Matteo M. Galizzi, Daniel Navarro-Martinez

On the external validity of social preference games: a systematic lab-field study

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

(Accepted for publication in *Management Science*).

Matteo M. Galizzi

Department of Psychological and Behavioural Science, London School of Economics and Political Science, London, UK.

E-mail address: m.m.galizzi@lse.ac.uk.

Daniel Navarro-Martinez

Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain;
Barcelona Graduate School of Economics, Barcelona, Spain;
Barcelona School of Management, Barcelona, Spain.
E-mail address: daniel.navarro@upf.edu.

Abstract

We present a lab-field experiment designed to systematically assess the external validity of social preferences elicited in a variety of experimental games. We do this by comparing behavior in the different games with several 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 explaining both social behaviors in the field and social behaviors from the past. We also include a systematic review and meta-analysis of previous literature on the external validity of social preference games.

1. Introduction

The last few 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) and 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, Schmittberger, and Schwarze 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 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) and *trust* (e.g., Berg et al. 1995; Ortmann et al. 2000). In a typical study, participants play these games in laboratory settings, where special care is taken to strip the games from contextual features that depart from the underlying game-theoretic structures on which they are based, and to provide 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; Henrich et al. 2001, 2005; Charness and Rabin 2002; Herrmann, Thoni, and Gachter 2008).[1]

Given such a major interest in the topic, it is surprising how little work has been done to investigate systematically the *external validity* of this experimental games approach to social preferences. This seems to us to be one of the most fundamental questions yet to be answered about the social preference paradigm. Specifically, to what extent do 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 previously warned 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 that may arise 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 actions are unlikely to remain anonymous; (iii) the context matters and cannot be completely controlled by the experimenters; (iv) the stakes are

[1] See Camerer (2003) for a comprehensive review of experimental social preference games; see 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.

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 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; Kessler and Vesterlund, 2015). 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 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, Hoffman, McCabe, and Smith 1996; Eckel and Grossman 1996; Dana, Cain, and Dawes 2006; Dana, Weber, and Kuang 2007; List 2007; Bardsley 2008; Franzen and Pointer 2012; Winking and Mizer 2013); the context and framing (Cherry, Frykblom, and Shogren 2002; List 2006; Branas-Garza 2007; Stoop, Noussair, and van Soest 2012; Stoop 2013); the size of the stakes (Slonim and Roth 1998; Cameron 1999; Munier and Zaharia 2003; Carpenter, Verhoogen, and Burks 2005; List and Cherry 2008); the subject pool (Gachter, Herrmann, and Thoni, 2004; List 2004, 2006; Carpenter and Seki 2005; Bellemare and Kroger 2007; Bellemare, Kroger, and van Soest, 2008; 2011; Carpenter, Connolly, and Myers 2008; Garbarino and Slonim 2009; Stoop et al. 2012; Cleave, Nikiforakis, and Slonim 2013; Exadaktylos, Espin, and Branas-Garza 2013; Kessler, 2013; Stoop 2013); and the self-selection into lab experiments (Krawczyk 2011; Falk, Meier, and Zehnder 2013; Slonim et al. 2013; Abeler and Nosenzo 2015).

All these factors have been shown to matter, at least in some cases, which calls into question 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; Ariely, Loewenstein, and Prelec 2006; Lichtenstein and Slovic 2006; Stewart, Reimers, and Harris 2015).

More closely related to the research presented here are a relatively small number of empirical studies that have examined the external validity of experimental social preference games by directly linking evidence from the lab and the field for the same pool of subjects. We have conducted a systematic review and meta-analysis of this literature, focusing on the games that we used in our *lab-field* experiment (i.e., the dictator game, the ultimatum game, the trust game, and the public good game, as explained further down). Appendix A contains the details of the systematic review and meta-analysis. Table A1 provides a synoptic summary of all the lab-field studies retrieved by our systematic review, ordered chronologically.[2]

As Table A1 shows, the currently accumulated evidence is clearly mixed. Some studies have found significant correlations between behavior in particular experimental games and specific field behaviors, others have found no correlation, and some others have obtained mixed findings. More specifically, our meta-analysis reveals that 39.7% of the reported lab-field correlations and 37.5% of the reported lab-field regressions find a statistically significant association between games and field behaviors. The overall average lab-field correlation reported is 0.14, and the overall correlation in the papers that report significant correlations is 0.27. So, analyzing all the papers together, there is currently only weak evidence of correlation between these social preference games and behavior in the field.

In addition, it is unclear how we should interpret the significant results obtained in this literature. There is a well-known bias to produce, submit, and publish significant results over insignificant ones, which leads to false positives and to an overrepresentation of spurious correlations (see, e.g., Rosenthal 1979; Ioannidis 2005; Simmons, Nelson, and Simonsohn 2011; Maniadis, Tufano, and List 2014; Miguel et al. 2014; Simonsohn, Nelson, and Simmons 2014). This has recently produced a wide-ranging replication crisis that represents a serious threat to the social and behavioral sciences (and to other sciences as well), with some analyses estimating the replicability of published research at less than 50% (see, e.g., Ioannidis 2005; Simmons et al. 2011; Open Science Collaboration 2015; Baker 2016).

This issue becomes much more problematic if the studies published in a specific field are not systematic, which we believe is currently the case with many studies in this lab-field research on the external validity of social preference games. As our systematic review shows (see Table A1), the typical study in this area reports the results of comparing one social preference game with one specific, or several related, field measures. It is also apparent that there is a very wide diversity in the variables, methodologies, and samples used in the different papers. A particularly worrying aspect is that the abstract and context-free nature of these games makes it very difficult to establish clear theoretical correspondences between the games and real social behaviors outside the lab, which can lead to an *ex post* rationalization of almost any social behavior as related to the games. We believe that this is clearly reflected in the wide variety of behaviors that have been associated to the same games, including things as diverse as: earning or spending money, fishing shrimps, drinking beer, participating in elections for parent representatives in schools, or registering books in a library. This lack of a strict theoretical mapping may also exacerbate publication bias by making it difficult for the

[2] See also Camerer (2011) for an interesting review of this and other literature related to external validity, and Dolan and Galizzi (2014) for a general discussion of lab-field experiments on pro-social behavior and in other areas.

researcher to justify specific null results as meaningfully related to the games. For all these reasons, we think it is crucial to conduct more systematic research on the external validity of social preference games.

In this paper, we present a systematic investigation of the external validity of social preference games, conducted by comparing behavior in a variety of games with a variety of situations created in the field, and also with 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-report measures include various pro-social tendencies shown in the past.

Rather than trying to establish a one-to-one correspondence between particular experimental games and specific field situations, which is (as indicated above) necessarily imprecise given the abstract and context-free nature of the games, we adopt the strategy of covering a variety of prominent social preference games and several relevant field behaviors to explore more broadly the extent to which the games are predictive of social behaviors shown in the field. The relationship between the games and the field situations we studied is further discussed in the next sections. The self-report measures of past social behaviors provide an additional layer to evaluate the explanatory ability of the games. To the best of our knowledge, this constitutes the most systematic and comprehensive lab-field 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 were 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 (together with our systematic review and meta-analysis) seriously question the external validity of social preference games 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 presenting 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 against actual social behaviors in the field and 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-report measures of past social behaviors. 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 complete a task that was unrelated to social preferences, and after exiting they were faced with an opportunity to behave pro-socially in one of five field situations.

This three-day lab-field structure allowed us to obtain all the information that we were interested in, while minimizing the possibility 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-report measures of past social behaviors

In the first experimental session of the week, upon their arrival to the lab, the participants were assigned anonymous 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 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. 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.[3]

In Session 1, the participants reported on their past pro-social behaviors using the Self-Report Altruism (SRA) Scale (Rushton, Chrisjohn, and Fekken 1981). This scale consists of 20 items, in which people are asked to state how frequently in the past they have done 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 B. Participants rated 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 completed another unrelated task, which consisted of watching and rating several videos.

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

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 cubicles, and then received more specific and detailed instructions for the tasks they would complete throughout the session. Because of the structure of some of the games played in the session, we needed the number of participants 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 in detail below) 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 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 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).

Participants first received general instructions on the seven-game structure and the general payment mechanism, followed by specific instructions before each game. 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 better fit 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 also makes the games more similar to the one-shot field situations they would face at the end of Session 3. Furthermore, one-shot games have been the focus of most of the previous lab-field studies on the external validity of social preference games.

Our participants were not given information or feedback about the results of the different games until the end, with the exception of the information they inevitably obtained from simply playing the games (i.e., 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 in the order specified below:

1) *Dictator Game 1 (DG1):* Two-player game in which Player 1 decides how to divide £10 between the self and Player 2. 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. 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 was determined by a participant who was Player 1 in a preliminary pilot session.[4]
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 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 was determined by a participant who was Player 1 in a preliminary pilot session.
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, and those who were Player 2 in UG1 were Player 2 again in TG1. Thus, 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.

[4] One randomly selected participant in Session 2 was then actually matched with that previous participant to determine his/her payoff in the game. 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. We used the same method in Trust Game 2 (see below).

These 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 include: *altruism* (Player 1 in DG1 and DG2); *positive 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 explain the behavior of all the players in all the games). This variety of games and behavioral constructs associated with 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 task they would complete during the session.

In this session, the participants worked on a 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 individually 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 encountering the field situations outside of the lab. We made sure that one participant left the lab approximately every three minutes to allow time for the previous participant to complete the field situation.

