social situations of broader interest to economics and other social and behavioral sciences.

A few researchers have warned before us about potential issues of external validity in research on social preferences (see Levitt and List, 2007a,b, 2008; List, 2009). In a particularly prominent paper, Levitt and List (2007a) discuss six potential complications arising when the findings of social-preference experiments are extrapolated outside the lab: (i) participants in the lab act under the scrutiny of the experimenters; (ii) their decisions and

[1] See Camerer (2003) for a comprehensive review of experimental social-preference games, and Zelmer (2003), Oosterbeek et al. (2004), Engel (2011), and Johnson and Mislin (2011), for more specific reviews on the Public Good, Ultimatum, Dictator, and Trust game, respectively.

actions are unlikely to remain anonymous; (iii) the context matters and cannot be completely controlled by the experimenters; (iv) the stakes are different from the ones in real life; (v) the participants in experiments differ from the groups of people engaged in most real-world behaviors; and (vi) there are artificial restrictions on choice sets and time horizons.

Some of the points raised by Levitt and List can actually be interpreted as general limitations of laboratory experimentation compared to field settings, and they have in fact initiated a broad-ranging methodological debate on the scope and limitations of laboratory experiments in economics (see also Falk and Heckman, 2009; Camerer, 2011). In that sense, it is important to clarify that our focus here is not on the external validity of laboratory experimentation as a whole, which is very diverse and, in our view, not only useful but also necessary in the social and behavioral sciences. We center exclusively on the more specific issue of the external validity of experimental social-preference games.

There has indeed been extensive research on some of the complications identified by Levitt and List in the realm of social preferences, including studies on the effects of anonymity and scrutiny (Hoffman et al., 1994, 1996; Eckel and Grossman, 1996; Dana et al., 2006; Dana et al., 2007; List, 2007; Bardsley, 2008; Franzen and Pointer, 2012; Winking and Mizer, 2013); the context and framing (Cherry et al., 2002; List, 2006; Branas-Garza, 2007; Stoop et al., 2012; Stoop, 2013); the size of the stakes (Slonim and Roth, 1998; Cameron, 1999; Munier and Zaharia, 2003; Carpenter et al., 2005; List and Cherry, 2008); the subject pool (List, 2004, 2006; Carpenter and Seki, 2005; Bellemare and Kroger, 2007; Carpenter et al., 2008; Garbarino and Slonim, 2009; Stoop et al., 2012; Cleave et al., 2013; Exadaktylos et al., 2013; Stoop, 2013); or the self-selection into lab experiments (Krawczyk, 2011; Falk et al., 2013; Slonim et al. 2013; Abeler and Nosenzo, forthcoming).

All these factors have been shown to matter, at least in some cases, which casts doubts on the idea that behavior in experimental social-preference games can be immediately representative of social behavior outside the lab. The role of the context is perhaps especially problematic, given that typical social-preference games are meant to be as context-free as possible, while much research in experimental economics and psychology has shown that preferences seem to be significantly shaped by the context in which they are elicited (see, e.g., Slovic, 1995; Loewenstein, 1999; Lichtenstein and Slovic, 2006; Ariely et al., 2006; Stewart et al., forthcoming).

Closer related to the research presented here are a relatively small number of empirical studies that have looked at the question of the external validity of experimental social-preference games by directly linking evidence from the lab and the field for the same pool of subjects (Glaser et al., 2000; Castillo and Carter, 2002; Fehr et al., 2003; Cardenas and Carpenter, 2005; Holm and Danielson, 2005; Karlan, 2005; Ruffle and Sosis, 2007; Benz and Meier, 2008; Laury and Taylor, 2008; Barr and Serneels, 2009; Barr and Zeitlin, 2010; Baran, Sapienza, and Zingales, 2010; Carpenter and Myers, 2010; Englmaier and Gebhardt, 2010; Rustagi et al., 2010; Serra et al., 2010; Carpenter and Seki, 2011; de Oliveira et al., 2011; Fehr and Leibbrandt, 2011; Lamba and Mace, 2011; Voors et al., 2011, 2012;

Leibbrandt, 2012; Cardenas et al., 2013; Franzen and Pointer, 2013; Bernold et al. 2014).[2] Table 1 provides a synoptic summary of these *lab-field* studies, ordered chronologically.[3]

As Table 1 shows, the evidence accumulated so far is somewhat mixed. Some studies have found significant correlations between behavior in particular experimental games and specific field behaviors (Englmaier and Gebhardt, 2010; Rustagi et al., 2010; Serra et al., 2010; Carpenter and Seki, 2011; de Oliveira et al., 2011; Fehr and Leibbrandt, 2011; Leibbrandt, 2012); some others have found no correlation (Bouma et al., 2008; Lamba and Mace, 2011; Voors et al., 2011, 2012; Bernold et al. 2014); and many have obtained mixed findings (Glaeser et al., 2000; Castillo and Carter, 2002; Fehr et al., 2003; Cardenas and Carpenter, 2005; Holm and Danielson, 2005; Karlan, 2005; Ruffle and Sosis, 2007; Benz and Meier, 2008; Laury and Taylor, 2008; Barr and Serneels, 2009; Barr and Zeitlin, 2010; Carpenter and Myers, 2010; Cardenas et al., 2013; Franzen and Pointer, 2013).

In terms of the number of studies, the balance seems to be on the side of the ones reporting at least some significant correlation, but the interpretation of this is unclear. This imbalance could reflect the well-known bias to submit and publish positive results over negative ones, and could over-represent spurious correlations. In addition, the near impossibility to establish a strict one-to-one correspondence between context-free experimental games and field settings could sharpen the bias, making it difficult to justify negative results and further impeding their publication.

For these reasons, a major shortcoming of the evidence available to date is that most of the studies are not systematic. They typically report some relationship between one experimental game and one field behavior, leaving us wondering what a more systematic or comprehensive study of games against field situations would reveal.

In this paper, we present a more systematic investigation of the external validity of social-preference games, conducted by comparing behavior in a variety of games against a variety of situations created in the field, and also against self-reported social behaviors performed in the past, all using the same sample of participants. The different social-preference games included (dictator game, ultimatum game, trust game, and public good game) cover a large proportion of experimental research on social preferences; the five different field situations tap into different types of pro-social behaviors related to giving money and helping others; and the self-reported measures include various pro-social tendencies shown in the past.

Rather than trying to establish one-to-one correspondences between particular experimental games and specific field situations, which is necessarily imprecise given the context-free nature of the games, we adopt the strategy of covering a variety of prominent social-preference games and a variety of relevant field behaviors to explore more broadly the extent to which the games are predictive of social behaviors shown in the field. The

[2] Camerer (2011) provides a thorough review of this and other related literature on external validity.

[3] See also Dolan and Galizzi (forthcoming) for a more general discussion of lab-field experiments on pro-social behavior and in other areas.

relationship between the games we used and the field situations we studied is further discussed in the next sections. The self-reported measures of past social behaviors give us an additional layer to evaluate the explanatory ability of the games. To the best of our knowledge, this constitutes the most systematic and comprehensive study of the external validity of experimental social-preference games available to date.

Our results show that the social-preference games do a poor job in explaining both the field behaviors and the self-reports. In a nutshell, none of the behaviors elicited in the field or reported from the past that we analyzed seem to be explained to a significant extent by behavior in the experimental games. We do not claim that this single study can establish any firm or final conclusions about the complex issue of the external validity of social-preferences games. We do believe, however, that our results are worrying and they call for more, and more systematic, research on this important issue.

The rest of the paper is organized as follows: Section 2 describes the methods used; Section 3 presents the results obtained; Section 4 discusses the results and concludes.

2. Methods

Our general approach to investigate the external validity of social-preference games involved facing the same sample of participants with the following three elements: (i) a set of questions about social behaviors exhibited in the past; (ii) a variety of social-preference games played in the laboratory; and (iii) several naturalistic situations related to social preferences that we created in the field. The main aim of this design was to evaluate the social-preference games by comparing them against actual social behaviors in the field and against self-reported social behaviors from the past, all using the same individuals.

This *lab-field* set-up was organized so that each individual participated in three separate sessions on three different days of the same week. On the first day, the participants came into the lab to do different tasks (some of them unrelated to social preferences), which included the self-reported measures of past social behaviors that we are interested in here. On the second day, the same participants played various social-preference games in the lab. On the third day, they came again to the lab to do a task that was unrelated to social preferences, and when they went out they were faced with one of several field situations that we created, in which they could behave pro-socially.

This three-day lab-field structure allowed us to obtain all the information that we were interested in, while minimizing the possibilities of cross-contamination between the different tasks. We now explain each one of these three main components in turn.

2.1. Session 1: The self-reported measures of past social behaviors

In the first experimental session of the week, on their arrival to the lab, the participants were anonymously indentified by using individual ID codes. They were then asked to read an informed consent form and sign it if they agreed to carry on with the experiment. The form reiterated important information that they had already seen on the invitation email. Specifically, it said that: the experiment would require coming to the lab for three separate sessions on three different days of the week; each session would last about one hour; they would receive a fixed amount of £30 for their participation in all the three sessions (to be paid at the end of the last session); and they would have the opportunity to get an extra payment depending on their performance in the tasks they had to do. The participants were then randomly assigned to different cubicles in the lab. Throughout the session, they were given more specific instructions for the different tasks they had to do.[4]

In Session 1, the participants provided the self-reported measures of past pro-social behaviors that constitute one of our main variables. We elicited those measures by using the Self-Report Altruism (SRA) Scale (Rushton et al., 1981). This scale consists of 20 items, in which people are asked to state how frequently they have done in the past different actions related to pro-social behaviors. Three examples are: *"I have given money to a charity"*, *"I have helped carry a stranger's belongings (books, parcels, etc.)"*, and *"I have given money to a stranger who needed it (or asked me for it)"*. A full list of the 20 SRA items is contained in Appendix A. People had to rate each statement on a scale from 1 ("never") to 5 ("very often"). This constitutes our primary measure of past pro-social behaviors.

