



TESI DOCTORAL UPF / 2020



Essays on the External Validity of Social Preference Games

Xinghua Wang

TESI DOCTORAL UPF / 2020



Universitat
Pompeu Fabra
Barcelona

Essays on the External Validity of Social Preference Games

Xinghua Wang

Essays on the External Validity of Social Preference Games

Xinghua Wang

TESI DOCTORAL UPF / 2020

Thesis supervisor

Dr. Daniel Navarro-Martinez

Department of Economics and Business

Acknowledgement

First of all, I want to thank my supervisor, Daniel Navarro-Martinez, for his generous support and great guidance. He led me into the wonderful world of Experimental Economics, he inspired me a lot when I was looking for a research topic for my doctoral thesis, and he spared a lot of time to discuss with me detailed issues that I encountered in my research. I am glad and lucky to have him as my supervisor and collaborator.

Second, I would like to express my appreciation to Gert Cornelissen, Gael Le Mens, Veronica Benet-Martinez, Jose Apesteguia, Larbi Alaoui and Francesco Cerigioni for their helpful comments and suggestions to my research. I also want to thank Yan Chen, Jan Stoop, and Matteo Galizzi for being my doctoral committee members. I am really honored to have you on my committee and thank you for your time and support.

Third, I would like to thank my friends, including those in our behavioral research group and those Chinese PhD candidates I met in Barcelona. They offered me a lot of help and enriched my personal life. Thank you, Rahil Hosseini, Josep Gisbert, Thomas Woiczuk, Elia Soler, David Puig, Christoph Albert, Liang Fu, Lang Zhou, Donghai Zhang, Renbin Zhang, Shangyu Liu, Erqi Ge, Tingting Wu, Shi Feng, Yifei Lu, Yanan Pang and many others.

Fourth, I want to express gratitude to my family. They have been the most important support for me. I feel sorry for spending very little time accompanying them in the last six years. Thank you for your understanding and support of all the time.

Finally, I want to thank our university staffs for their assistance, including Marta Araque, Laura Agusti, Sara Fernandez, Pablo Lopez-Aguilar, Bruno Guallar, and Marta Molina. I also want to thank the Spanish Ministry of Economics and Competitiveness and Barcelona GSE for providing funding for my studies and research.

Abstract

The three chapters of this thesis investigate the external validity of social preference games. Chapter 1 reveals the context-dependent nature of human social behavior and shows that it is possible to make lab games much more predictive of field behavior by bringing the right contextual elements from the field into the lab. Chapter 2 shows that social preference games reach moderate correlations with a collection of daily pro-social behaviors when these behaviors are observed and averaged over a longer time period. This suggests that context-free games capture some fundamental aspects of pro-sociality in daily life and that future research on the external validity of these games should pay more attention to social behavior over extended time periods. Chapter 3 presents a systematic investigation of the external validity of social preference games at the workplace by comparing game behavior with organizational behaviors in a group of hotels. The results show that social preference games have a low predictive power in relation to organizational behavior.

Resumen

Los tres capítulos de esta tesis investigan la validez externa de los juegos de preferencias sociales. El Capítulo 1 revela la naturaleza dependiente del contexto del comportamiento social humano y muestra que es posible hacer que los juegos de laboratorio sean mucho más predictivos del comportamiento de campo si se traen los elementos contextuales correctos del campo al laboratorio. El capítulo 2 muestra que los juegos de preferencias sociales alcanzan correlaciones moderadas con una colección de comportamientos pro-sociales diarios cuando estos comportamientos son observados y agregados a lo largo de un período de tiempo más largo. Esto sugiere que los juegos libres de contexto capturan ciertos aspectos fundamentales de la pro-socialidad en la vida diaria y que la investigación futura sobre la validez externa de estos juegos debería prestar más atención al comportamiento social durante períodos más prolongados. El Capítulo 3 presenta una investigación sistemática de la validez externa de los juegos de preferencias sociales en el lugar de trabajo, comparando el comportamiento en los juegos con comportamientos organizacionales en un grupo de hoteles. Los resultados muestran que los juegos de preferencia social tienen un bajo poder predictivo en relación al comportamiento organizacional.

Preface

There is a growing concern about the extent to which laboratory games generalize to social behaviors outside the lab. The three chapters of this thesis systematically address this issue. Chapter 1 explores the reasons for the low external validity of context-free social preference games and discusses how to increase the external validity of laboratory studies that use these games. Chapter 2 tests whether social preference games can predict daily pro-social behaviors when they are aggregated over time. Chapter 3 systematically assesses the external validity of social preference games using a comprehensive data set collected at the workplace.