When the participants left the laboratory, they encountered one of five naturalistic field situations that provided an opportunity to behave pro-socially. Two involved helping and the other three involved donating money. The five situations were run consecutively, in the order specified below:

1) *Boxes:* A research assistant stood in an area outside the lab and told the participants that he needed help carrying 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 after they exited the lab 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 needed 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) *Children's Charity:* A research assistant stood outside the lab 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) *Environmental Charity:* This situation was exactly like the previous one, but with a different charity. This organization was a leading charity dedicated to protecting the environment. The money donated was actually sent to the charity.
5) *Lab Donation:* This situation was 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.

The participants always encountered the field situations in the same location, which was out of the lab premises, and one floor above the lab, in a place they had to cross on the ground floor of the building (the lab is located in the basement of one of the core teaching buildings of the university). The area selected is often used by the students' union and its associated charities to advertise events and conduct different reach-out initiatives, including signing petitions or raising funds. The location was therefore a very natural place for students to go through, and there was no apparent reason to think that the participants related the situations they encountered with the lab experiment.

The different field situations cover a variety of naturalistic environments, in which the participants were able to express their social inclinations. Helping others and giving money to others are representative of a large number of real-world circumstances related to social preferences. In addition, donations have been extensively used in behavioral and social science research (see Bekkers and Wiepking 2011; List 2011; Oppenheimer and Olivola 2011), and a number of authors have explicitly assumed that donations should be related to behavior in social preference games (see DellaVigna 2009). Helping behaviors have also been widely studied in the behavioral and social sciences (see, e.g., Darley and Batson 1973; Eagly and Crowley 1986; Levine, Norenzayan, and Philbrick 2001), and helping has been explored in the field as well, often using methodologies that are very much in line with ours (see, e.g., MacRae and Johnston 1998; Levine et al. 2001).

As explained in the introduction, we deliberately do not want to establish a strict one-to-one correspondence between our field situations and the behavior of specific players in specific games. The abstract and 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 several 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 *Children's Charity* situations, where there is no element of strategic interaction or reciprocity involved; (ii) positive reciprocity is likely to be part of behavior in the *Lab Donation* situation, because participants have just received money from the lab; (iii) cooperation is likely to be a relevant motive in the *Environmental Charity* 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 *Children's Charity*, where inequality between the children in need and the participants is an important feature of the situation.

2.4. Participants and sessions

All experimental sessions were conducted 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 a university-wide subject pool, which comprises about 5,000 people, mostly current and former students of the university. We used no other eligibility or exclusion criteria to select participants. All the experimental procedures were approved by the research ethics committee of the institution.

3. Results

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

3.1. Self-report 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 C1 in Appendix C.

[Insert Figure 1 here]

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.

[Insert Figure 2a here]
[Insert Figure 2b here]

The results are broadly in line with the patterns usually reported in the literature.[5] 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 £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.

[5] The variety of specific estimates found in the literature makes it difficult to discuss particular numerical results, but it is clearly possible to identify more general patterns typically observed in these games, which are very much in line with our results (see Camerer et al., 2003, for a review).

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

Table 1 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 Player 2 in the ultimatum games, reflecting that people who accept allocations in the ultimatum games are more likely 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.

[Insert Table 1 here]

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 monetary situations together (*Children's Charity*, *Environmental Charity* and *Lab Donation*). The number of participants in each situation, after removing the missing data, was 50 in *Boxes*, 44 in *Phone*, 59 in *Children's Charity*, 73 in *Environmental Charity*, and 48 in *Lab Donation*.[7]

[Insert Figure 3 here]

As Panel A shows, 88% of the participants that faced the *Boxes* situation helped the research assistant to carry the boxes to the basement, and only 12% chose not to help. In Panel B, we see that people were more divided in the *Phone* situation: 70% of the people lent their phones to the research assistant and 30% did not. Panel C shows the distribution of contributions in the *Children's Charity* 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 67% of the people did not give anything to the *Environmental Charity*. The other 33% gave amounts between £0.05 and £2.10, with a higher spike at £1. Finally, in Panel E, we have the 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 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 behaviors?

[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.

We now turn to the main question of whether the game decisions explain the self-report measures and the field behaviors. To being with, Table 2 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 related to money (Items 4, 5 and 13). The game variables were then correlated with both scores.

[Insert Table 2 here]

Table 2 shows that three of the eight game variables are significantly correlated with total SRA scores at the 5% level, and only one of the eight variables is significantly correlated with the monetary SRA score. The significant correlations are relatively low, with no correlation greater than 0.2. The game decisions that correlate significantly 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 relate to 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 3 contains pairwise correlations (Spearman's ρ) between the eight game variables and the five different field behaviors. We have also included three additional field variables that group together different field conditions: one joining the two situations related to helping (*Box* and *Phone*, with n = 94); one joining the three situations that had to do with donating money (*Children's Charity*, *Environmental Charity* and *Lab Donation*, with n = 180); and one putting all field situations together (n = 274). This increases considerably the power of our analyses, which is more limited in the individual conditions. Table 3 contains as well the average of each column (labeled as *Average 1*), representing the average correlation obtained in each field condition (individually or grouped); and it also contains the average of each column but excluding the two variables related to the behavior of player 2 in the ultimatum game, which are the only ones for which a negative correlation was found in Table1 (labeled as *Average 2*).

[Insert Table 3 here]

As Table 3 shows, only one out of the 64 correlations is statistically significant at the 5% level. It is a correlation of 0.54 between TG1 P.1 and behavior in the *Children's Charity* situation. This is likely to be a spurious correlation produced by randomness, given that it is only one out of 64 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).[8] None of the correlations obtained grouping the field variables to gain statistical power are significant. In addition, the correlations within the same variables change from positive to negative throughout the table 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 the behavior in the field situations that we analyzed. It is also interesting to note that the overall average correlation obtained in our systematic lab-field experiment (0.03) is clearly lower than the overall correlation resulting from the papers that report significant correlations in our systematic review and meta-analysis of the previous literature (0.27).[9]

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

Table 4 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 with the following entered as explanatory variables: (i) 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) the decisions made by only one half of the sample (UG1 P.1 and TG1 P.1); and (iii) the decisions made by 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 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.[10]

[8] We also calculated p-values correcting for the multiple comparisons performed in Table 3 using six established methods: Bonferroni (1935), Holm (1979), Hochberg (1988), Hommel (1988), Benjamini and Hochberg (1995), and Benjamini and Yekutieli (2001). The last of these methods (Benjamini and Yekutieli 2001) takes explicitly into account dependence between the variables (which makes it similar to other possible approaches, such as the ones discussed in Romano and Wolf 2005, 2010; Romano Shaikh, and Wolf 2008; or List, Shaikh, and Xu 2016). All six correction methods used further reduce the significance of the results to the extent of having no correlation that is statistically significant at the 5% level. Specifically, the p-value for the only significant correlation obtained in Table 3 (which is 0.0018 without correction) becomes 0. 0720 with all methods except Benjamini and Yekutieli (2001), where it becomes 0.3081.

[9] The 95% confidence interval of the overall average correlation obtained is [-0.12, 0.18], which shows that the average correlation of 0.27 obtained from the previous papers that report significant correlations is clearly rejected by our experiment.

[10] Note that the variables used in the second and third regressions in each column could never be included in the same model, because there is no overlap in observations between them. In addition, putting the variables of

[Insert Table 4 here]

As Table 4 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 other variables in the first column of Table 4 (DG1&2 P.1 and TG2 P.2) are significant at the 10% level. These results are broadly in line with the correlations reported in Table 2.

More importantly, the proportions of variance explained by the models in Table 4 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 5 contains a summary of the regression results obtained for the field behaviors. The table has seven columns, corresponding to the five different field situations plus two additional variables, one putting together the two situations related to helping (*Boxes* and *Phone*) and another joining the three situations that have to do with donating money (*Children's Charity*, *Environmental Charity* and *Lab Donation*). Each of the columns is constructed following the same three-regression structure explained for Table 5 with behavior in the field situation as the dependent variable. 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.

[Insert Table 5 here]

The results in Table 5 show that only two out of the 56 coefficients are statistically significant at the 5% level. One corresponds to the only correlation that was significant at 5% in Table 3 (TG1 P.1 in the *Children's Charity* column), and the other corresponds to a correlation that was significant at 10% in Table 3 (TG2 P.2 in the *Lab Donation* column). As mentioned in reference to Table 3, this statistical significance is likely to be the result of spurious correlations. None of the coefficients obtained grouping the field variables to gain statistical power are significant. Overall, the results in Table 5 are broadly consistent with the correlations reported in Table 3.

More substantially, the proportions of variance explained by the regression models are again very low. Most of them are below 0.07 (13 out of 21 overall, and 5 out of 6 in the regressions combining field situations to gain power) and many of them are close to zero. The variation in these proportions does not seem to follow any meaningful pattern and may also be the result of randomness.

either of those two regressions together with the variables in the first model would sacrifice half of the observations contained in the sample.

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.[11]

Finally, it is also interesting to look at the correlations between the SRA scores and the field variables, which are all non-significant at the 5% level, generally small, and they change signs within the same variables with no apparent meaningful pattern. The average correlation with the total SRA scores is 0.02; with the items that are related to money it is 0.09. This does not affect the analyses presented above, but it is an interesting element to take into account when interpreting them and drawing conclusions. We elaborate further on this aspect in the next section.