In addition to the SRA scale, in Session 1, the participants responded to other questionnaires (not part of the present study) and they did another unrelated task, which consisted in watching and rating several videos.

2.2. Session 2: The social-preference games

In Session 2, the subjects returned to the lab, were again assigned individual ID codes and randomly allocated to their cubicles, and then received more specific and detailed instructions for the tasks they had to do throughout the session. Because of the structure of some of the games played in the session, we needed the participant number to be a multiple of four. To this end, we allocated the remaining people to a separate task (conducted in a different room and unrelated to this study) for the rest of the session.

In this session, the subjects participated in seven different games (explained further down) that are widely used in economics and other social and behavioral sciences to study social preferences. All the games were one-shot (i.e., the subjects only played them once) and independent from each other. In each of the seven games, the participants were randomly

[4] All the instructions given to participants are available from the authors upon request.

matched (anonymously) with other participants in the session, under the constraint that they never interacted with the same person more than once. At the end, one of the seven games was randomly selected and all the participants were actually paid the amount they earned in that particular game. All the games were computerized, and they were programmed and implemented using Z-Tree (Fischbacher, 2007).

All the procedures were explained in detail to the participants. Specifically, they received general instructions on the seven-game structure and the general payment mechanism, followed by specific instructions about each game before playing it. All the instructions given to the subjects included examples to illustrate the games and the consequences of playing different strategies, and there was always explicit room for questions.

One aspect of this design that may be worth stressing is that we used only one-shot games. While we acknowledge that repeating some experimental games can show interesting patterns of behavior, we deliberately avoided repetition because one-shot situations fit better our purposes of eliciting social preferences using a variety of games. In particular, this minimizes unwanted cross-contamination effects produced by learning, feedback, income, and reputation building (see, e.g., Goeree and Holt, 2001, 2004). This set-up makes also the games played by the participants more similar to the one-shot field situations they faced at the end of Session 3. In addition, although the games used here have been repeated in experiments, they have also been widely used as one-shot games. Furthermore, one-shot games have been the focus of most the previous lab-field studies on the external validity of social-preference games (Glaeser et al., 2000; Castillo and Carter, 2002; Fehr et al., 2003; Karlan, 2005; Ruffle and Sosis, 2007; Benz and Meier, 2008; Laury and Taylor, 2008; Barr and Serneels, 2009; Barr and Zeitlin, 2010; Carpenter and Myers, 2010; Rustagi et al., 2010; Fehr and Leibbrandt, 2011; Lamba and Mace, 2011; Leibbrandt, 2012; de Oliveira et al., 2011; Cardenas et al., 2013; Franzen and Pointner, 2013; Bernold et al. 2014).

Our participants were not given information or feedback about the results of the different games until the end, with the exception of the cases in which they inevitably had the information from simply playing the games (which happened in the case of player 1 in the dictator games and player 2 in the ultimatum and trust games explained below). All the games were played always in the order specified below.

The seven games played in the session were as follows:

1) *Dictator Game 1 (DG1):* Two-player game in which Player 1 decides how to divide £10 between him/her and Player 2. Then Player 2 simply receives the allocation established by Player 1. Half of the participants were Player 1 and the other half Player 2.
2) *Dictator Game 2 (DG2):* Like Dictator Game 1, but switching the roles (and matching people with different partners).
3) *Ultimatum Game 1 (UG1):* Two-player game in which Player 1 decides how to divide £10 between him/her and Player 2. Then Player 2 decides whether to accept the allocation or

not. If the allocation is rejected, both players get nothing. Half of the participants were Player 1 and the other half Player 2.

4) *Ultimatum Game 2 (UG2):* Like Ultimatum Game 1, but all the participants were Player 2 and all of them had to respond to the same allocation of £5 for Player 2, which corresponded to the allocation decided by a participant who was Player 1 in a preliminary pilot session. One randomly selected participant in Session 2 was then actually matched with that previous participant to determine his/her payoff in the game.[5]
5) *Trust Game 1 (TG1):* Two-player game in which Player 1 has an endowment of £10 and decides how much of it to send over to Player 2. The amount sent over is multiplied by three and the given to Player 2, who has to decide how much of it to send back to Player 1. Half of the participants were Player 1 and the other half Player 2.
6) *Trust Game 2 (TG2):* Like Trust Game 1, but all the participants were Player 2 and all of them had to respond to the same amount of £5 sent over by Player 1, which corresponded to the amount decided by a participant who was Player 1 in a preliminary pilot session. One randomly selected participant in Session 2 was then actually matched with that previous participant to determine his/her payoff in the game.
7) *Public Good Game (PGG):* Four-player game in which all the players have an endowment of £10 and have to decide simultaneously how much of it to contribute to a common group fund. The overall money in the group fund is then multiplied by two and split between the four players.

Note that these seven games involve six different decisions per participant, and eight different decisions overall, as follows: (i) Player 1 in DG1 (half of the subjects) or in DG2 (half of the subjects); (ii) Player 1 in UG1 (half of the subjects); (iii) Player 2 in UG1 (half of the subjects); (iv) Player 2 in UG2; (v) Player 1 in TG1 (half of the subjects); (vi) Player 2 in TG1 (half of the subjects); (vii) Player 2 in TG2; and (viii) one of the players in PG. The allocation of participants was arranged so that those who acted as Player 1 in UG1 acted as Player 1 again in TG1, which also implies that those who were Player 2 in UG1 were Player 2 again in TG1. So, every participant was Player 1 in a dictator game; Player 2 in an ultimatum and in a trust game with a fixed amount of £5; and one of the players in a public good game. In addition, half the participants were Player 1 in an ultimatum game and in a trust game, and the other half were Player 2 in those games.

These different types of experimental games cover a substantial proportion of research on social preferences and they address many of the main behavioral constructs invoked in the literature to explain social behaviors. Those constructs (which refer sometimes to alternative explanations of the same behaviors) include: *altruism* (Player 1 in DG1 and DG2); *positive*

[5] This method is a simple way of eliciting Player 2 behaviors in the ultimatum game presenting all the participants with the same situation (instead of having varying offers by Player 1), and it avoids the additional complications of techniques like the strategy method, which would have made the experimental session excessively complex.

reciprocity (Player 2 in TG1 and TG2); *negative reciprocity* (Player 2 in UG1 and UG2); *anticipation of positive reciprocity* (Player 1 in TG1); *anticipation of negative reciprocity* (Player 1 in UG1); *trust* (Player 1 in TG1); *cooperation* (PGG); and *inequality aversion* (which could be used to address all the players in all the games). This variety of games and behavioral constructs associated to social preferences constitutes our benchmark to compare behavior in social-preference games to the self-reported social behaviors from the past and to the social behaviors exhibited in the field situations.

2.3. Session 3: The field situations

In Session 3, the participants again returned to the lab, were assigned individual ID codes and randomly allocated to their cubicles, and then received more specific and detailed instructions for the tasks they had to do throughout the session.

In this session, the participants had to do one single task that was unrelated to the present study. The task consisted in making choices between different consumer products. At the end of the task, the subjects were paid one by one the £30 they were entitled to for having participated in the three sessions. The £30 were always paid using exactly the same bill and coin denominations, namely: two £10 bills, one £5 bill, three £1 coins, two 50 pence coins, and five 20 pence coins. This was done to make sure that all the participants had available cash in various denominations before facing the field situations they encountered outside of the lab. We made sure that one participant left the lab every three minutes approximately, which was enough for the field situation that the previous participant faced outside to finish.

When the participants went out of the laboratory, they encountered one of several naturalistic field situations that we created for them, in which they could behave pro-socially. We prepared five different field situations overall and each participant faced only one of them. Two of them involved helping and the other three donating money. The five situations were run consecutively, in the order specified below.

The five field situations were as follows:

1) *Boxes:* A research assistant stood in an area outside the lab and told the participants that he needed help to carry two voluminous (but light) boxes to the basement of the university building where the lab was located. He explicitly asked the participants one by one as they went out if they could help. If the participants said yes, they actually helped him carry the boxes downstairs.
2) *Phone:* A research assistant stood outside the lab and said to the participants that he was in need to make a quick phone call but his phone was out of battery. He explicitly asked the participants if they could lend him their phone for a minute to make the call. If the participants lent him the phone, he simply made a call, hung up, and said that there was no answer.

3) *Charity Children:* A research assistant stood outside the lab and was collecting money for a leading charity dedicated to helping children in developing countries. He explicitly asked the participants if they wanted to contribute money to the charity. The research assistant was wearing an official university T-shirt and a professional (sealed) charity bucket of the type commonly used to collect donations, with a large sticker with the logo of the charity. He also had color-printed leaflets with a brief description of the charity and its activities. The money given by people was then actually sent to the charity.
4) *Charity Environment:* This situation was exactly like the previous one, but with a different charity. This time it was a leading charity dedicated to protect the environment.
5) *Lab Donation:* This situation was also analogous to situations 3 and 4, but this time the research assistant was asking for money to support research projects conducted in our lab. The money given by people was actually added to the research funds of the lab.

These different field situations cover a variety of naturalistic environments, in which the participants had to express their social inclinations. Helping others and giving money to others are actually representative of a large number of real-world circumstances related to social preferences.

We deliberately do not want to establish strict one-to-one correspondences between these field situations and the behavior of specific players in specific games. The context-free nature of the games makes such correspondences necessarily imprecise. Our strategy here is rather to cover a variety of relevant social-preference games and a variety of relevant field situations and explore the extent to which the games are predictive of social behaviors in the field.

Nevertheless, the different field situations could be related to some of the behavioral constructs presumably captured by the games, as follows: (i) altruism (in different forms) is likely to be related to decisions in the *Box*, *Phone*, and *Charity Children* situations; (ii) positive reciprocity is likely to be part of behavior in the *Lab Donation* situation; (iii) cooperation is likely to be a relevant motive in the *Charity Environment* situation, which focuses essentially on a contribution to a public good; and (iv) inequality aversion may be, to some extent, relevant in all the situations, but possibly especially so in *Charity Children*.