Chapter 1 first confirms the general finding in the literature: context-free games have limited external validity when addressing behavior in specific field domains. Then the chapter further shows that it is possible to make lab games much more predictive of field behavior by bringing the right contextual elements from the field into the lab. For this, we focus on dictator games and charitable giving in the field. We present a series of three experiments in which we face the same participants with different versions of the dictator game and with two different field situations outside the lab: one in which they can donate money to a charity and one in which they can show interest in volunteering. The games are designed to include step by step contextual elements that make them similar to our first field situation. We find a dramatic increase in the lab-field correlation with the first situation as contextual elements are incorporated. However, making the games more similar to the first situation does not increase correlation with the second one, showing that pro-social behavior is highly context-specific. Our results demonstrate that context-free games have limited predictive power in the field, but this gap can be substantially bridged by incorporating appropriate contextual elements from the field. This has wide-ranging implications for experiments in economics and it highlights the importance of establishing close links between lab and field research.

Chapter 2 shows that social preference games can achieve significant correlations with a collection of daily pro-social behaviors when these behaviors are observed and aggregated over a longer time period. In this study, we compared social preferences elicited in lab games with daily pro-social behaviors reported over one or two weeks. We included the dictator game, dictator game with taking option and public goods game; and we used the day reconstruction method and an experience-sampling method to track our participants' pro-social behavior. For each reported pro-social behavior, we have very detailed information, e.g., the general description, the recipient of the behavior, the time spent on it, the monetary cost, and a subjective evaluation of the size of the favor. All this detailed information was coded and rated by two impartial judges to get different pro-sociality scores. We found that, when the period of aggregation was long enough, the games significantly (although moderately) correlate with daily behaviors, even though those behaviors are diverse and do not share obvious characteristics with the games. This suggests that social preference games capture some underlying component of pro-sociality that is relevant across situations.

In chapter 3, our systematic investigation shows that social preference games have a low predictive power in relation to organizational behaviors. More specifically, in this

project, we worked with 5 high-end hotels in China. A large number of ordinary employees and managers working there participated in our study. All the participants played the same set of social preference games. They also answered survey questions about perceived respect and fairness inside the company, together with a big-five personality questionnaire. At the same time, all participants were rated by their supervisors in relation to their work performance using a standard Organizational Citizenship Behavior scale. Using these data, we compared decisions in different social preference games with different measures of work performance, controlling for perceived respect and fairness in the organization. We found that social preference games capture some characteristics of the participants at the group level, but these games do not have a strong predictive power in relation to organizational behaviors.

Table of contents

	Pag.
Abstract.....	v
Preface	vii
1. BRIDGING THE GAP BETWEEN THE LAB AND THE FIELD: DICTATOR GAMES AND DONATIONS	1
1.1. Introduction	1
1.2. Methods	5
1.2.1. The field situations	7
1.2.2. The lab games: Bridging the gap between the lab and the field.....	9
1.2.3. The structure of our study	11
1.3. Results	14
1.3.1. Donations in the field	14
1.3.2. Interest in volunteering.....	16
1.3.3. Decisions in the lab games	17
1.3.4. Do the lab games predict the field behaviors? ...	18
1.3.5. Relationship between the two field behaviors....	22
1.3.6. Predictive power of the big-five and empathy in the field	23
1.4. Conclusions	27
2. COMPARING BEHAVIOR IN THE LAB AND THE FIELD OVER TIME	35
2.1. Introduction	35
2.2. Study 1	38
2.2.1. Methods	38
2.2.2. Results	43
2.3. Study 2	53
2.3.1. Methods	54
2.3.2. Results	56
2.4. Conclusions	65
3. SOCIAL PREFERENCE GAMES GO TO WORK: TESTING FOR EXTERNAL VALIDITY IN THE WORKPLACE	75
3.1. Introduction	75
3.2. Methods	77
3.2.1. The social preference games	77
3.2.2. Organizational citizenship behavior	80
3.2.3. Survey data: other variables	81
3.2.4. Participants	82
3.3. Results	83
3.3.1. Game decisions	83
3.3.2. Organizational citizenship behavior	88
3.3.3. Perceived respect and fairness	89