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-report measures and the field behaviors.

The overarching conclusion is that the games do a poor job explaining both the self-report 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 call for more, and more systematic, research on this issue.

Our results are particularly troubling in light of our systematic review and meta-analysis of the previous lab-field literature related to the games that we used in the experiment (see Appendix A).

[11] The regressions discussed in this section have been further investigated employing a broad range of statistical methods, including: regressions with one game variable at a time, 'stepwise' regressions with game variables added in sequentially, robust standard errors, log-transforming 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.

That review and meta-analysis shows that the typical study in this line of research reports the results of comparing one lab game to one specific, or several related, field variables. There is also a very wide diversity in the type of variables, methods, and samples used. And more importantly, the lack of a clear theoretical mapping of the context-free games into real-world situations, allows for very different variables to be rationalized as potentially related to the games (fishing shrimps, drinking beer, voting in school elections, etc.). This raises concerns about the possibility that some of the previous findings may be the result of spurious correlations. As explained in the introduction, there is a well-known bias to produce, submit, and publish significant results over insignificant ones, which can be particularly problematic if the studies are not sufficiently systematic. Even without taking this into account, our systematic review and meta-analysis show that only 39.5% of the lab-field correlations and 37.5% of the lab-field regressions reported show significant associations.

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 present study compared a variety of games with a variety of naturalistic field situations using the same sample of participants. Given that we focused on one-shot games and that we did not cover all existing social preference games, or all relevant field situations, our strategy could be extended in subsequent studies, for instance by exploring repeated games, looking into other game structures, or creating other field situations, 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 with patterns of reciprocal behavior occurring in different field environments (see List 2006; Stoop et al. 2012; Kessler 2013; 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 not answer the question of whether the games predict social behaviors that are relevant outside the lab. As discussed above, this lack of a clear theoretical mapping from games into field situations is a problem that plagues the previous literature.

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 with this specific claim. In our experience, few experimental economists would feel comfortable with the idea that they are merely studying how people play games that have no relevance to 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 to say 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, 2008; 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; Kessler and Vesterlund, 2015).

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), List and Reiley (2008), Roth (2008, 2015), Falk and Heckman (2009), Hennig-Schmidt et al. (2010), Camerer (2011), Harrison (2013), Al-Ubaydli and List (2015), Kagel (2015), or Kessler and Vesterlund (2015), 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 causal relationships, to closely reproduce conditions of theoretical models, and to replicate past findings. Furthermore, they can provide insights into 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. Thus, 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.

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 and punishment (e.g., Henrich et al. 2001; Andreoni and Miller 2002; Charness and Rabin, 2002; Fehr and Gachter 2000, 2002; Herrmann, Thoni, and Gachter 2008). There is, however, a more specific issue of whether the particular type of lab experimentation being conducted 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 such a revision should be done, but part of the answer may involve bringing more context into the lab, and constructing experimental environments that more closely resemble naturalistic situations of interest. After all, experimental economics and psychology have widely documented that subtle differences in the context can have profound effects on how people behave (e.g., Ross and Ward 1996; Cherry et al. 2002; Ariely et al. 2006; List 2007; Bardsley 2008; Stewart et al. 2015). This conclusion seems further reinforced by the fact that we did not find a significant correlation between our self-report measure of past social behaviors (the SRA scale) and behavior in our field situations. This is reminiscent of the person *versus* situation debate in personality and social psychology, and of the conclusion that personality measures are a poor predictor of behavior in specific situations. On the other hand, personality measures do a much better job predicting the average of various behaviors over time (see Fleeson 2004). This suggest another potentially interesting avenue for future research on the external validity of social preference games, namely looking into the prediction of average social behaviors over longer time periods. In any case, it is important to keep in mind that, 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).

Acknowledgments

Both authors contributed equally to this paper and are listed in alphabetical order.

The authors are very grateful to Paul Dolan for his encouragement and support when they embarked on this project at the LSE. They thank Abigail Barr, Pablo Branas-Garza, Antonio Cabrales, Monica Capra, Juan Camilo Cardenas, Gary Charness, Yan Chen, Robin Cubitt, Susann Fiedler, Antonio Filippin, Urs Fischbacher, Toke Fosgaard, Simon Gachter, Andreas Glockner, Daniela Grieco, Benedikt Herrmann, Heather Kappes, Menusch Khadjavi, Andreas Lange, John List, Sara Machado, Robert Metcalfe, Rosemarie Nagel, Chris Olivola, Giovanni Ponti, Wojtek Przepiorka, Holger Rau, Paul Smeets, Jan Stoop, Fabio Tufano, Roberto Weber, and the participants at the SPI Hub Conference in Chicago, 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, the London Experimental Workshop at Middesex University, and the NIBS conference at the University of Nottingham 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).

References

Abeler, J., and Nosenzo, D. (2015). Self-selection into laboratory experiments: Pro-social motives versus monetary incentives. *Experimental Economics*, 18, 195-214.

Al-Ubaydli, O., and List, J.A. (2015). On the generalizability of experimental results in economics. In G. Frechette and A. Schotter (Eds.), *Handbook of Experimental Economic Methodology*. Oxford University Press: Oxford.

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.

Ashraf, N., Bohnet, I., and Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, 9, 193-208.

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.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452-454.

Baran, N.M., Sapienza, P., and Zingales, L. (2010). Can we infer social preferences from the lab? Evidence from the trust game. NBER Working Paper 15654.

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.

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

Barr, A., Packard, T., and Serra, D. (2014). Participatory accountability and collective action: experimental evidence from Albania. *European Economic Review*, 68, 250-269.

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.

Benjamini, Y, and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1), 289-300.

Benjanini, Y, and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4), 1165-1188.

Bekkers, R., and Wiepking, P. (2011). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly*, 40(5), 924-973

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

Bellemare, C., Kroger, S., and van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4), 815-839.

Bellemare, C., Kroger, S, and van Soest, A. (2011). Preferences, intentions, and expectation violations: a large-scale experiment with a representative subject pool. *Journal of Economic Behavior and Organization*, 78, 349-365.

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.

Bluffstone R., Dannenberg, A., Martinsson, P., Jha, P., and Bista, R. (2015). Cooperative behavior and common pool resources: experimental evidence from community forest user groups in Nepal. World Bank Group Policy Research Working Paper 7323.

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

Bonferroni, C.E. (1935). *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato.

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. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. (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. (2011). 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.

Darley, J.M., and Batson, C.D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, 27(1), 100-108.

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.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic literature*, 47(2), 315-372.

Dolan, P., and Galizzi, M.M. (2014). Getting policy-makers to listen to field experiments. *Oxford Review of Economic Policy*, 30(4). 725-752.

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

Eagly, A.H., and Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological bulletin*, 100(3), 283-308.

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. (2016). Social dilemmas in the laboratory and in the field. *Journal of Economic Behavior & Organization*, 128, 85-96.

Ermisch, J., Gambetta, D., Laurie, H., Siedler, T., and Uhrig, S.C.N. (2009). Measuring people's trust. *Journal of Royal Statistical Society A*, 172(4), 749-769.

Exadaktylos, F., Espin, A.M., and Branas-Garza, P. (2013). Experimental subjects are not different. *Scientific reports*, 3.

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.

Fleeson, W. (2004). Moving personality beyond the person-situation debate the challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, 13(2), 83-87.

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.

Goeschl, T., Kettner, S.E., Lohse, J., and Schwieren, C. (2015). What do we learn from public good games about voluntary climate action? Evidence from an artefactual field experiment. University of Heidelberg, Department of Economics Discussion Paper 595.

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

Gurven, M., Winking, J. (2008). Collective action in action: prosocial behavior in and out of the laboratory. *American Anthropologist*, 110(2), 179-190.

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.

Hennig-Schmidt, H., Rockenbach, B., and Sadrieh, A. (2010). In search of workers' real effort reciprocity: A field and laboratory experiment. *Journal of the European Economic Association*, 8(4), 817-837.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, 91, 73-78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C.F., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N.S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F.W., Patton, J.Q., and Tracer, D. (2005). "Economic main" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795-815.

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.

Hill, K., and Gurven, M. (2004). Economic experiments to examine fairness and cooperation among the Ache Indians of Paraguay. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis (Eds.), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford University Press.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), 800-802.

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.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2), 383-386.

Hopfensitz, A., Miguel-Florensa, J. (2015). Mill ownership and farmer's cooperative behavior: the case of Costa Rica coffee farmers. Toulouse School of Economics.

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Med*, *2*, e124.

Kagel, J.H. (2015). Laboratory experiments: The lab in relationship to field experiments, field data, and economic theory. In G. Frechette and A. Schotter (Eds.), *Handbook of Experimental Economic Methodology*. Oxford University Press: Oxford.

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

Kessler, J.B. (2013). When will there be gift exchange? Addressing the lab-field debate with a laboratory gift exchange experiment. Wharton School at the University of Pennsylvania.

Kessler, J.B., and Vesterlund, L. (2015). The external validity of laboratory experiments: The misleading emphasis on quantitative effects. In G. Frechette and A. Schotter (Eds.), *Handbook of Experimental Economic Methodology*. Oxford University Press: Oxford.