2.4. Participants and sessions

All experimental sessions were conducted at the London School of Economics and Political Science (LSE) Behavioural Research Lab (BRL) between June and September 2012. A total of 363 people participated in the experiment in a total of 35 experimental sessions. The participants were volunteers recruited from the BRL subject pool, which comprises about 5,000 people, mostly current and former students of the University of London. We used no other eligibility or exclusion criteria to select participants. All the experimental procedures were approved by the LSE Research Ethics Committee.

3. Results

We will present our results in four separate sections. We will start by describing briefly the results obtained in the three main elements that we elicited (self-reported measures of past social behaviors, social-preference games, and field situations). Then (in Section 3.4) we will focus on the main research question of the paper, which is the extent to which the games explain the self-reported measures and the field behaviors.

3.1. Self-reported measures of past social behaviors

Figure 1 shows the distribution of total scores obtained by the participants on the SRA Scale. SRA responses are normally combined in one single SRA score, with no multi-factor structure. The means and standard deviations obtained for the different items are shown in Table B1 in Appendix B.

As Figure 1 shows, there was a wide variety in the total SRA scores obtained, with more scores concentrated around the center of the distribution and a slight positive skew.

3.2. Social-preference games

Figures 2a and 2b consist of 4 panels each (Panels A, B, C, and D in Figure 2a, and Panels E, F, G, and H in Figure 2b), which together show the distribution of responses in the 8 different decisions obtained from the games.

The results are broadly in line with the patterns usually reported in the literature. Panel A shows that 37% of the people acting as Player 1 in DG1 and DG2 gave £0 to Player 2. The rest made contributions greater than £0, with most people giving amounts between £1 and £5, and showing a high 25% spike at £5. Contributions above zero in this type of game are typically interpreted as altruism.[6] Panel B shows a different picture for Player 1 in UG1. In that case, only 3% of the people allocated £0 to Player 2, with most people contributing amounts between £1 and £5, and a high 37% spike at £5. This difference between Panels A and B is typically interpreted as an anticipation of negative reciprocity in Player 2 that could lead him/her to reject small contributions. In Panel C, we can see that 14% of the people acting as Player 2 in UG1 rejected the allocations established by Player 1; the rest of the participants accepted them. This rejection behavior is usually interpreted as negative reciprocity on the side of Player 2. Panel D shows approximately the same percentages as Panel C for the case of Player 2 in UG2.

In Panel E, we find contributions scattered all across the range from £0 to £10 for Player 1 in TG1. The highest bar is at £2 (22%), with other bars above 10% at £0, £3, £5, and

[6] Note, however, that this, and also all the other behaviors observed, may also be interpreted as inequality aversion.

£10. Contributions above zero by Player 1 in this type of game are typically interpreted as an anticipation of positive reciprocity in Player 2 (or trust). In Panel F, we see the amounts sent back by Player 2 in TG1. They show a high spike of 47% at £0. The rest of the participants contributed varying amounts across the range from £1 to £15, most of them between £1 and £5. Contributions greater than zero here are typically interpreted as positive reciprocity by Player 2. Panel G shows a very similar pattern for Player 2 in TG2, with slightly fewer people at £0, more people between £1 and £5, and no one at £15. Finally, Panel H shows a tri-modal distribution of contributions in PGG, with two high bars at £0 (21%) and £10 (21%), and a lower spike at £5 (13%). The remaining contributions are scattered across the whole range, with more contributions between £1 and £4 than between £6 and £9. Amounts greater than zero can be interpreted here as cooperative behavior.

Table 2 shows all the pairwise correlations (Spearman's ρ) between the different game decisions. The majority of the correlations are statistically significant at the 5% level (32 out of 48, removing the correlations between the same variables) and positive (26 out of 32). All the negative correlations involve the behavior of Payer 2 in the ultimatum games, reflecting that people who are more likely to accept allocations in the ultimatum games tend to make lower contributions in the other game decisions. Some of the correlations are also relatively high, with 6 of them above 0.4. This shows that there was a relatively high degree of internal consistency in the decisions that the participants made in the games.

3.3. Field situations

Figure 3 shows the distribution of behaviors in the five different field situations, organized in five different panels. It also shows an additional Panel F, which displays the three situations that had to do with money together (*Charity Children*, *Charity Environment* and *Lab Donation*). The number of participants in each situation, after removing the missing data, was 50 in *Boxes*, 44 in *Phone*, 59 in *Charity Children*, 73 in *Charity Environment*, and 48 in *Lab Donation*.[7]

As Panel A shows, 88% of the participants that faced the *Boxes* situation decided to help the research assistant to carry the boxes to the basement, and only 12% denied the help. In Panel B, we see that people were more divided in the *Phone* situation. 70% of the people decided to lend their phones to the research assistant, but 30% did not do it. Panel C shows the distribution of contributions in the *Charity Children* situation. 42% of the participants did not give any money to the charity, and the other 58% gave varying amounts between £0.15 and £5, with higher bars at £1 and £2. Panel D shows that, in the *Charity Environment* situation 67% of the people did not give anything to the charity. The other 33% gave different amounts between £0.05 and £2.10, with a higher spike at £1. Finally, in Panel E, we have the

[7] The main reasons for missing data in the field situations were attrition (i.e., people not completing the three experimental sessions) and incidental factors of the situation that made it impossible for the research assistants to approach particular participants.

contributions made by the participants in the *Lab Donation* situation. 50% of the people did not give money to the lab, and the other 50% contributed varying amounts between £0.20 and £2.00, with higher bars at £1 and £2.

3.4. Do the games explain the past and the field?

We now turn to the main question of whether the game decisions explain the self-reported measures and the field behaviors. To being with, Table 3 contains pairwise correlations (Spearman's ρ) between the eight game variables and the SRA scores. SRA responses are typically aggregated into one total score (*SRAtotal*), but to extend the analysis we calculated a second score (*SRAmoney*), including only the three items that are related to money (Items 4, 5 and 13). The game variables were then correlated with both scores.

Table 3 shows that three out of the eight game variables are significantly correlated with total SRA scores at the 5% level, and only one out of the eight variables is significantly correlated with the second score that we calculated. The significant correlations are relatively low, with no correlation greater than 0.2. The game decisions that show the significant correlation with SRA scores are those of Player 1 in DG1 and DG2 (labeled as DG1&2 P.1), Player 2 in TG2 (TG2 P.2) and the players in PGG. This suggests that these correlations with SRA scores may come from motivations that have to do with altruism, positive reciprocity or cooperative tendencies, which seems consistent with the types of items included in the SRA Scale. Overall, we interpret this as evidence that there is only a weak relationship between social-preference games and SRA responses.

Table 4 contains pairwise correlations (Spearman's ρ) between the eight game variables and the five different field behaviors, including one additional field variable that groups together the contributions made in the three situations that had to do with donating money (*Charity Children, Charity Environment* and *Lab Donation*).

As Table 4 shows, only one of the 48 correlations is statistically significant at the 5% level. It is a correlation of 0.54 between TG1 P.1 and behavior in the *Charity Children* situation. This is likely to be a spurious correlation, given that it is only one out of 48 and that there is no theoretical reason to expect that this game variable would be correlated with this field situation to a larger extent than some of the other game variables (e.g., DG1&2 P.1). In addition, the correlations within the same variables change from positive to negative with no apparent meaningful pattern, which suggests randomness and a lack of consistent relationships. Overall, we interpret this as evidence that there is no systematic relationship between the game decisions and behavior in the field situations that we analyzed.

To extend these initial correlations, we next present a regression analysis that puts together different game variables in the same models and shows how much of the variance in the self-reported measures and the field behaviors is explained by the games.

Table 5 contains a summary of the regression results obtained for the SRA scores. The table consists of two columns, one of them for the results using the total SRA scores

(*SRAtotal*) as the dependent variable and the other for the results using only the three items that have to do with money (*SRAmoney*). The results in each column are obtained from three separate linear (Ordinary Least Squares) regressions: (i) one regression using as explanatory variables the game decisions in which we have responses from the full sample of participants (DG1&2 P.1, UG2 P.2, TG2 P.2, and PGG); (ii) another regression that has as explanatory variables the decisions in which we only had one half of the sample (UG1 P.1 and TG1 P.1); and (iii) one more regression using as explanatory variables the decisions in which we had the other half of the sample (UG1 P.2 and TG1 P.2). The coefficient shown in the table for each variable corresponds to the coefficient obtained for that variable in the corresponding regression in which it was included. In addition to the coefficients, each column also shows the proportion of variance explained by the explanatory variables in each of the regression models, in the form of R^2.[8]

As Table 5 shows, only one variable in the first column and one in the second column appear as statistically significant at the 5% level (UG1 P.1 in the first column and TG2 P.2 in the second). In addition, two of the other variables showing significant correlations in the first column of Table 4 (DG1&2 P.1 and TG2 P.2) appear as significant at the 10% level. These results are broadly in line with the correlations reported in Table 3.

More importantly, the proportions of variance explained by the models in Table 5 are very low. All of them are below 0.07, and most of them are actually very close to zero. We interpret this as evidence that the game variables have a very limited power to explain the SRA scores.

Table 6 contains a summary of the regression results obtained for the field behaviors. The table has six columns, which correspond to the five different field situations plus one additional variable that brings together the three situations that have to do with contributing money (*Charity Children*, *Charity Environment* and *Lab Donation*). Each of the columns is constructed following the same three-regression structure explained for Table 6. In the case of the situations with binary dependent variables (*Boxes* and *Phone*), the models are standard logistic regressions, and the measures of variance explained correspond to McFadden's Pseudo-R^2.