3.3.4. Do game decisions predict performance in the workplace?	91
3.4. Conclusions	100
Bibliography.....	105

1. BRIDGING THE GAP BETWEEN THE LAB AND THE FIELD: DICTATOR GAMES AND DONATIONS

Joint with Daniel Navarro-Martinez

1.1. Introduction

Research on what is known as “social preferences” has attracted great interest in the last decades and has been one of the main building blocks of behavioral and experimental economics. Social preferences mainly fall into two categories: distributive preferences and reciprocal preferences. For agents with distributive preferences, their utilities are not narrowly defined over their own payoffs but over distributions of payoff in certain groups. That is, for the well-being of these agents, not only material wealth matters, but also some spiritual pursuit, e.g., equity, efficiency, and altruism. Reciprocal preferences then take into account decision makers’ desire to reward or punish others beyond mere material consequences, e.g., gift-exchange and punishing those who are unfair to the decision maker, even at a cost.

A central feature of this line of research is that it has been strongly based on studying behavior in “social preference games”, experimental games designed to capture different aspects of social behavior, such as altruism, inequity aversion, reciprocity, trust, cooperation, etc. Examples of these games include dictator game, ultimatum game, trust game, and public goods game, among others (see Camerer, 2003, for a review). These social preference games typically share some common characteristics: they are closely based on game-theoretic structures and have clear game-theoretic equilibria; they are deliberately as context-free as possible; and their outcomes consist in monetary payoffs for the players involved. This approach to social behaviors has been hugely influential and has become a benchmark for the study of human interaction in economics and beyond.

Context-free games have some clear virtues. They provide very stylized and controlled environments that allow researchers to achieve high levels of internal validity in the laboratory. That is, the variables being modified in experiments and the elements

being affected by them can be very clear and can be very tightly controlled. Moreover, this facilitates that different research groups use exactly the same tools in the same way and that evidence can quickly accumulate.

However, the artificial and context-free nature of these games raises one very fundamental question: are social preference games externally valid? In other words, do social preference games tap into the principles that determine social behavior outside the laboratory? To the extent that the aim of economics, and of the social and behavioral sciences more generally, is to understand real-world social behavior outside the lab and not how people play artificial games, this is a crucial question to respond. There is currently increasing concern that social preference games cannot deliver on external validity. Levitt and List (2007) were the first to systematically discuss this concern (but see Loewenstein, 1999, for a shorter discussion). The authors presented a theoretical framework that can be used to organize different factors that are likely to limit the external validity of social preference games, including differences between lab games and field settings in terms of scrutiny, anonymity, context, stakes, participants, and restrictions on choice sets and time horizons. After Levitt and List's paper, however, there has only been a relatively small number of papers that have investigated the issue of external validity (for reviews, see Camerer, 2011; Galizzi and Navarro-Martinez, 2019). In a meta-analysis of this literature, Galizzi and Navarro-Martinez (2019) concluded that only 39.7% of the lab-field correlations reported in the papers they analyzed were statistically significant, and the average lab-field correlation obtained was 0.14. The authors also conducted a systematic experiment comparing various lab games against several field behaviors and found no significant correlation between them. It is probably too soon to reach final conclusions about the external validity of all these games, but it seems safe to say that social preference games have issues of external validity.

We actually do not find these external validity problems too surprising, given that behavioral research has widely documented that context plays a crucial role in economic behavior (see, e.g., Slovic, 1995; Ariely et al, 2003, 2006; Lichtenstein and Slovic, 2006; List, 2007; Stewart et al., 2015). This also relates to the long-standing person-situation debate in psychology, in which social and personality psychologists have shown that the cross-situational consistency of behavior is typically very low and behavior is highly

dependent on particular situational cues (see Mischel, 1968; Ross and Nisbett, 1991; Fleeson, 2001, 2004). Overall, if behavior is so determined by contextual elements, it seems logical that context-free lab games cannot provide a good account of behavior when it is put in context outside the lab.