Kolstad, J.R., and Lindkvist, I. (2012). Pro-social preferences and self-selection into the public health sector: evidence from an economic experiment. *Health Policy and Planning*, 1-8.

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

Lagarde, M., and Blaauw, D. (2014). Pro-social preferences and self-selection into jobs: evidence from South Africa nurses. *Journal of Economic Behavior & Organization*, 107, 136-152.

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.

Levine, R.V., Norenzayan, A., and Philbrick, K. (2001). Cross-cultural differences in helping strangers. *Journal of Cross-Cultural Psychology*, 32(5), 543-560.

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. (2011). The market for charitable giving. *The Journal of Economic Perspectives*, 25(2), 157-180.

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.

List, J.A., and Reiley, D. (2008). Field experiments in economics. In S.N. Durlauf and L.E. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Palgrave Macmillan Publishing.

List, J.A., Shaikh, A.M., and Yang, X. (2016). Multiple hypothesis testing in experimental economics. NBER Working Paper 21875.

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

Macrae, C.N., and Johnston, L. (1998). Help, I need somebody: Automatic action and inaction. *Social Cognition*, 16(4), 400-417.

Maniadis, Z., Tufano, F., and List, J.A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review*, 104, 277-290.

Markus, A.A.M., and Potgieter, J.L. (2015). Will you give him some money? External validity of social preferences in dictator games. University of Utrecht Bachelor's Thesis.

Miguel, E., Camerer, C.F., Casey, K., Cohen, J., Esterling, K.M., Gerber, A., Glennerster, R., Green, D.P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B.A., Petersen, M., Sedlmayr, R., Simmons, J.P., Simonsohn, U., and Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, 343(6166), 30-31.

Moher, D., Liberati, A., Teztlaff, J., Altmann, D.G., The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA Statement. *PLoS Medicine* 6(7): e1000097. doi:10.1371/journal.pmed1000097.

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.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Oppenheimer, D.M., and Olivola, C.Y. (Eds.). (2011). *The science of giving: Experimental approaches to the study of charity*. Psychology Press.

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

Riedl, A., and Smeets, P. (2015). Why do investors hold socially responsible mutual funds? University of Maastricht.

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.

Romano, J., Shaikh, A.M., and Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17(3), 417-442.

Romano, J., and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469), 94-108.

Romano, J., and Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics*, 38(1), 598-633.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638-641.

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.

Roth, A.E. (2008). What have we learned from market design? *Economic Journal*, 118, 285-310.

Roth, A.E. (2015). Is experimental economics living up to its promise? In G. Frechette and A. Schotter (Eds.), *Handbook of Experimental Economic Methodology*. Oxford University Press: Oxford.

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.

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366.

Simonsohn, U., Nelson, L.D., and Simmons, J.P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9, 666-681.

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. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, 61, 687-705.

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*.

Torres-Guevara, L.E., Schluter, A. (2016). External validity of artefactual field experiments: a study on cooperation, impatience, and sustainability in an artisanal fishery in Colombia. *Ecological Economics*, 128, 187-201.

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. (2012). 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: Systematic review and meta-analysis methodology

In conducting and reporting our systematic review and meta-analysis, we followed as closely as possible the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist and guidelines (Moher et al., 2009), as explained below.

A.1. Search strategy and key terms

Google Scholar was searched in July-November 2016 using the following combinations of key terms:

1) “social preference” AND “game” AND “external validity” (field TX all text) OR
2) “dictator game” AND “external validity” (field TX all text) OR
3) “ultimatum game” AND “external validity” (field TX all text) OR
4) “trust game” AND “external validity” (field TX all text) OR
5) “public good game” AND “external validity” (field TX all text) OR
6) “social preference” AND “game” AND “field behavior” (field TX all text) OR
7) “dictator game” AND “field behavior” (field TX all text) OR
8) “ultimatum game” AND “field behavior” (field TX all text) OR
9) “trust game” AND “field behavior” (field TX all text) OR
10) “public good game” AND “field behavior” (field TX all text).

A.2. Selection and exclusion criteria

The authors reviewed and assessed all the references systematically, following a two-stage strategy. In the first stage, the inclusion criteria were applied to the title, the keywords, and the abstract; in the second stage, the criteria were applied to the abstract and the full text. All the papers were independently assessed for inclusion by each of the authors. Differences in opinions between the authors were solved through discussion.

The two stages worked as follows. In the first stage, a study was included only if it satisfied the following three criteria:

1) The study was available (no broken link).
2) The study was written in English.
3) The study presented original empirical evidence. This criterion excluded theoretical or conceptual papers, reviews, commentaries, editorials, letters, or similar items.

Each article was sequentially evaluated against the three criteria, starting with criterion one and ending on criterion three. Whenever a criterion was not met, the article was excluded.

In the second stage, the abstract and the full text of the studies shortlisted in the first stage were screened and evaluated according to a further set of two criteria, so that a study was included only if it satisfied both of the following criteria:

1) The study looked into external validity using a within-subject design, i.e. considering both lab games and field behaviors for the same pool of subjects.
2) The study considered a dictator game, an ultimatum game, a trust game, or a public good game (or combinations of those games). This criterion focused the systematic review and meta-analysis on the games used in our experiment.

Each article was sequentially evaluated against the two criteria. Whenever either criterion was not met, the article was excluded. All articles meeting both inclusion criteria were retrieved and included in our meta-analysis.

We included both published and unpublished studies, for example studies in working paper or in dissertation form. If both published and unpublished versions of the study were available, we considered the published version. If different dates of the unpublished versions were available, we considered the most recent one.

To ensure that the set of studies retrieved was exhaustive and comprehensive, for each included study we also back-tracked and screened all the references cited in the article, applying the same inclusion criteria explained above.

A.3. Search results

The initial Google Scholar search resulted in a total number of n = 5,723 entries. After duplicates were removed, the resulting number of studies was n = 972. We then excluded the papers that were not accessible (n = 39), were not written in English (n = 1), or did not present any original evidence (n = 380). A total of 552 articles met all three criteria in this first stage of our selection strategy.

The abstract and the full text of the 552 studies shortlisted were then screened and evaluated. We first excluded the studies which did not use a within-subject design (i.e. combining lab games and field behaviors for the same pool of subjects) (n = 488). Then we excluded the studies that did not consider the dictator game, the ultimatum game, the trust game, or the public good game (n = 29). A total of n = 35 studies matched all the inclusion criteria in this second stage.

Back-tracking, screening, and evaluating the references cited in these 35 articles against the same inclusion criteria retrieved further n = 4 studies. So, at the end of the whole process, the systematic review resulted in a total of n = 39 selected studies.

Table A1 reports a complete list of these papers. Of the n = 39 papers, n = 29 are published: n = 27 in scientific journals (mainly economics journals, n = 23), and n = 2 in books. One study is a Bachelor’s thesis, and n = 9 are in working paper form.

The selection process and the number of papers excluded and included in each stage are summarised in the PRISMA flow chart in Figure A1.

[Insert Table A1 here]

[Insert Figure A1 here]

A.4. Meta-analysis

We then performed a meta-analysis of the n = 39 papers resulting from our systematic review, based on the information reported in the papers. First, each study was closely inspected and a number of pieces of information were extracted and collated into Table A1. In particular, Table A1 includes information on: the type and number of subjects, the setting, the lab games and field variables used, and whether or not a significant association between the lab and the field was found (categorized as "Yes", "No", "Marginally significant", or "Mixed evidence").

Second, we systematically extracted a number of statistical figures and collated them into an Excel file, included as a supplementary material. The Excel file consists of three different spreadsheets, as detailed below.

The first spreadsheet reports all the lab-field correlations reported in each study. In particular, it specifies whether or not each study reports the relevant correlation information and, if so, it reports the lab games used and the measures derived from them, the field variables analysed, the number of subjects, the correlations reported, the type of correlation (e.g., "Pearson", "Spearman"), the p-value, and also additional descriptive notes about the studies.

The second spreadsheet reports all the lab-field regressions reported in each study. In particular, it specifies the source of the information (i.e. the relevant table and page within the paper), the lab games used and the measures derived from them, the field variables analysed, the type of regression (e.g., "OLS", "Probit"), whether or not the regression has controls and how many, the estimated coefficients, the standard errors, the corresponding values of the statistics, the reported level of significance, the p-values (whenever reported in the paper), the number of subjects, the associated degrees of freedom, and also descriptive notes about the studies. The information that was missing in the papers but was possible to reconstruct from the available information is highlighted in red.

The third spreadsheet reports several summary statistics based on the lab-field correlation and regression figures collated in the other two spreadsheets. Specifically, we calculated, for each paper, the proportion of correlation and regression results reported that are significant at the standard 5% level, and also the average correlation reported. We also calculated the overall proportion of papers reporting correlation information (46.3%), the overall proportions of reported lab-field correlations and lab-field regressions that are statistically significant at the standard 5% level (39.7% and 37.5% respectively), and the overall average lab-field correlation obtained in all the papers (0.14) and in the papers that report significant correlations (0.27).

Appendix B: The Self-Report Altruism (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 C: Additional tables

Table C1: 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: 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*	*PGG*
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.

DG1&2 P.1 stands for Player 1 in Dictator Games 1 and 2; UG1 P.1 for Player 1 in Ultimatum Game 1; UG1 P.2 and UG2 P.2 for Player 2 in Ultimatum Games 1 and 2 respectively; TG1 P.1 for Player 1 in Trust Game 1; TG1 P.2 and TG2 P.2 for Player 2 in Trust Games 1 and 2 respectively; and PGG for the Public Good Game.