The results in Table 6 show that only two out of 48 coefficients are statistically significant at the 5%. One of them corresponds to the only correlation that was significant at 5% in Table 4 (TG1 P.1 in the *Charity Children* column), and the other one corresponds to a correlation that was significant at the 10% level in Table 4 (TG2 P.2 in the *Lab Donation* column). As mentioned in reference to Table 5, this statistical significance may well be the result of spurious correlations. Overall, the results in Table 6 are broadly consistent with the correlations reported in Table 4.

[8] Note that the variables used in the second and third regressions performed in each column could never be together in the same model, because there is no overlap in observations between them. In addition, putting the variables of one of those two regressions together with the variables in the first model would sacrifice half of the observations contained in the sample.

More substantially, the proportions of variance explained by the regression models are again very low. Most of them are below 0.07 (11 out of 18) and many of them are close to zero. The variation in these proportions does not seem to follow any meaningful patterns and may also be the result of randomness. We interpret these results as evidence that the game decisions have a very limited power to explain the field behaviors that have been investigated here.

Finally, note that all the regressions discussed in this section have been further investigated employing a broad range of different methods, including: regressions with one game variable at a time, 'stepwise' regressions with game variables added in sequentially, robust standard errors, taking logarithms of the dependent variables, Tobit models, non-parametric techniques, and two-stage approaches. The main results remain essentially the same across all these methods, and we have therefore opted to present the results in the simplest and clearest way possible. The outcomes of additional analyses are available from the authors upon request.

4. Discussion and conclusions

We have presented the results of a large lab-field experiment that constitutes arguably the most systematic assessment of the external validity of experimental social-preference games available to date. In particular, we elicited self-reported social behaviors performed in the past, decisions in seven experimental social-preference games, and behaviors in five naturalistic field situations that we created. In this context, we investigated the extent to which the games can explain the self-reported measures and the field behaviors.

The overarching conclusion is that the games do a poor job in explaining both the self-reported measures and the field behaviors. It is particularly striking that they do not seem to explain to any significant extent any of the behaviors observed in the field. Our results seem to support the conclusions by Voors et al. (2012) that, in social-preference games, *"play in lab experiments has no predictive power for behavior in naturally occurring settings"* (p.310); or by Laury and Taylor (2008) that *"one should be cautious when using the results from laboratory […] experiments to make inferences about altruism outside the laboratory"* (p.29).

Evaluating the external validity of social-preference games is, of course, a vast and difficult task, which requires a full research program and can potentially be tackled in a number of different ways. We do not claim to have established any firm or final conclusions about it with this single paper, but we do believe that our results are worrying and they call for more, and more systematic, research on this issue.

We find particularly worrying that most of the available studies that relate behavior in social-preference games with field behaviors report only the results of one game and one field situation. This raises serious concerns about the possibility that some of those results are produced by spurious correlations. There is a well-known and strong bias to write up and to

publish positive results and not negative ones. In this particular case, the bias is likely to be even stronger, because the lack of a clear one-to-one correspondence between experimental games and field situations makes it difficult to justify negative results as being meaningful.

For these reasons, we believe that more systematic studies investigating the external validity of social-preference games are needed. Systematization can be achieved in different ways. The strategy followed in this study has been to compare a variety of games against a variety of field situations using the same sample of participants (and adding also an additional comparison layer with the self-reported measures of past behaviors). This strategy can be extended further in following studies, but there are also other possibilities. One could, for example, compare the patterns observed in reciprocal (or altruistic, or trusting, or cooperative) behavior in the lab to patterns of reciprocal behavior occurring in different field environments (see List, 2006; Stoop et al., 2012; Stoop, 2013).

One potential limitation of our approach is that there is (deliberately) no clear theoretical mapping from one specific game to one specific field situation. While we acknowledge this limitation, we also believe that such a mapping is virtually impossible to achieve with standard social-preference games because of their artificiality and lack of context, unless field situations are stylized to be mere replications of the games. Under those circumstances, however, one could never answer the question of whether the games predict social behaviors of any relevance outside the lab.

Another potential response to our results is that the issue of the external validity of social-preference games does not really matter. For instance, Camerer (2011) argues that there is *"consensus among most experimental economists that realism, generalizability, or external validity are not especially important"* (p.7). While we agree with many of the arguments in Camerer (2011), we respectfully disagree on this specific claim. In our experience, few experimental economists would feel comfortable with the idea that they are simply studying how people play games that have nothing to do with behavioral principles of any relevance in the world outside the lab. That is also clearly not the spirit in which experimental results are presented and discussed in academic journals and conferences. We would even venture saying that the interest that most experimental economists (let alone other types of economists) have in economic experiments comes mainly from external validity, in the sense of being able to learn something about human behavior beyond the specific games played in the lab (see arguments along these lines in Roth, 1988, 1995; Davis and Holt, 1993; Loewenstein, 1999; Starmer, 1999a,b; Hertwig and Ortmann, 2001; Smith, 2002, 2003; Harrison and List, 2004; Bardsley, 2005; Guala, 2005; Schram, 2005; Bardsley et al. 2009; Croson and Gachter, 2010).

We will finish by stressing two important points. First, we do not see our research as addressing any dispute about lab versus field experimentation. As noted by authors like Harrison and List (2004), List and Levitt (2007b), Falk and Heckman (2009), Camerer (2011), or Harrison (2013), among others, the relationship between lab and field experiments is a symbiotic one, with the two approaches complementing each other. Both lab and field

experiments have their own strengths and weaknesses. Lab experiments, for instance, are important because of their ability to tightly control the environment and isolate stylized causal relationships, to closely reproduce conditions of theoretical models, and to replicate past findings. Furthermore, they can provide insights on important behavioral patterns prior to moving into the field (Levitt and List, 2007b). There are indeed countless types of laboratory experiments in the social and behavioral sciences, and many of them have proved to be invaluable in uncovering behavioral principles of relevance for real-world phenomena outside the lab.

From this point of view, we would interpret a mismatch between lab behavior and related field behaviors of interest as an indication of a potentially inadequate experimental framework, rather than as a general disqualification of laboratory research. So, our conclusions here are not at all on the adequacy of laboratory experimentation as a whole, but on the external validity of experimental social-preference games, which constitute the bedrock of modern research on social preferences in economics and other related disciplines. As Falk and Heckman (2009) point out, most potential limitations of the social-preference experimental paradigm *"raise questions that can be very well analyzed with lab experiments, suggesting the wisdom of conducting more lab experiments, not fewer"* (p.535).

Second, we do not see our study as dismissing the important contributions of the literature on social preferences. It is undeniable that the social-preference paradigm has provided groundbreaking insights into phenomena like cooperation or punishment (e.g., Andreoni and Miller, 2002; Fehr and Gachter, 2000, 2002; Herrmann et al., 2008). There is, however, a more specific issue of whether the particular type of lab experimentation being done in this paradigm is capturing the actual underpinnings of real-life social behavior, which may have to lead to a revision of some of the experimental methods used in the paradigm.

It may still be too early to say how that should be done, but part of the answer may involve bringing more context into the lab, and constructing experimental environments that resemble more closely naturalistic situations. After all, experimental economics and psychology have widely documented that subtle differences in the context can have profound effects on how people behave (see, e.g., Ross and Ward, 1996; Cherry et al., 2002; Ariely et al., 2006; List, 2007; Bardsley, 2008; Stewart et al., forthcoming). As pointed out by Harrison and List (2008), *"it is not the case that abstract, context-free experiments provide more general findings if the context itself is relevant to the performance of the subjects"* (p.840).

References

Abeler, J., and Nosenzo, D. (Forthcoming). Self-selection into laboratory experiments: Pro-social motives versus monetary incentives. *Experimental Economics*.

Andreoni, J. (1988). Privately provided public goods in a large economy: The limits of altruism. *Journal of Public Economics*, 35, 57-73.

Andreoni, J., and Miller, J.H. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70, 737-753.

Ariely, D., Loewenstein, G., and Prelec, D. (2006). Tom Sawyer and the construction of value. *Journal of Economic Behavior & Organization*, 60(1), 1-10.

Bandiera, O., Barankay, I., and Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics*, 120(3), 917-962.

Bardsley, N. (2005). Experimental economics and the artificiality of alteration. *Journal of Economic Methodology*, 12, 239-251.

Bardsley, N. (2008). Dictator Game giving: Altruism or artefact? *Experimental Economics*, 11(2), 122-133.

Bardlsey, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., and Sugden, R. (2009). *Experimental economics: Rethinking the rules*. Princeton University Press: Princeton,

Barr, A., and Serneels, P. (2009). Reciprocity in the workplace. *Experimental Economics*, 12(1), 99-112.

Barr, A., and Zeitlin, A. (2010). Dictator games in the lab and in nature: External validity tested and investigated in Ugandan primary schools. CSAE Working Paper 2010-11.

Bellemare, C., and Kroger, S. (2007). On representative social capital. *European Economic Review*, 51, 183-202.

Benz, M. and Meier, S. (2008). Do people behave in experiments as in the field? Evidence from donations. *Experimental Economics*, 11, 268-281.

Berg, J., Dickhaut, J.W., and McCabe, K.A. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 90, 166-93.

Bernold, E., Gsottbauer, E., Ackermann, K., and Murphy, R. (2014). Social framing and cooperation: The roles and interaction of preferences and beliefs. ETH Zurich.

Bolton, G.E. and Ockenfels, A. (2000). ERC - A theory of equity, reciprocity, and competition. *American Economic Review*, 10, 122-42.

Bouma, J., Bulte, E., and van Soest, D. (2008). Trust and cooperation: Social capital and community resource management. *Journal of Environmental Economics and Management*, 56, 155-166.

Branas-Garza, P. (2007). Promoting helping behavior with framing in dictator games. *Journal of Economic Psychology*, 28(4), 477-486.

Camerer, C.F. (2011). The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. SSRN working paper.

Camerer, C. and Thaler, R. (1995). Anomalies: Ultimatums, dictators and manners. *Journal of Economic Perspectives*, 9, 209-219.