Given the context-free nature of social preference games and the context-dependent nature of behavior, one interesting possibility arises: it might be possible to bridge the gap between the lab and the field and improve the external validity of the games by introducing contextual elements in the lab. This brings us to the main goal of our paper. In this paper, we show that it is possible to make lab games much more predictive of field behavior by taking the right contextual elements from the field and incorporating them to the games. For this, we focus on the dictator game (Kahneman et al., 1986) and on charitable giving in the field. The dictator game is one of the most influential games in social preference research and arguably the simplest one. It is also a game that has been conceptually related to real-world social behaviors, such as charitable giving (e.g. Eckel and Grossman, 1996; Branas-Garza, 2006; Benz and Meier, 2008; Carpenter et al., 2008; Carpenter and Myers, 2010; Konow, 2010; Kolstad and Lindkvist, 2012), and a game in which behavior has been shown to be sensitive to contextual elements. For instance, giving in the game has been shown to be affected by the degree of anonymity (Dana et al., 2006; Franzen and Pointner, 2012), adding a taking option (List, 2007; Bardsley, 2008), earning the monetary endowment (List, 2007), the presence of verbal feedback and face to face interaction (Ellingsen and Johannesson, 2008; Andreoni and Rao, 2011; DellaVigna et al. 2012; Andreoni et al., 2017), or the type of recipient of the shared amount (Eckel and Grossman, 1996; Konow, 2010). All this makes the dictator game a suitable candidate for our investigation.

In the field, we decided to focus on charitable giving, which is a domain of notable economic and social relevance. The estimated amount of donation to charities in the USA in the year 2018 was \$428 billion (Giving USA 2019 report); the annual estimated amount in Europe is € 87.5 billion (Hoolwerf and Schuyt, 2017). Charitable giving is also a widely researched topic in economics (see, e.g., Andreoni, 1989, 1990; Glazer and Konrad, 1996; Okten and Weisbrod, 2000; Auten et al., 2002; Karlan and List, 2007; List, 2011; DellaVigna et al., 2012; Andreoni and Payne, 2013; Andreoni et al., 2017), and like we said, a domain that has been directly linked to the dictator game.

So, we investigated if it is possible to make the dictator game more predictive of charitable giving in the field. To this end, we run a series of three interconnected experiments in which we presented the same participants with different versions of the dictator game and with two naturalistic field situations that we created, in which they could behave pro-socially. In the first field situation, participants were approached by a solicitor and had the opportunity to donate money to charity; in the second one, they could show interest in volunteering by checking information about volunteering opportunities at various charities. Our different versions of the dictator game were designed to incorporate step by step additional contextual elements that made the games more similar to our first field situation. In particular, we incorporated the following three elements: a recipient in real need (as opposed to another student), a monetary endowment that was earned by the participant (as opposed to simply assigned by the experimenter), and face to face interaction (as opposed to anonymous giving).

Our approach bears some similarities to that of List (2006), Stoop et al. (2012) and Stoop (2014), where the authors designed lab games to match particular field environments. None of these papers, however, correlated lab and field behaviors for the same sample of participants.

We found a dramatic increase in the correlation between the lab games and our first field situation as more contextual elements were incorporated. These elements, however, did not increase the correlation with our second field situation (which they were not intended to address), confirming that pro-social behavior is highly context-specific. Our results show that context-free lab games have very limited predictive power in the field, but this power can be very substantially increased by incorporating appropriate contextual elements from the field to the games. Our interpretation of these findings is that social preference research should move in the direction of including more context in laboratory experiments and be more explicit about the types of field behaviors it aims to address. Combining lab and field experiments can also be an important way to reduce external validity issues. Remaining in the realm of context-free games could put social preference research at risk of being disconnected from real-world behaviors.

The rest of the paper is organized as follows: Section 2 describes our research methods, Section 3 presents the results, and Section 4 discusses our findings and concludes.

1.2. Methods

We conducted a series of three interconnected experiments, so we will explain the methods of all three together. We have three main aspects to explain: 1) the two field situations that will serve as our benchmark of field behavior that the lab games will aim to predict, 2) the different lab games that we created to approach step by step one of the field situations, and 3) how all these elements were organized in the three different experiments and in the different experimental sessions. These three things are explained in subsections 1.2.1, 1.2.2 and 1.2.3, respectively.

1.2.1. The field situations

1.2.1.1. Charitable giving

The first situation took place on a large square inside a university campus. A research assistant solicited donations for a well-known charity devoted to helping children in developing countries (to distinguish this charity from others we refer to later, let us denote this one as “Charity F”).¹ This square is a popular location in the university, where students gather between classes, and it is often used to organize various activities, including collecting donations for different causes. So, this was a natural place to find a charity solicitor. The square was also next to one of the rooms where we conducted our experiment, so that we knew participants had to walk through the square when they left the room at the end of one of the sessions.