Table 2: 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.

SRAtotal stands for the total Self-Report Altruism (SRA) score; SRAmoney for a score including only the SRA items related to money.

DG1&2 P.1 stands for Player 1 in Dictator Games 1 and 2; UG1 P.1 for Player 1 in Ultimatum Game 1; UG1 P.2 and UG2 P.2 for Player 2 in Ultimatum Games 1 and 2 respectively; TG1 P.1 for Player 1 in Trust Game 1; TG1 P.2 and TG2 P.2 for Player 2 in Trust Games 1 and 2 respectively; and PGG for the Public Good Game.

Table 3: Correlations between game decisions and field behaviors (Spearman's ρ)

	Boxes	*Phone*	*Children's Charity*	*Environ. Charity*	*Lab Donation*	*All Helping*	*All Donations*	*All Conditions*
DG1&2 P.1	0.04	0.06	-0.25*	0.20*	-0.05	0.05	-0.06	-0.06
UG1 P.1	0.18	0.27	0.18	0.04	0.15	0.21	0.15	0.12
UG1 P.2	-0.09	0.14	0.22	-0.05	-0.06	0.08	0.04	0.07
UG2 P.2	-0.13	0.11	0.02	-0.13	0.06	0.05	-0.04	-0.01
TG1 P.1	0.15	0.28	0.54***	0.00	-0.12	0.22	0.15	0.13
TG1 P.2	0.35	-0.08	0.05	-0.12	-0.24	0.1	-0.11	-0.07
TG2 P.2	0.18	0.27*	-0.18	0.13	0.29*	0.21*	0.03	0.03
PGG	0.14	-0.04	-0.03	0.02	-0.08	0.04	-0.02	0
Average 1	0.10	0.13	0.07	0.01	-0.01	0.12	0.02	0.03
Average 2	0.19	0.15	0.05	0.05	-0.01	0.14	0.02	0.03

Notes:

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

DG1&2 P.1 stands for Player 1 in Dictator Games 1 and 2; UG1 P.1 for Player 1 in Ultimatum Game 1; UG1 P.2 and UG2 P.2 for Player 2 in Ultimatum Games 1 and 2 respectively; TG1 P.1 for Player 1 in Trust Game 1; TG1 P.2 and TG2 P.2 for Player 2 in Trust Games 1 and 2 respectively; and PGG for the Public Good Game.

All Helping, All Donations, and All Conditions group together, respectively, the two conditions related to helping, the three conditions related to giving money, and all conditions.

Average 1 is the overall average of the column; Average 2 is the average excluding the variables related to player 2 in the ultimatum games.

Table 4: 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:

The numbers reported in the first eight rows are regression coefficients from standard OLS regressions.

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

SRAtotal and SRAmoney are the dependent variables, and they stands for the total Self-Report Altruism (SRA) score and a score including only the SRA items related to money.

DG1&2 P.1 stands for Player 1 in Dictator Games 1 and 2; UG1 P.1 for Player 1 in Ultimatum Game 1; UG1 P.2 and UG2 P.2 for Player 2 in Ultimatum Games 1 and 2 respectively; TG1 P.1 for Player 1 in Trust Game 1; TG1 P.2 and TG2 P.2 for Player 2 in Trust Games 1 and 2 respectively; and PGG for the Public Good Game.

Var. Explained 1, 2 and 3 stand for the proportions of variance explained (R^2) in the three different regressions conducted for each dependent variable: one using as explanatory variables the game decisions for which we have observations for the full sample of participants (DG1&2 P.1, UG2 P.2, TG2 P.2, PGG); a second one with the variables with observations for half of the participants (UG1 P.1 and TG1 P.1); and a third one with the variables with observations for the other half of the participants (UG1 P.2 and TG1 P.2).

Table 5: Regression analysis field behaviors

	Boxes	*Phone*	*Children's Charity*	*Environ. Charity*	*Lab Donation*	*All Helping*	*All Donations*
DG1&2 P.1	-0.07	-0.08	-0.11	0.03	-0.06	-0.09	-0.04
UG1 P.1	0.24	0.11	0.07	0.01	0.14	0.19	0.03
UG1 P.2	142.81	0.59	0.76	-0.08	-0.08	0.79	0.08
UG2 P.2	-15.37	0.60	0.11	-0.19	-0.00	0.54	-0.05
TG1 P.1	0.05	0.27	0.20***	-0.01	-0.06	0.14	0.05
TG1 P.2	16.43	-0.05	-0.00	-0.02	-0.05	0.03	-0.03*
TG2 P.2	0.34	0.53*	0.01	0.01	0.22***	0.44*	0.01
PGG	0.09	-0.04	0.00	-0.01	-0.03	0.00	-0.00
Var. Explained 1	0.19	0.12	0.06	0.04	0.16	0.14	0.01
Var. Explained 2	0.04	0.10	0.32	0.00	0.11	0.06	0.04
Var. Explained 3	0.27	0.02	0.04	0.03	0.08	0.01	0.04

Notes:

The numbers reported in the first eight rows are regression coefficients, from standard logistic regressions in the case of the binary variables (not helping = 0, helping = 1), and from standard OLS regressions in the case of the continuous variables.

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

DG1&2 P.1 stands for Player 1 in Dictator Games 1 and 2; UG1 P.1 for Player 1 in Ultimatum Game 1; UG1 P.2 and UG2 P.2 for Player 2 in Ultimatum Games 1 and 2 respectively; TG1 P.1 for Player 1 in Trust Game 1; TG1 P.2 and TG2 P.2 for Player 2 in Trust Games 1 and 2 respectively; and PGG for the Public Good Game.

All Helping and All Donations group together, respectively, the two conditions related to helping and the three conditions related to giving money.

Var. Explained 1, 2 and 3 stand for the proportions of variance explained (R^2 for the OLS regressions, McFadden’s Pseudo-R^2 for the logistic regressions) in the three different regressions conducted for each dependent variable: one using as explanatory variables the game decisions for which we have observations for the full sample of participants (DG1&2 P.1, UG2 P.2, TG2 P.2, PGG); a second one with the variables with observations for half of the participants (UG1 P.1 and TG1 P.1); and a third one with the variables with observations for the other half of the participants (UG1 P.2 and TG1 P.2).

Table A1: Summary of lab-field studies on external validity of social preference games

	Subjects	*N*	*Setting*	*Lab game*	*Field variable*	*Lab-field significant association?*
Glaeser et al. (2000)	Undergraduate students	97	Harvard University	TG	i) Reservation value for an Envelope Drop game; ii) GSS survey question about trust; iii) index based on GSS survey question about trust; iv) survey question on trusting strangers; v) behavioral index; vi) self-reported trustworthiness; vii) honesty index; viii) hours volunteering	i): No for TG P.1; not reported for TG P.2. ii): No for TG P.1; yes for TG P.2. iii): No for TG P.1; yes for TG P.2. iv): Yes for TG P.1; not reported for TG P.2; v): Marginally significant for TG P.1; no for TG P.2; vi): Not reported for TG P.1; no for TG P.2; vii): Not reported for TG P.1; no for TG P.2. viii): No for TG P.1; yes for TG P.2.
Castillo & Carter (2002)	Urban and rural community members	283	KwaZulu-Natal, South Africa	DG, TG	Per capita household expenditure	Yes for DG and TG P.2 in urban areas; marginally significant for TG P.1 in urban areas; yes but negative for TG P.1 in rural areas; no for DG and TG P.2 in rural areas.
Fehr et al. (2003)	Representative sample of adults	147	Germany	TG	i) Survey question about belief that people are fair; ii) survey question about trustworthiness of others; iii) survey question about trust in others and in institutions; iv) survey question about benefit from generosity of strangers in the past; v) frequency of past trustful behavior.	i): No for TG P.1; no for TG P.2. ii): Mixed evidence for TG P.1; no for TG P.2. iii): No or marginally significant for TG P.1; no for TG P.2. iv) No for TG P.1; no for TG P.2. v) Yes for TG P.1; no for TG P.2.
Gachter et al. (2004)	University students and	277	Samara, Kursk, Zheleznogorsk,	PGG	i) GSS survey question about trust; ii) index based	i): No. ii): Yes. iii): Yes. iv): Marginally significant. v): Yes.