Cameron, L. (1999). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry*, 37, 47-59.

Cardenas, J.C. (2003). Real wealth and experimental cooperation: Experiments in the field lab. *Journal of Development Economics*, 104, 16-33.

Cardenas, J.C., and Carpenter, J. (2005). Three themes on field experiments and economic development. In J. Carpenter, J. List and G. Harrison (Eds.), *Field Experiments in Economics*. JAI Press: Greenwich.

Cardenas, J.C., Chong, A., and Nopo, H. (2013). Stated social behaviour and revealed actions: Evidence from six Latin American countries. *Journal of Development Economics*, 104, 16-33.

Carpenter, J.P., Connolly, C., and Myers, C.K. (2008). Altruistic behavior in a representative dictator experiment. *Experimental Economics*, 11, 282-298.

Carpenter, J.P., and Myers, C.K. (2010). Why volunteer? Evidence on the role of altruism, reputation, and incentives. *Journal of Public Economics*, 94, 911-920.

Carpenter, J.P. and Seki, E. (2005). Competitive work environments and social preferences: Field experimental evidence from a Japanese fishing community. *The B.E. Journal of Economic Analysis & Policy*, 5, 1-25.

Carpenter, J.P., and Seki, E. (2010). Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry*, 49(2), 612-630.

Carpenter, J.P., Verhoogen, E., and Burks, S. (2005). The effect of stakes in distribution experiments. *Economics Letters*, 86, 393-398.

Castillo, M., and Carter, M.R. (2002). The economic impacts of altruism, trust and reciprocity: An experimental approach to social capital. AAE Staff Papers, University of Wisconsin-Madison.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117, 817-69.

Cherry, T.L., Frykblom, P., and Shogren, J.F. (2002). Hardnose the dictator. *American Economic Review*, 92(4), 1218-1221.

Cleave, B., Nikiforakis, N., and Slonim, R. (2013). Is there selection bias in laboratory experiments? The case of social and risk preferences. *Experimental Economics*, 16, 372-382.

Cox, J.C. (2004). How to identify trust and reciprocity. *Games and Economic* Behavior, 46, 260-281.

Croson, R., and Gachter, S. (2010). The science of experimental economics. *Journal of economic Behaviour and Organization*, 73(1), 122-131.

Dana, J.R., Cain, D., and Dawes, R. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator game. O*rganizational Behavior and Human Decision Processes*, 100, 193-201.

Dana, J.R., Weber, R., and Kuang, J. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33, 67-80.

Davis, D.D., and Holt, C.A. (1993). *Experimental Economics*. Princeton University Press: Princeton.

de Oliveira, A.C.M, Croson, R.T.A., and Eckel, C.C. (2011). The giving type: Identifying donors. *Journal of Public Economics*, 95, 428-435.

Dolan, P., and Galizzi, M.M. (Forthcoming). Getting policy-makers to listen to field experiments. *Oxford Review of Economic Policy*.

Durkheim, E. (1893). *The division of labor in society*. Paris: Alcan.

Eckel, C.C., and Grossman, P.J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, 16, 181-191.

Englmaier, F., and Gebhardt, G. (2010). Free riding in the lab and in the field. SFB/TR 15 Discussion Paper No 344.

Falk, A., Meier, S., and Zehnder, C. (2013). Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association*, 11(4), 839-852.

Falk, A. and Heckman, J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326, 535-538.

Fehr, E. and Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90, 980-994.

Fehr, E., and Gachter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137-140.

Fehr, E., Fischbacher, U., Von Rosenbladt, B., Schupp, J., and Wagner, G.G. (2003). A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative surveys. CESIfo Working Paper No 866.

Fehr, E., and Leibbrandt, A. (2011). A field study on cooperativeness and impatience in the Tragedy of the Commons. *Journal of Public Economics*, 95, 1144-1155.

Fehr, E., and Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114, 173-68.

Fehrler, S., and Przepiorka, W. (2013). Charitable giving as a signal of trustworthiness: disentangling the signaling benefits of altruistic acts. *Evolution and Human Behavior*, 34, 139-145.

Fischbacher, U. (2007). Z-Tree: Zurich Toolbox for Readymade Economic Experiments. Experimenter's manual. *Experimental Economics*, 10, 171-178.

Fischbacher, U., and Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public good experiments. *American Economic* Review, 100, 541-556.

Forsythe, R., Horowitz, J.L., Savin, N.E., and Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior*, 6, 347-369.

Franzen, A., and Pointner, S. (2012). Anonymity in the dictator game revisited. *Journal of Economic Behavior and Organization*, 81(1), 74-81.

Franzen, A., and Pointner, S. (2013). The external validity of giving in the dictator game: A field experiment using the misdirected letter technique. *Experimental Economics*, 16, 155-169.

Gachter, S., Herrmann, B., and Thoni, C. (2004). Trust, voluntary cooperation and socio-economic background: Survey and experimental evidence. *Journal of Economic Behavior and Organization*, 55, 505-531.

Garbarino, E., and Slonim, R. (2009). The robustness of trust and reciprocity across a heterogeneous population. *Journal of Economic Behavior and Organization*, 69(3), 226-240.

Glaeser, E.L., Laibson, D.I., Scheinkman, J.A., and Soutter, C.L. (2000). Measuring trust. *Quarterly Journal of Economics*, 115, 811-846.

Guala, F. (2005). *The Methodology of Experimental Economics*. Cambridge University Press: Cambridge.

Guth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3, 367-88.

Harrison, G.W., List, J.A., and Towe, C. (2007). Naturally occurring markets and exogenous laboratory experiments: A case study of risk aversion. *Econometrica*, 75(2), 433-458.

Harrison, G.W., and List, J.A. (2004). Field experiments. *Journal of Economic Literature*, 42, 1009-1055.

Harrison, G.W., and List, J.A. (2008). Naturally occurring markets and exogenous laboratory experiments: A case study of the winner's curse. *The Economic Journal*, 118, 822-843.

Herrmann, B., Thoni, C. and Gachter, S. (2008). Anti-social punishment across societies. *Science*, 319, 1362-1367.

Hertwig, R., and Ortmann, A. (2001). Experimental practices in economics: A challenge for psychologists? *Behavioral and Brain Sciences*, 24, 383-451.

Hoffman, E., McCabe, K., Schachat, J., and Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7, 346-380.

Hoffman, E., McCabe, K., and Smith, V. (1996). Social distance and other regarding behavior in dictator games. *American Economic Review*, 86(3), 653-660.

Holm H.J., and Danielson, A. (2005). Tropic trust versus Nordic trust: Experimental evidence from Tanzania and Sweden. *The Economic Journal*, 115(503), 505-532.

Karlan, D.S. (2005). Using experimental economics to measure social capital and predict financial decisions. *American Economic Review*, 95, 1688-1699.

Krawczyk, M. (2011). What brings subjects to the lab? *Experimental Economics*, 14(4), 482-489.

Lamba, S. and Mace, R. (2011). Demography and ecology drive variation in cooperation across human populations. *Proceedings of the National Academy of Sciences of the USA*, 108, 14426-14430.

Laury, S.K., and Taylor, L.O. (2008). Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good? *Journal of Economic Behavior & Organization*, 65, 9-29.

Leibbrandt, A. (2012). Are social preferences related to market performance? *Experimental Economics*, 15, 589-603.

Levitt, S., and List, J.A. (2007a). What do laboratory experiments measuring social preferences reveal about the real world. *Journal of Economic Perspectives,* 21, 153-174.

Levitt, S., and List, J.A. (2007b). Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics*, 40(2), 347-370.

Levitt, S. and List, J.A. (2008). Homo economicus evolves. *Science*, 319, 909-910.

Levitt, S., List, J.A., and Reiley, D.H. (2010). What happens in the field stays in the field: Exploring whether professionals play minimax in laboratory experiments. *Econometrica*, 78(4), 1413-1434.

Lewin, K., Lippitt, R., and White, R.K. (1939). Patterns of aggressive behavior in experimentally created "social climates". *Journal of Social Psychology*, 10, 171-299.

Lichtenstein, S., and Slovic, P. (Eds.). (2006). *The construction of preference*. Cambridge University Press.

List, J.A. (2004). Young, selfish, and male: field evidence on social preferences. *Economic Journal*, 114, 121-149.

List, J.A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy*. 114(1), 1-37.

List, J.A. (2007). On the interpretation of giving in Dictator Games. *Journal of Political Economy*, 115(3), 482-493.

List, J.A. (2008). Introduction to field experiments in economics with applications to the economics of charity. *Experimental Economics*, 11, 203-212.

List, J.A. (2009). Social preferences: Some thoughts from the field. *Annual Review of Economics*, 1, 563-579.

List, J.A., and Cherry T.L. (2008), Examining the role of fairness in high stakes allocation decisions. *Journal of Economic Behavior & Organization* 65(1), 1-8.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, 109, 25-34.

Munier, B., and Zaharia, C. (2003). High stakes and acceptance behavior in ultimatum bargaining: A contribution from an international experiment. *Theory and Decision*, 53, 187-207.

Ortmann, A., Fitzgerald, J., and Boeing, C. (2000). Trust, reciprocity, and social history: A re-examination. *Experimental Economics*, 3, 81-100.

Oosterbeek, H., Sloof, R., and van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7, 171-188.

Palacios-Huerta, I., and Volij, O. (2008). Experientia docet: Professionals play minimax in laboratory experiments. *Econometrica*, 76(1), 71-115.

Roe, B.E., and Just, D.R. (2009). Internal and external validity in economics research: Trade-offs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics*, 91(5), 1266-1271.

Ross, L., and Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E.S. Reed and E. Turiel (Eds.), *Values and knowledge* (pp. 103-135). Erlbaum: Hillsdale, NJ.

Roth, A.E. (1988). Laboratory experimentation in economics: A methodological overview. *Economic Journal*, 98, 974-1031.