¹ To choose the four different charities involved in our experiments, we run a pilot survey with 106 students, in which we asked them some questions about the popularity of 12 charities operating in the city (including a hypothetical allocation of €100 among these charities). All four charities chosen for the experiments were known by over 96% of the students. Moreover, these four charities were allocated similar amounts of money.

The assistant (always female to control for and check gender differences) was wearing an official university T-shirt and was standing next to a table with a professional charity bucket that had a large sticker with the logo of the charity. She also had a laminated color-printed leaflet with information about the charity and its projects. Participants left the room one by one at the end of the session and, as they approached the location of our assistant, she spoke to them using always the same initial words: "Hello, I am collecting donations for Charity F, would you like to learn more?" In case the participants wanted more information, she was instructed to show them the leaflet. If the participants decided to donate, they simply put the money they wanted inside the bucket, and this was typically the end of the interaction. Then the assistant recorded the donations made by the participants, together with any other relevant comments about their behavior.

Before implementing this field situation, we were granted approval by the university and the charity. At the end of the study, all the money donated was sent to the charity.

1.2.1.2. Interest in volunteering

For the second field situation, a few weeks after the experiment, we sent our participants an email with a link to a website where they could check information about volunteering opportunities at three different charities. The email was sent from an address linked to the university but unrelated to our lab or to our experiments. The subject of the email was the name of the university plus the word "volunteering". The email said:

"We have constructed a platform that provides information to students at our university about volunteering opportunities at charities. Each time we select three volunteering opportunities that suit students. If you are interested, please click the following link to learn more."

If participants clicked the link, they were directed to a website where they could see the names of three real volunteering projects managed by three real charities. Participants were able to click the volunteering project they wanted and go to a page where they could see a brief description of the project. At the bottom of that page,

people were asked “Are you interested in volunteering for this program?” They could answer “yes” or “no” to this question. This was followed by a page where participants could click a link to go to the charity’s website if they wanted to learn more about that particular volunteering project or to sign up for it. After having checked the information about one volunteering project, participants were automatically redirected back to the initial page where they could click on the links for the other volunteering projects if they wanted. This set-up allowed us to measure people's interest in volunteering by tracking their information search behavior (i.e., the links they clicked).

The three charities we worked with in this field situation were Charity F and two other well known charities (selected as indicated in footnote 1). One of the projects was about “digital volunteering”. This allowed enrollees volunteering at home and collaborating in many activities, such as translating documents, helping to create and manage websites, or working with old people and AIDS patients from a long distance. The other two projects needed volunteers to help organize events or do office work.

1.2.2. The lab games: Bridging the gap between the lab and the field

To investigate whether dictator games can be made more predictive of field behavior by bringing contextual elements from the field into the lab, we created a series of four different games. Each of them incorporates one additional field element that makes the lab game more similar to our first field situation. Our methodology requires us to focus on one single field situation to try to approximate its specific characteristics. We chose our first field situation for this because, as explained in the introduction, charitable giving is a domain of high economic and social relevance, an extensively researched topic, and a setting that has been directly linked to the dictator game. Our second field situation, which has markedly different characteristics, will serve as a comparison benchmark to illustrate the importance of the specific contextual features of the situation.

To create our games, we focused on three main aspects in which the standard dictator game (as typically implemented in economic experiments) differs from our first field situation: 1) participants in the standard dictator game have to allocate "house money" assigned to them by the experimenters, as opposed to "earned money" that belongs to

them; 2) they share the money with another study participant, as opposed to people in serious need; and 3) they do not engage in face-to-face interaction when making their donation. There are of course more differences between the standard dictator game and our first field situation, but we decided to focus on these three because we believe they are of particular relevance, and all three have been shown to significantly affect behavior in dictator game settings (e.g., see List, 2007; Eckel and Grossman, 1996; Andreoni et al., 2017).

Table 1 summarizes the structure of our four games. Game 1 in the series was the standard dictator game. So, one player (the "dictator") was endowed with €5 and had to split that amount between herself/himself and a passive second player (the "recipient"). In this case, both players were randomly matched participants. In Game 2, we changed the recipient of the money to a charity involved in medical-humanitarian actions, which we will call "Charity L" to distinguish it from the charity we worked with in the field. In Game 3, participants first earned their money (always €5) in a real-effort task and then decided how much to give to the same charity. The real-effort task involved entering sequences of letters and symbols on the computer. In Game 4, we further included face-to-face interaction. Specifically, instead of inputting the money they wanted to donate on the computer, participants had to give the money they wanted to donate to an assistant (always female) collecting donations for the same charity, who was standing outside the lab with a bucket. More details of these games will be explained in the following subsection.