	non-students		Russia; Minsk, Grodno, Belarus		on GSS survey question about trust; iii) GSS survey question about fairness; iv) GSS survey question about helpfulness; v) survey question about trusting strangers; vi) index of 3 self-reported trusting behaviors: leaving door unlocked, lending money to friends, lending possessions to friends; vii) survey question about trustworthiness	vi): No. vii) No.
Hill & Gurven (2004)	Ache Indians (hunter-gatherers)	30	Arroyo Bandera, Paraguay.	UG, PGG	i) Individual total food production in kilograms; ii) household total food production in kilograms; iii) percentage of individual total food production shared outside nuclear family; iv) percentage of household total food production shared outside nuclear family.	i)-iv): No.
Cardenas & Carpenter (2005)	Urban slum dwellers	186	Bangkok, Thailand and Ho Chi Minh City, Vietnam	VCM (PGG)	Natural log of household expenditure	Yes for Bangkok; no for Ho Chi Minh City.
Holm & Danielson (2005)	Undergraduate economics students	86 + 51	Dar Es Salaam University, Tanzania; Lund University, Sweden	TG	i) GSS survey question about trust; ii) index based on GSS survey question about trust; iii) survey questions about trusting strangers; iv) survey question about past lending	i): No for TG P.1; no for TG P.2 in Tanzania; yes for TG P.2 in Sweden. ii): No for TG P.1 in Tanzania; no for TG P.2 in Tanzania; yes for TG P.1 in Sweden; yes for TG P.2 in Sweden. iii) No for TG P.1; not

					money; v) self-reported trustworthiness.	reported for TG P.2. iv): No for TG P.1; marginally significant for TG P.2 in Tanzania; no for TG P.2 in Sweden. v): Not reported for TG P.1; yes for TG P.2 in Tanzania; no for TG P.2 in Sweden.
Karlan (2005)	Female members of non-profit village banking organization (FINCA)	864	Ayacucho, Peru	TG, PGG	i) Default on the loan; ii) drop-out from the loan due to default or discipline; iii) total voluntary savings.	i): No for TG P.1; yes for TG P.2; no for PGG. ii): Marginally significant for TG P.1; yes for TG P.2; no for PGG. iii): Yes for TG P.1; yes for TG P.2; no for PGG.
Ashraf et al. (2006)	University students	177	Moscow, Russia; Capetown, South Africa; Boston, US.	TG (also DG and triple DG, but not reported)	Survey question about trust others not to cheat.	No for TG P.1; yes for TG P.2.
Bellemare & Kroger (2007)	Representative sample of adults	276+223	CentERpanel, the Netherlands	TG	Past life experiences when trusting others.	No for TG P.1; no for TG P.2.
Benz & Meier (2008)	University students	i)-ii): 99. iii)-iv): 83.	University of Zurich	DG	i) Past donations to social fund supporting students in financial difficulties or foreigners; ii) future donations to social fund supporting students in financial difficulties or foreigners; iii) past donations to charities; iv) future donations to charities.	i): Yes. ii): Yes. iii): Mixed evidence. iv): Yes.
Bouma et al. (2008)	Rural villagers in watersheds	92	Five villages in India	TG	i) Household investment in soil and water conservation; ii) household contribution to soil and water	i): No for TG P.1; not reported for TG P.2. ii): No for TG P.1; not reported for TG P.2.

					maintenance.	
Gurven & Winking (2008)	Tsimane forager-horticulturalist villagers	71 (DG) + 67 (UG)	Cosincho, Amazonian forest, Bolivia	DG, UG	i) Recorded number of days participating in construction of community well; ii) observed food sharing as percentage of production given to others outside nuclear family; iii) observed average number of interacting partners outside nuclear family; iv) recorded time spent in social visitation outside residential cluster; v) observed average beer provisioning to extra-household individuals; vi) observed average number of others' beer drinking parties outside nuclear family.	i): No for DG; no for UG. ii): No for DG; yes for UG. iii)-v) No for DG; no for UG. vi): Marginally significant for DG; no for UG.
Laury & Taylor (2008)	University students	125 + 68	Georgia State University	PGG	Price for contributing to local non-profit organization elicited through discrete choice experiment.	Yes.
Barr & Serneels (2009)	Manufacturing workers	164	Ghana	TG	i) Outcome per worker; ii) logarithm of earnings.	i): Yes. ii): Mixed evidence.
Ermisch et al. (2009)	Former respondents to the British Household Panel Study	173 (TG P.1) + 85 (TG P.2)	UK	Binary TG	i) Being active in an organization on a regular basis; ii) willingness to take risks in trusting strangers.	i): No for TG P.1; no for TG P.2. ii) Marginally significant for TG P.1, not reported for TG P.2.
Baran et al. (2010)	MBA students	463	Chicago Booth Business School	TG	i) Original gift amount to Chicago Booth Business School; ii) Original gift amount to Chicago Booth	i): Yes, sometimes marginally significant. ii): Yes. iii): No. iv): Mixed evidence.

					Business School if paid out outright; iii) Original gift amount to Chicago Booth Business School if defaulted on pledge; iv) Final gift amount to Chicago Booth Business School.	
Barr & Zeitlin (2010)	Primary school teachers	487	Uganda	DG	Proportion of contracted time allocated to teaching in previous month.	Yes
Carpenter & Myers (2010)	Volunteer firefighters, non-volunteer community members	205 + 189	Vermont, US	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-volunteer community member.	i): No. ii): Yes. iii): No. iv): No or marginally significant.
Rustagi et al. (2010)	Commons user groups	49	Bale region, Ethiopia	PGG	i) Forest management activities; ii) time spent on monitoring forest.	i): Yes. ii): Yes.
Carpenter & Seki (2011)	Fishermen catching shrimp in ‘pooling’ and ‘non-pooling’ boats	115	Toyama Bay, Japan	VCM (PGG)	Productivity of the fishing boats in terms of kilograms of fish caught per trip.	No.
Fehr & Leibbrandt (2011)	Fishermen selling shrimp in open-air markets	114	Villages near a lake in North Eastern Brazil	PGG	i) Hole size in shrimp traps; ii) survey measure of mesh size of fishnet; iii) real measure of mesh size of fishnet.	i): Yes. ii): Yes. iii): Marginally significant.
Lamba & Mace (2011)	Villagers in small-scale forager-	160	16 villages in Central India	PGG	Salt taken from a common pool.	Mixed evidence.

	horticulturist Pahari Korwa society					
Voors et al. (2011)	Poor rural villagers	99	South Eastern Sierra Leone	PGG	Survey attitudinal measures of i) illegal commercial mining, logging, and hunting; ii) illegal hunting of endangered species; iii) support to forest conservation.	i): Marginally significant. ii): No. iii): No or marginally significant.
Kolstad & Lindkvist (2012)	Medical and nursing students	40 + 40	Muhimbili University for Health and Allied Sciences, Dar Es Salaam, Tanzania.	DG (also a TG, but not reported)	Survey question on self-reported annual donations to the poor.	Yes.
Leibbrandt (2012)	Fishermen selling shrimp in open-air markets	148	Villages near a lake in North Eastern Brazil	PGG	i) Achieved price (for shrimp of similar quality); ii) stability of trade relations; iii) duration of trade relations; iv) trustworthiness signaling ability; v) quality misrepresentation.	i): Marginally significant. ii): Yes. iii): Marginally significant. iv): Yes. v): No.
Voors et al. (2012)	Poor rural villagers	453	South Eastern Sierra Leone	PGG	i) Contribution to community project fund for the village; ii) survey attitudinal measure of community labor; iii) survey attitudinal measure of village farm labor.	i): No. iii): No. iii): No.
Franzen & Pointner (2013)	University students	27 + 75	Universities of Cologne and Bern	DG	Returning a misdirected letter with money.	Yes.

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 (PGG)	Survey attitudinal measures of i) participation in any social organization; ii) meeting attendance; iii) participation in their decision planning; iv) hours in a month spent in them.	i)-iv): Yes for TG P.1; no for TG P.2.; no for VCM (PGG).
Barr et al. (2014)	Parents in primary schools	1,800	Albania	Binary PGG (also a DG and a third-party punishment game, but not reported)	i) Membership in voluntary organizations; ii) participation in community activities in past year; iii) participation in elections for parent representatives at individual level; iv) participation in national elections at individual level; v) participation in elections for parent representatives at district level; vi) participation in national elections at district level.	i): No. ii): No. iii): Yes, sometimes marginally significant. iv): Yes. v): Marginally significant. vi): Mixed evidence.
Lagarde & Blaauw (2014)	Final-year nursing students	343	South Africa	DG (also a modified DG with a patient receiver, and a modified DG with a poor receiver)	i) Self-reported job in rural area (i.e., a deep rural village, a rural village, or a small town in a rural area); ii) self-reported job in a rural health center.	i): No. ii). No.
Bernold et al. (2015)	University students	41 + 48 + 45 + 44	University of Zurich, ETH Zurich	PGG	i) Donation to an environmental charity within a neutral frame; ii) within a community frame; iii) within a Wall Street	i)-iv): No.

					frame; iv) within an environmental frame.	
Bluffstone et al. (2015)	Forest user group members	327	20 villages in Nepal	PGG	i) Likelihood of planting any trees in community forest in previous month; ii) likelihood of attending community forest user group meeting in previous month; iii) likelihood of adopting biogas; iv) number of hours spent monitoring and guarding community forest in previous month; v) number of trees planted by subject or their families in their private land in last five years; vi) number of trees planted by the subject in community forest in previous month.	i): Marginally significant. ii): No. iii): No. iv): No. v): Marginally significant. vi): Marginally significant.
Goeschl et al. (2015)	University students and subjects from the general population.	43 + 92	University of Heidelberg	PGG	Contribution to a project to reduce global CO_2 emissions.	No.
Hopfensitz & Miquel-Florensa (2015)	Coffee farmers members of a cooperative	46	Tarrazu and Turrialba regions, Costa Rica	PGG	Self-reported side-selling coffee in free market.	Mixed evidence.
Markus & Potgieter (2015)	University students	146	University of Utrecht	DG	i) Self-reported likelihood of being a blood donor; ii) an organ donor; iii) a volunteer; iv) a member of an association.	i)-iv): No.