Roth, A.E. (1995). Introduction to experimental economics. In J. Kagel and A.E. Roth (Eds.), *The Handbook of Experimental Economics* (pp. 3-110). Princeton University Press: Princeton.

Ruffle, B.J., and Sosis, R. (2007). Does it pay to pray? Costly ritual and cooperation. *The B.E. Journal of Economic Analysis and Policy*, 7(1) (Contributions), Article 18.

Rushton, J.P., Chrisjohn, R.D., and Fekken G.C. (1981). The altruistic personality and the self-report altruism scale. *Personality and Individual Differences*, 2, 293-302.

Rustagi, D., Engel, S., and Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science*, 330, 961-965.

Schram, A. (2005). Artificiality: The tension between external and internal validity in economic experiments. *Journal of Economic Methodology*, 12, 225-237.

Serra, D., Serneels, P., and Barr, A. (2010). Intrinsic motivations and the non-profit health sector: Evidence from Ethiopia. CSAE Working Paper 2010-04.

Slonim, R., and Roth, A. (1998). Learning in high stakes ultimatum games: An experiment in the Slovak Republic. *Econometrica*. 66, 569-596.

Slonim, R., Wang, C., Garbarino, E., and Merrett, D. (2013). Opting-in: Participation bias in economic experiments. *Journal of Economic Behavior and Organization*, 90, 43-70.

Slovic, P. (1995). The construction of preference. *American psychologist*, 50(5), 364-371.

Smith, A. (1759). *The theory of moral sentiments*. A. Millar: London.

Smith, V.L. (2002). Method in experiment: Rhetoric and reality. *Experimental Economics*, 5, 91-110.

Smith, V.L. (2003). Constructivist and ecological rationality. *American Economic Review*, 93(2), 465-508.

Starmer, C. (1999a). Experimental economics: Hard science or wasteful tinkering? *Economic Journal*, 109, 5-15.

Starmer, C. (1999b). Experiments in economics: Should we trust the dismal scientists in white coats? *Journal of Economic Methodology*, 6, 1-30.

Stewart, N., Reimers, S., and Harris, A. J. (Forthcoming). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*.

Stoop, J., Noussair, C.N., and van Soest, D. (2012). From the lab to the field: Cooperation among fishermen. *Journal of Political Economy*, 120(6), 1027-1056.

Stoop, J. (2013). From the lab to the field: Envelopes, dictators and manners. *Experimental Economics*.

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *American Journal of Psychology*, 9, 507-533.

Voors, M., Bulte, E., Kontoleon, A., List, J.A., and Turley, T. (2011). Using Artefactual Field Experiments to learn about the incentives for sustainable forest use in developing economies. *American Economic Review*, 101(3), 329-333.

Voors, M., Turley, T., Kontoleon, A., Bulte, E., and List, J.A.(2011). Exploring whether behavior in context-free experiments is predictive of behavior in the field: Evidence from lab and field experiments in rural Sierra Leone. *Economics Letters*, 114, 308-311

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6, 299-310.

Appendix A: The SRA Scale

INSTRUCTIONS: Tick the category on the right that conforms to the frequency with which you have carried out the following acts.

	Never	*Once*	*More than once*	*Often*	*Very often*
1. I have helped push a stranger's car out of the snow.					
2. I have given directions to a stranger.					
3. I have made change for a stranger.					
4. I have given money to a charity.					
5. I have given money to a stranger who needed it (or asked me for it).					
6. I have donated goods or clothes to a charity.					
7. I have done volunteer work for a charity.					
8. I have donated blood.					
9. I have helped carry a stranger's belongings (books, parcels, etc.).					
10. I have delayed an elevator and held the door open for a stranger.					
11. I have allowed someone to go ahead of me in a lineup (at Xerox machine, in the supermarket).					
12. I have given a stranger a lift in my car.					
13. I have pointed out a clerk's error (in a bank, at the supermarket) in undercharging me for an item.					
14. I have let a neighbour whom I didn't know too well borrow an item of some value to me (e.g., a dish, tools, etc.).					
15. I have bought 'charity' Christmas cards deliberately because I knew it was a good cause.					
16. I have helped a classmate who I did not know that well with a homework assignment when my knowledge was greater than his or hers.					
17. I have, before being asked, voluntarily looked after a neighbour's pets or children without being paid for it.					
18. I have offered to help a handicapped or elderly stranger across a street.					
19. I have offered my seat on a bus or train to a stranger who was standing.					
20. I have helped an acquaintance to move households.					

Appendix B: Additional tables

Table B1: Means and Standard Deviations (SD) SRA Scale

Item	*Mean*	*SD*
1	1.65	1.01
2	3.86	1.06
3	3.28	0.97
4	1.35	0.82
5	2.40	1.13
6	2.42	1.14
7	2.07	1.22
8	3.65	0.94
9	2.03	1.24
10	2.66	1.25
11	3.79	0.97
12	4.09	0.80
13	2.52	1.25
14	2.69	1.17
15	3.35	0.90
16	2.65	1.09
17	3.40	1.05
18	3.29	1.21
19	1.77	1.20
20	2.91	1.14
Total	55.83	10.98

Tables and figures

Table 1: Summary of lab-field studies on external validity of social-preference games

	Subjects	*N*	*Setting*	*Lab game*	*Field variable*	*Lab-field correlation?*
Glaeser et al. (2000)	Undergraduate students	196	Harvard University	Modified TG	i) Twelve survey questions about trusting attitudes; ii) survey questions about past trusting behavior	i): No for TG P.1 in ten questions out of twelve; some, but weak, for TG P.2. ii): Some, but weak, for TG P.1; no for TG P.2
Castillo & Carter (2002)	Urban and rural community members	Not reported	KwaZulu-Natal, South Africa	DG, TG	i) Per capita household expenditure; ii) density of associational life	i): Yes for DG and TG P.1 in urban areas; No for DG and TG P.1 in rural areas. ii): No
Fehr et al. (2003)	Representative sample of adults	429	Germany	TG	Six survey questions about trusting attitudes	No for TG P.1 in five questions out of six; no for TG P.2.
Cardenas & Carpenter (2005)	Urban slum dwellers	240	Bangkok, Thailand and Ho Chi Minh City, Vietnam	VCM	Household expenditure	Yes for Thailand; No for Vietnam
Holm & Danielson (2005)	Undergraduate economics students	200 + 110	Dar Es Salaam University, Tanzania; Lund University, Sweden	TG, DG	i) Survey questions about trusting attitudes; ii) survey question about past trusting behavior (lending money)	i): No for TG P.1, TG P.2, and DG in Tanzania; some for TG P.1, TG P.2, and DG in Sweden; ii): No for TG P.1, TG P.2, and DG
Karlan (2005)	Women members of non-profit village banking organization (FINCA)	864	Ayacucho, Peru	TG, PGG	i) Repaying micro-loans; ii) savings, iii) likelihood of default.	Yes for TG P.2; Yes for TG P.1 but in 'opposite' direction; No for PGG

Ruffle & Sosis (2007)	Members of religious kibbutzim	208	Israel	CPRG	Synagogue attendance	Yes for males; No for females
Benz & Meier (2008)	University students	99 + 83	University of Zurich	DG framed as contribution to either a 'social fund' or a 'charity'	Donations to two university funds supporting students in financial difficulties or foreigners	Some, but weak for the 'charity' DG
Bouma et al. (2008)	Rural villagers in watersheds	92	Five villages in India	TG	i) Investment in soil and water conservation; ii) efforts to operate and maintain conservation infrastructures	No
Laury & Taylor (2008)	University students	125 + 68	Georgia State University	PGG	DCE for contribution to local non-profit organization protecting the metropolitan environment by planting trees	Some, but not in a uniform way (e.g., positive relationship between number of rounds being a 'weak free-rider' and probability of purchasing a tree)
Barr & Serneels (2009)	Manufacturing workers	424	Ghana	TG	i) Firm productivity; ii) individual earnings	i): No for mean return in TG P.2; Yes for proportion of 'high reciprocators' in TG P.2. ii): Yes for 'high reciprocators'
Baran et al. (2010)	MBA students	462	Chicago Booth Business School	TG	Donation to their university	Yes
Barr & Zeitlin (2010)	Primary school teachers	487	Uganda	DG	Proportion of contracted time allocated to teaching in previous month	Some, but weak.
Carpenter & Myers (2010)	Volunteer firefighters, non-volunteer community members	205 + 189	Vermont, US	DG, 'charity' DG	i) Self-reported training hours; ii) self-reported call hours; iii) recorded response to calls; iv) odds of being a volunteer firefighter relative to a non-	Yes for i) and iv); No for ii) and iii).

					volunteer community member	
Englmaier & Gebhardt (2010)	University students	20	University of Munich	PGG	Average speed in registering books in institute library in a group task	Yes
Rustagi et al. (2010)	Commons users groups	49	Bale region, Ethiopia	PGG	i) Forest management activities; ii) time spent on monitoring forest	Yes (based on number of 'conditional cooperators' in the group)
Serra et al. (2010)	Nursing students, medical students	219 + 90	Ethiopia	GTG	Choice to work in non-profit healthcare sector	Yes
Carpenter & Seki (2011)	Fishermen catching shrimps in 'pooling' and 'non-pooling' boats	14 + 13	Toyama Bay, Japan	VCM, VCM with 'social disapproval'	Kilograms of fish caught per trip	Yes (based on conditional cooperation measures, and propensity to disapprove)
de Oliveira et al. (2011)	Respondents in low-income African-American neighbourhood	190	Dallas, US	VCM	Amounts donated to three neighbourhood charities (providing health, education, and job training services)	Yes
Fehr & Leibbrandt (2011)	Fishermen selling shrimps in open-air markets	114	Villages near a lake in North Eastern Brazil	PGG	i) Hole size in shrimp traps, ii) fishnet mesh size	Yes
Lamba & Luce (2011)	Villagers in small-scale forager-horticulturist Pahari Korwa society	413	16 villages in Central India	PGG	Individual deviation from equal share in distribution of salt among the villagers	No
Voors et al. (2011)	Poor rural villagers	99	South Eastern Sierra Leone	PGG	Survey attitudinal measures on i) illegal commercial	No

					mining, logging, and hunting; ii) illegal hunting of endangered species; iii) support to forest conservation	
Leibbrandt (2012)	Fishermen selling shrimps in open-air markets	143	Villages near a lake in North Eastern Brazil	PGG	i) Achieved price (for shrimps of similar quality); ii) duration and stability of trade relations	Yes
Voors et al. (2012)	Poor rural villagers	453	South Eastern Sierra Leone	PGG	Survey attitudinal measure on contribution to community project fund for the village	No
Franzen & Pointner (2013)	University students	27 + 75	Universities of Cologne and Bern	DG	Returning a misdirected letter	Some, but weak
Cardenas et al. (2013)	General population respondents in six capital cities	567 + 498 + 488 + 541 + 580 + 435	Bogota, Buenos Aires, Caracas, Lima, Montevideo, San Jose	TG, VCM	i) Participation in any social organization; ii) attendance to their meeting; iii) participation in their decision planning; iv) hours in a month spent in them.	Yes for TG P.1, No for TG P.2. No for VCM
Bernold et al. (2014)	University students	47 + 47 + 47 + 47	University of Zurich, ETH Zurich	PGG	Donation to an environmental charity	No

Notes:

DG stands for Dictator Game; TG for Trust Game; TG P.1 for Trust Game Player 1; TG P.2 for Trust Game Player 2; GTG for Generalized Trust Game; PGG for Public Good Game; VCM for Voluntary Contribution Mechanism; CPRG for Common Pool Resource Game; DCE for Discrete Choice Experiment.

The '+' sign in the N column separates the sample sizes of different experimental treatments/pools/locations in the same study.

Table 2: Pairwise correlations between game decisions (Spearman's ρ)

	DG1&2 P.1	*UG1 P.1*	*UG1 P.2*	*UG2 P.2*	*TG1 P.1*	*TG1 P.2*	*TG2 P.2*	*PG*
DG1&2 P.1	1.00***	0.48***	-0.09	-0.18***	0.26***	0.32***	0.50***	0.36***
UG1 P.1	0.48***	1.00***	-	-0.20**	0.26***	-	0.35***	0.25***
UG1 P.2	-0.09	-	1.00***	0.09	-	0.00	0.05	0.00
UG2 P.2	-0.18***	-0.20**	0.09	1.00***	-0.02	-0.11	-0.15**	-0.09
TG1 P.1	0.26***	0.26***	-	-0.02	1.00***	-	0.43***	0.25***
TG1 P.2	0.32***	-	0.00	-0.11	-	1.00***	0.38***	0.30***
TG2 P.2	0.50***	0.35***	0.05	-0.15**	0.43***	0.38***	1.00***	0.34***
PGG	0.36***	0.25***	0.00	-0.09	0.25***	0.30***	0.34***	1.00***

Notes:

"*", "**" and "***" stand for statistical significance at the 10%, 5% and 1% levels respectively.

"-" indicates that the correlation cannot be computed because there is no overlap between participants in the pair of variables.

Table 3: Correlations between game decisions and SRA scores (Spearman's ρ)

	SRAtotal	*SRAmoney*
DG1&2 P.1	0.20***	0.04
UG1 P.1	0.16*	0.06
UG1 P.2	-0.05	0.07
UG2 P.2	-0.05	0.02
TG1 P.1	0.03	0.03
TG1 P.2	0.06	-0.01
TG2 P.2	0.20***	0.15**
PGG	0.14**	0.00

Notes:

"*", "**" and "***" stand for statistical significance at the 10%, 5% and 1% levels respectively.

Table 4: Correlations between game decisions and field behaviors (Spearman's ρ)

	Boxes	*Phone*	*Charity Children*	*Charity Environ.*	*Lab Donation*	*Monetary Together*
DG1&2 P.1	0.04	0.06	-0.25*	0.20*	-0.05	-0.06
UG1 P.1	0.18	0.27	0.18	0.04	0.15	0.15
UG1 P.2	-0.09	0.14	0.22	-0.05	-0.06	0.04
UG2 P.2	-0.13	0.11	0.02	-0.13	0.06	-0.04
TG1 P.1	0.15	0.28	0.54***	0.00	-0.12	0.15
TG1 P.2	0.35	-0.08	0.05	-0.12	-0.24	-0.11
TG2 P.2	0.18	0.27*	-0.18	0.13	0.29*	0.03
PGG	0.14	-0.04	-0.03	0.02	-0.08	-0.02

Notes:

"*", "**" and "***" stand for statistical significance at the 10%, 5% and 1% levels respectively.

Table 5: Regression analysis SRA scores

	SRAtotal	*SRAmoney*
DG1&2	0.61*	-0.01
UG1 P.1	1.25**	0.12
UG1 P.2	-0.78	0.41
UG2 P.2	-1.16	0.01
TG1 P.1	-0.16	-0.00
TG1 P.2	0.02	-0.03
TG2 P.2	0.65*	0.18**
PGG	0.18	-0.01
Var. Explained 1	0.06	0.02
Var. Explained 2	0.04	0.01
Var. Explained 3	0.00	0.01

Notes:

"*", "**" and "***" stand for statistical significance at the 10%, 5% and 1% levels respectively.

Table 6: Regression analysis field behaviors

	Boxes	*Phone*	*Charity Children*	*Charity Environ.*	*Lab Donation*	*Monetary Together*
DG1&2 P.1	-0.07	-0.08	-0.11	0.03	-0.06	-0.04
UG1 P.1	0.24	0.11	0.07	0.01	0.14	0.03
UG1 P.2	142.81	0.59	0.76	-0.08	-0.08	0.08
UG2 P.2	-15.37	0.60	0.11	-0.19	-0.00	-0.05
TG1 P.1	0.05	0.27	0.20***	-0.01	-0.06	0.05
TG1 P.2	16.43	-0.05	-0.00	-0.02	-0.05	-0.03*
TG2 P.2	0.34	0.53*	0.01	0.01	0.22***	0.01
PGG	0.09	-0.04	0.00	-0.01	-0.03	-0.00
Var. Explained 1	0.19	0.12	0.06	0.04	0.16	0.01
Var. Explained 2	0.04	0.10	0.32	0.00	0.11	0.04
Var. Explained 3	0.27	0.02	0.04	0.03	0.08	0.04

Notes:

"*", "**" and "***" stand for statistical significance at the 10%, 5% and 1% levels respectively.