Englmaier & Gebhardt (2016)	University students	20 + 16 + 13	University of Munich	PGG (also a ‘reverse’ PGG)	i) Average skill-adjusted speed in registering books in institute library in first 30 minutes of group task; ii) of individual task; iii) of task with no incentives.	i): Yes. ii): No. iii): No.
Riedl & Smeets (2016)	Individual investors	625	Mutual fund provider, the Netherlands	TG P.2	i) Likelihood of holding a Socially Responsible Investment (SRI) fund; ii) share invested in SRI fund.	i): Yes. ii): No.
Torres-Guevara & Schluter (2016)	Fishermen	152	Tasajera, Caribbean Coast, Colombia	PGG	i) Fishing impact index based on administrative data on fishing activities in past five years, estimated with experienced fishermen’s opinions; ii) estimated with scientists’ opinions.	i): No. ii): 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 (normally a synonym for PGG).

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

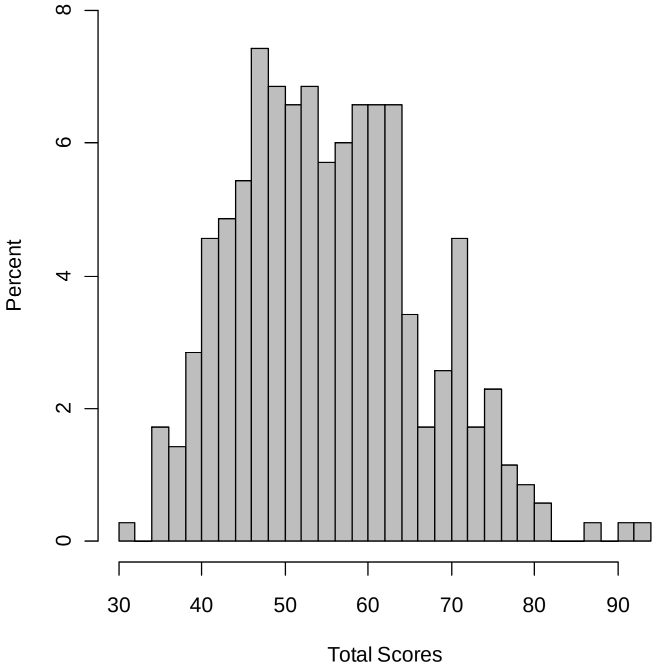


Figure 1: Total SRA scores

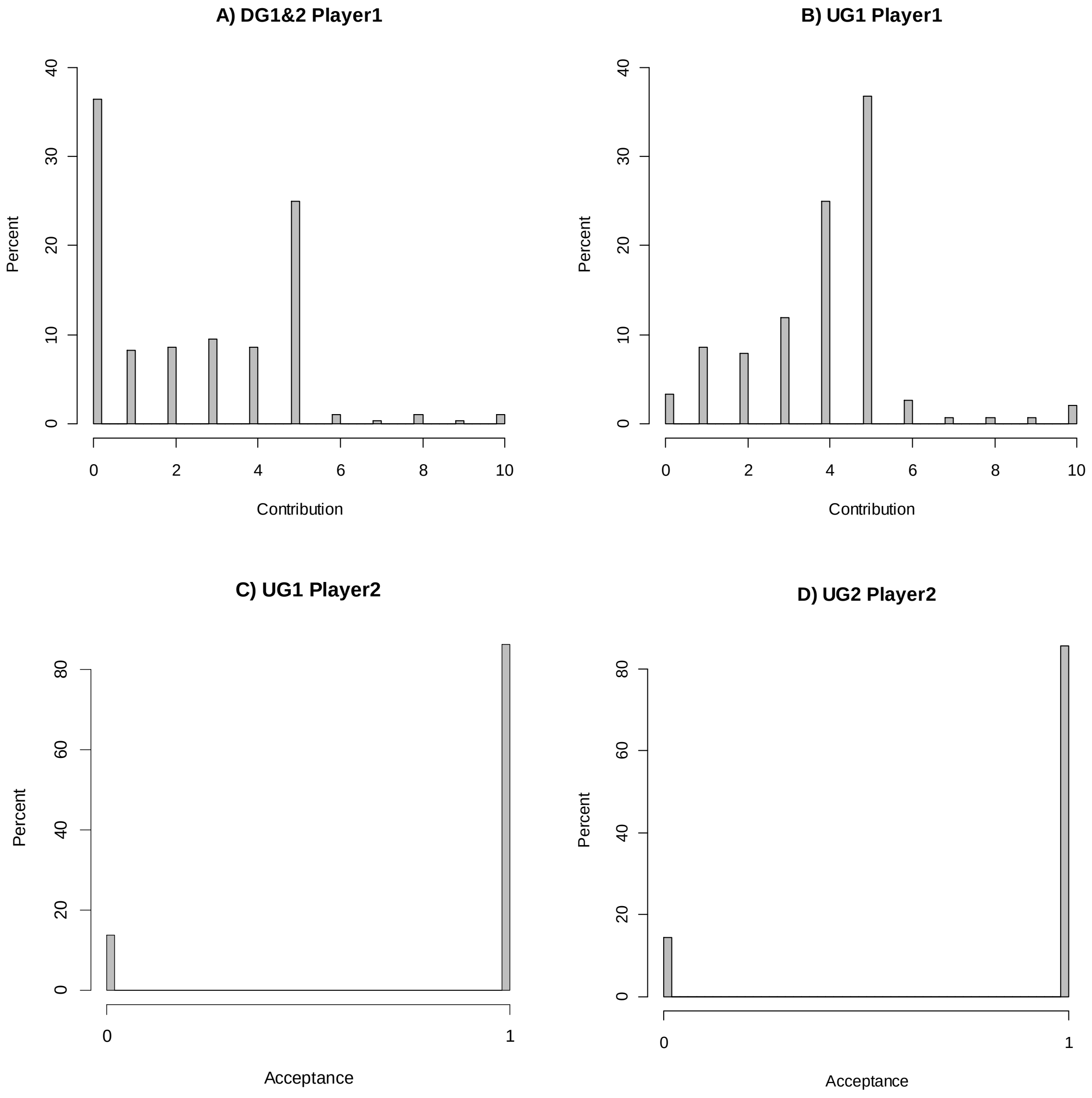
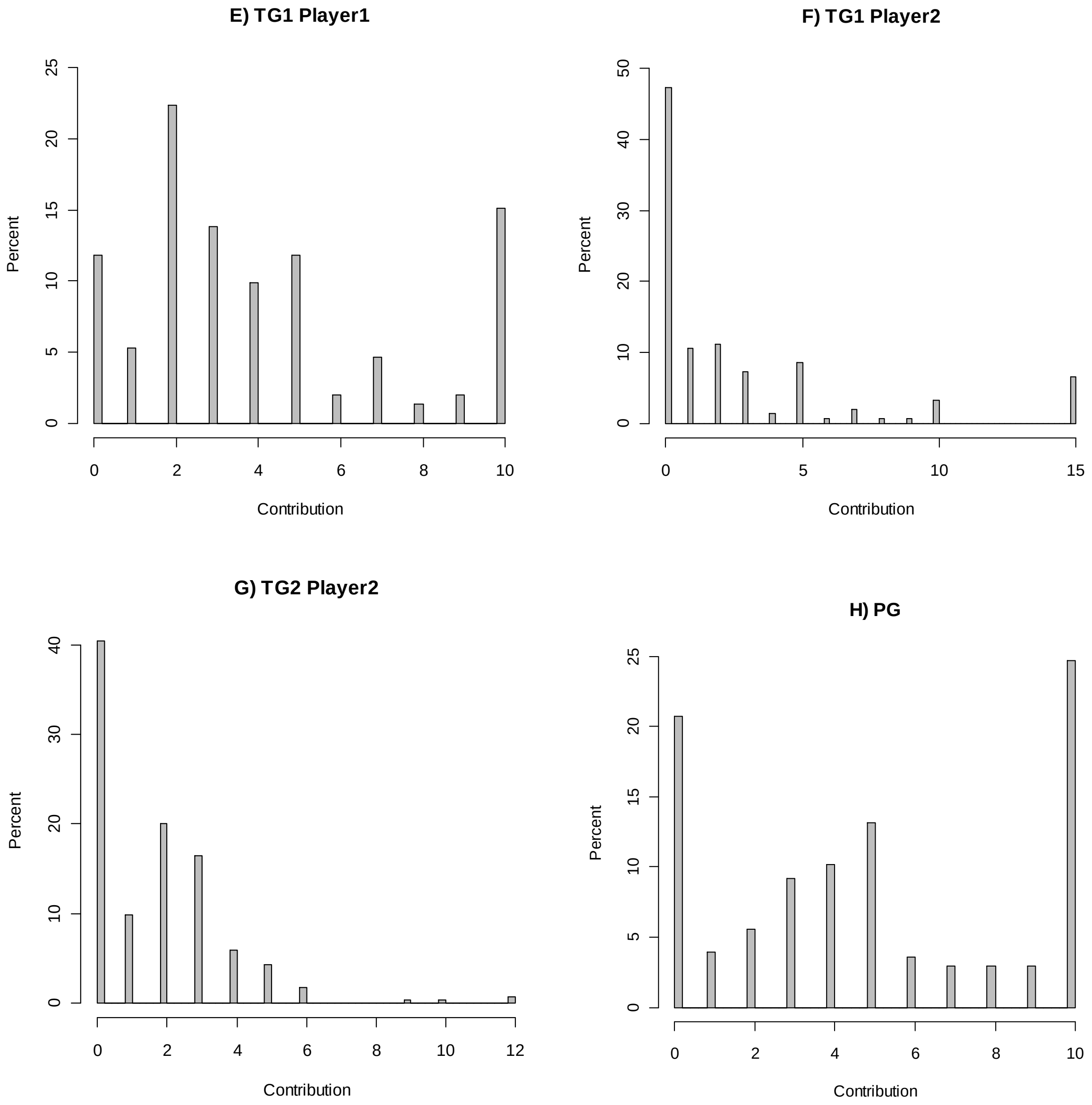


Figure 2a: Distribution of responses in first four game decisions



E) TG1 Player1

F) TG1 Player2

G) TG2 Player2

H) PG

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

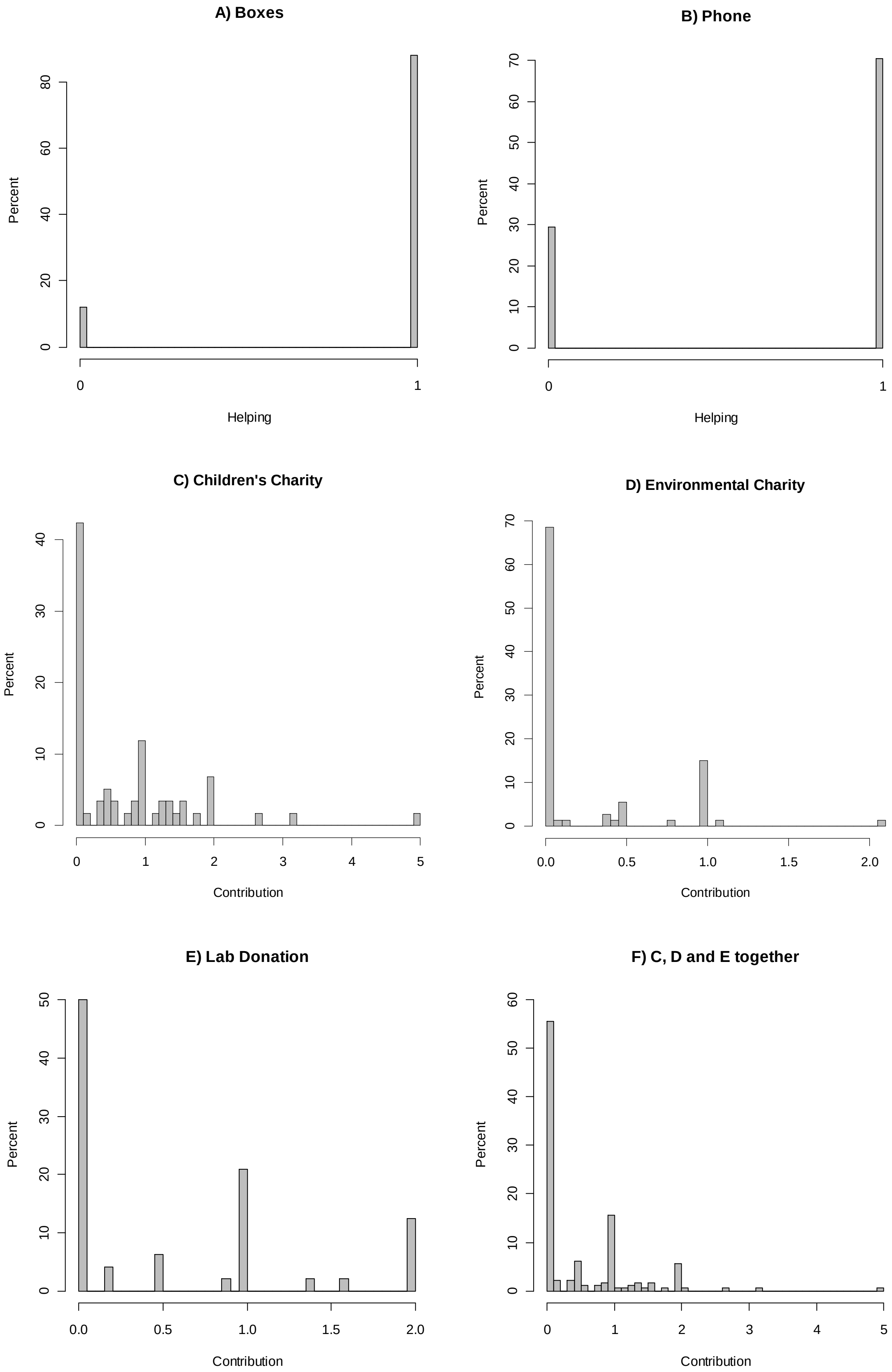


Figure 3: Distribution of behaviors in the field situations

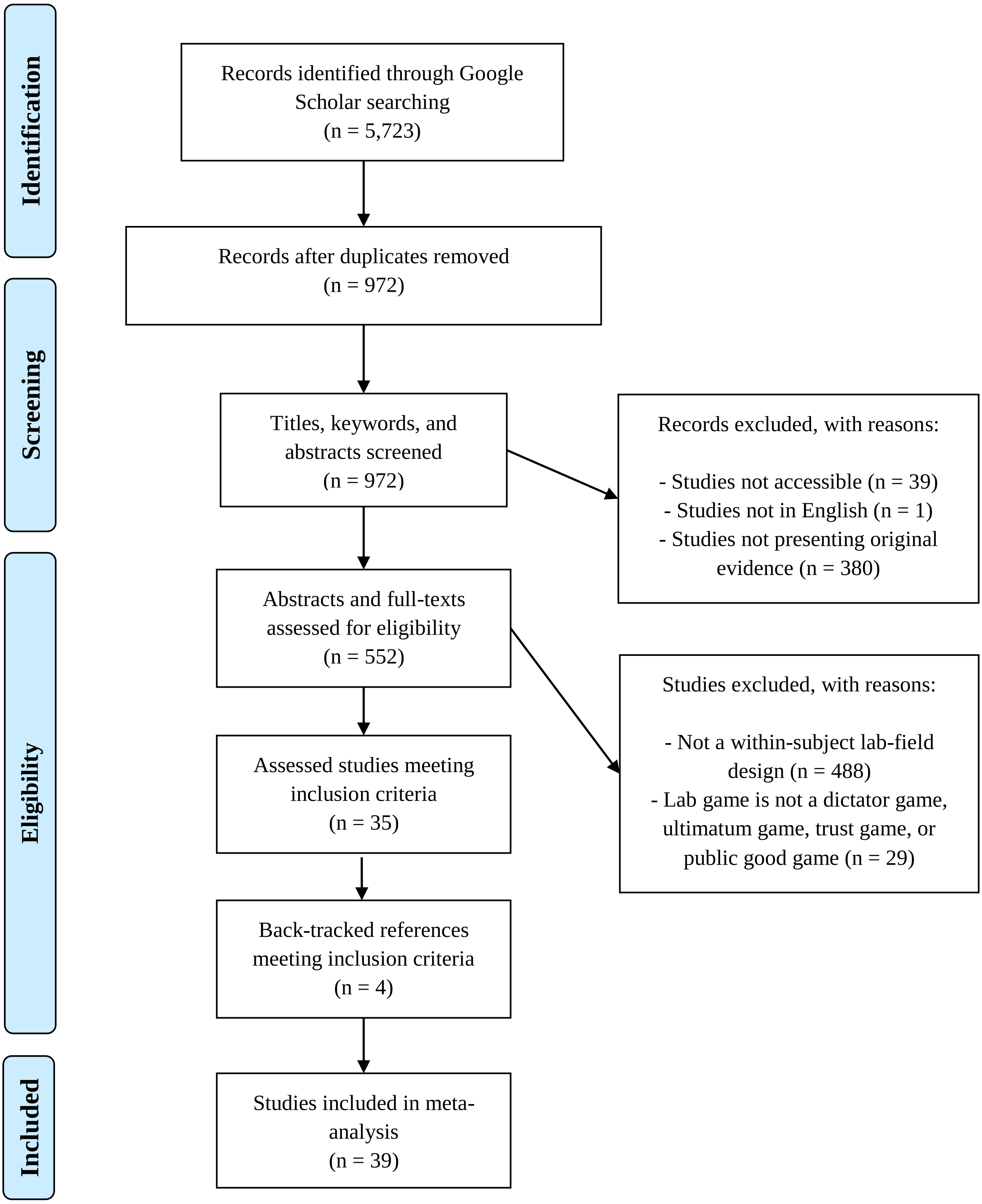

Identification
Screening
Eligibility
Included
Records identified through Google Scholar searching (n = 5,723)
Records after duplicates removed (n = 972)
Titles, keywords, and abstracts screened (n = 972)
Records excluded, with reasons:
- Studies not accessible (n = 39)
- Studies not in English (n = 1)
- Studies not presenting original evidence (n = 380)
Abstracts and full-texts assessed for eligibility (n = 552)
Studies excluded, with reasons:
- Not a within-subject lab-field design (n = 488)
- Lab game is not a dictator game, ultimatum game, trust game, or public good game (n = 29)
Assessed studies meeting inclusion criteria (n = 35)
Back-tracked references meeting inclusion criteria (n = 4)
Studies included in meta-analysis (n = 39)

Figure A1: PRISMA flow diagram of systematic review