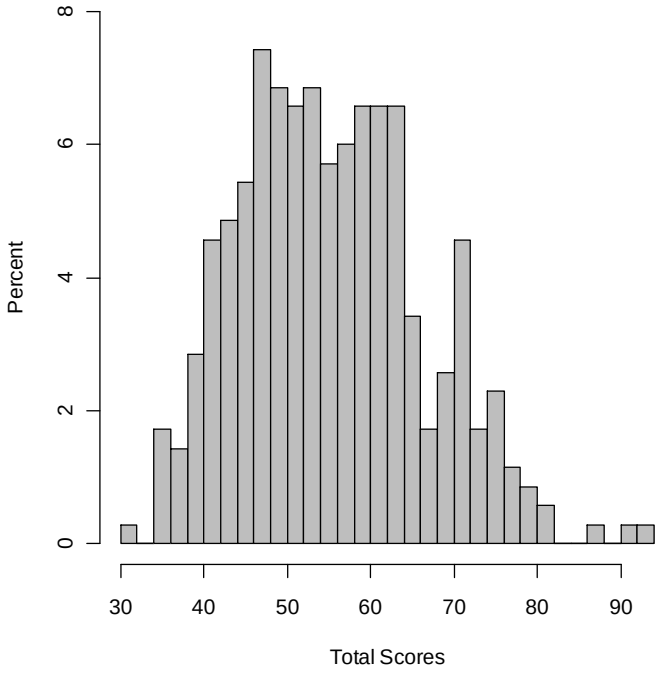

Figure 1: Total SRA Scores

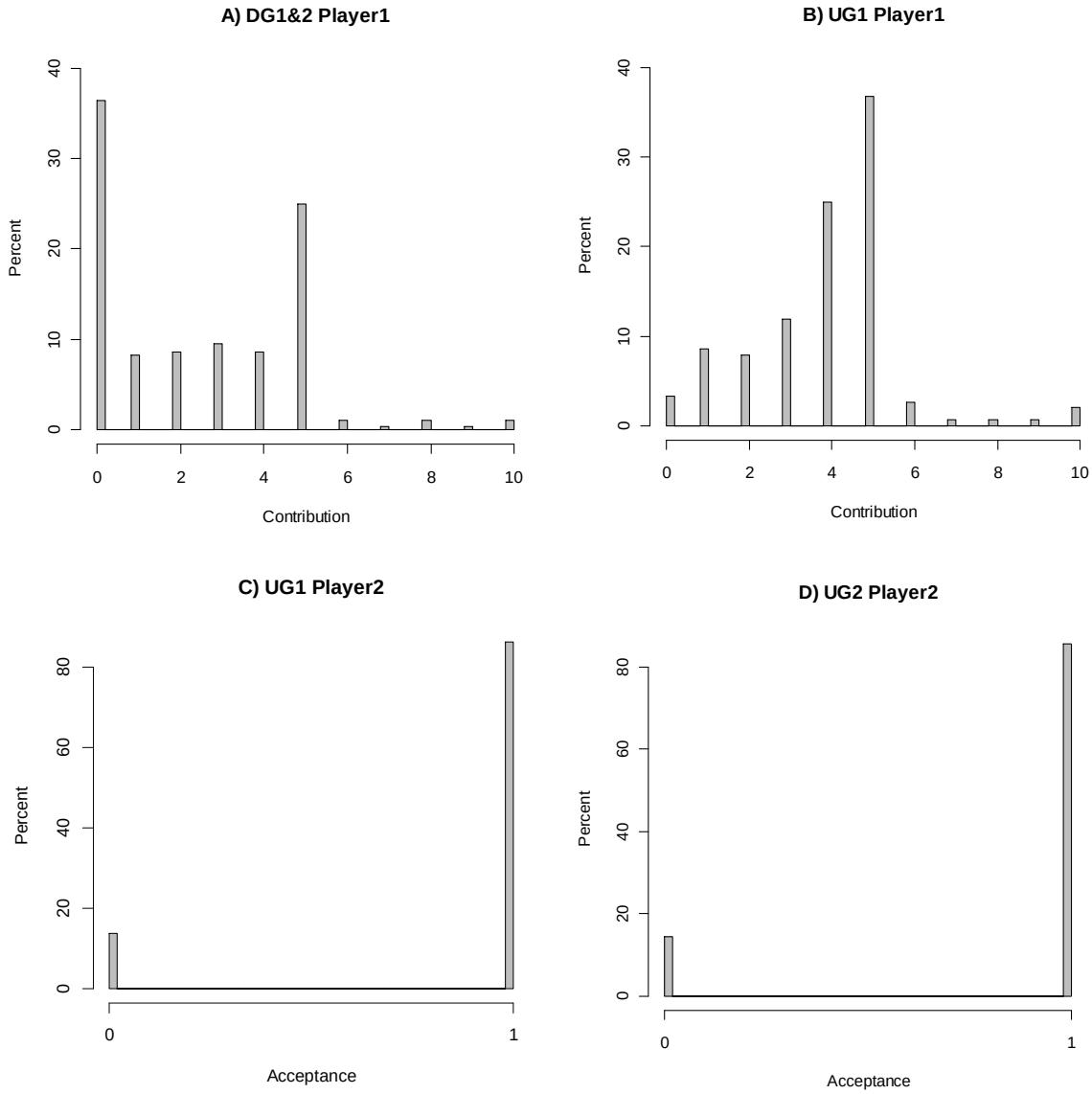

Figure 2a: Distribution of responses in first four game decisions

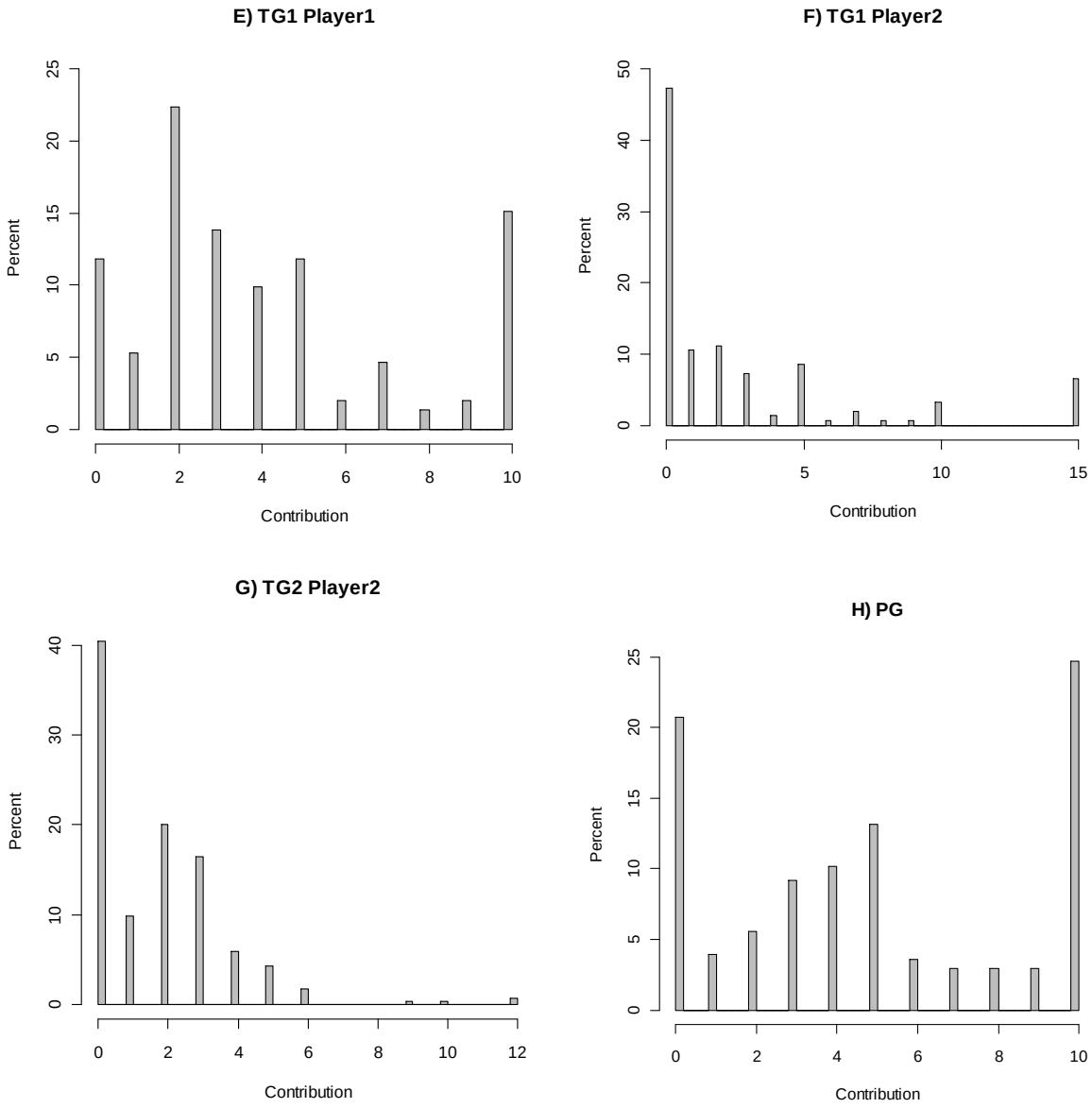

Figure 2b: Distribution of responses in last four game decisions

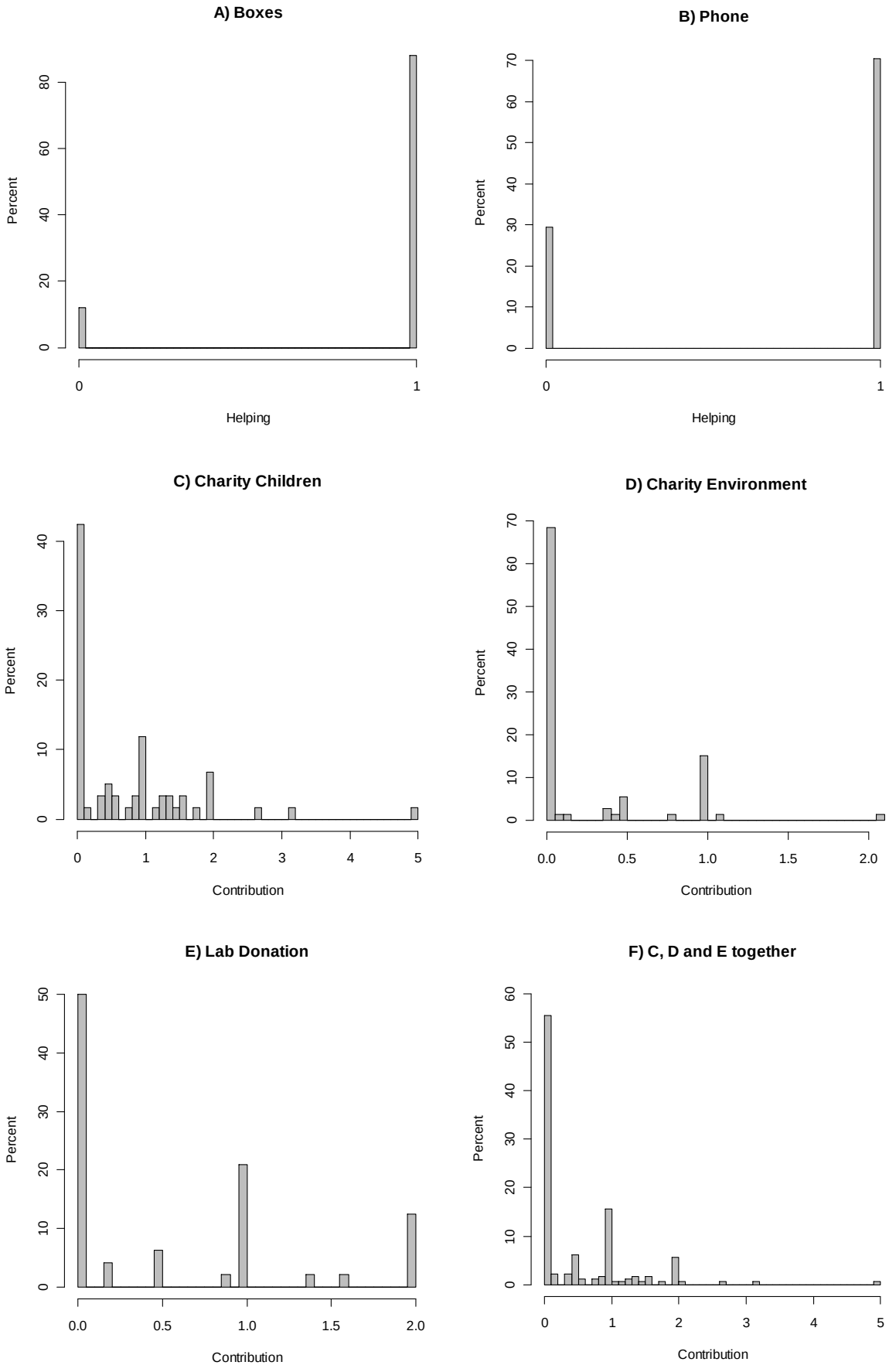

Figure 3: Distribution of behaviors in the field situations