This pyramidal structure of our experimental design allows us to investigate the role of the aforementioned three factors in affecting the predictive power of the game decisions. Table 1 summarizes the main features of the games.

Table 1. The Lab Games

Games	Description
Game 1	Standard dictator game
Game 2	Charity as a recipient
Game 3	Charity + earned money
Game 4	Charity + earned money + face to face interaction

1.2.3. The structure of the study

In this subsection, we will explain how all the lab games and the field situations were organized in the three different experiments that we conducted and in the different experimental sessions. We will start by describing the general structure and procedures and then we will go into the details of each experiment. Table 2 summarizes the overall structure of the three experiments.

Table 2. Structure of the study

	Day 1 (Lab games)	Day 2 (Field)	Day 3 (Field)	Observations
Experiment I	Games 1 & 2 (randomized)	Solicitor	Email	n = 102
	Control group			n = 46
Experiment II	Game 3			n = 101
Experiment III	Game 4			n = 98
	Control group	n = 45		

1.2.3.1. General structure and procedures

As Table 2 shows, each experiment comprised three different days. Participants were required to come to the lab on the first two days. In the first session (Day 1), people had to play one or two of our four lab games (as shown in Table 2 and explained in detail in the next subsections).² In addition, they responded to some psychological questionnaires.. They filled out the Interpersonal Reactivity Index scale (Davis, 1980), which measures four dimensions of empathy: Perspective Taking, Empathic Concern, Fantasy, and Personal Distress. These aspects of empathy are important determinants of social interaction and they have been shown to correlate with pro-social behavior (e.g.,

² All the lab games were implemented using oTree (Chen et al., 2016).

Eisenberg and Miller, 1987; Dovidio and Penner, 2001; Borman et al., 2001; Penner, 2002). They also responded to a questionnaire to measure the Big-Five personality dimensions (John, 1991): Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These five aspects of personality are fundamental constructs in psychology and they have been shown to explain a wide variety of behaviors (e.g., Barrick and Mount, 1991; Paunonen, 2003; Giluk and Postlethwaite, 2015). The goal of these questionnaires is to allow us to conduct a deeper analysis of the determinants of the pro-social behavior we observe in the lab and in the field, and to compare their predictive power in the field with that of the games. We also elicited some basic demographic variables (age, gender, study program, and citizenship).

In Experiments I and III (the ones with the games that were least and most similar to our first field situation), we also included control groups of subjects who did not play the games on Day 1, but participated in the rest of the experiment. This allows us to check if their field behaviors on Days 2 and 3 were different from the ones of participants who had previously played the games. On Day 1, instead of the games, these participants responded to other tasks unrelated to social preferences.

A few days later, for the second session (Day 2), the participants came to a different room, which was temporarily used as our lab and was next to the university square described in Section A.1. On this day, all participants responded to tasks that were unrelated to social preferences (and will not be used as part of this paper). These tasks included cognitive ability and cognitive reflection questions, and hypothetical choices between gambles.

At the end of the second session, people were paid € 15 in cash for their participation in both lab sessions. These € 15 were always paid using exactly the same bill and coin denominations: two € 5 bills, one € 2 coin, two € 1 coins, one 50 cents coin, two 20 cents coins, and one 10 cents coin. These denominations can be used to make up any amount from € 0 to € 15 in increments of € 0.1, which assured that all participants had cash to make any donation to charity up to € 15 when they faced the first field situation. We paid people one by one and made sure that each participant left the room approximately every three minutes, so that there was enough time to complete

the field situation. When participants left the room (after the session and the payment had finished) and were walking through the square, they encountered the first field situation as described in Section 1.2.1.1.

A few weeks later, the participants received the email with the link to information about volunteering opportunities that we used to implement the second field situation (as described in Section 1.2.1.2). In case some participants did not notice the email, we sent an identical reminder about one month after the first email.

1.2.3.2. Experiment I

In Experiment I, participants played two of our four lab games on Day 1: the standard dictator game (Game 1) and a modified game in which the recipient was Charity L (Game 2). In these two games, people made decisions about splitting a € 5 endowment between themselves and either another randomly matched participant (Game 1) or Charity L (Game 2). Game 1 was played for two rounds. In the second round, the roles of first and second player were reversed and people were re-matched with other participants. So, all the subjects got to play the game as the first player, who had to decide how to split the money. In Game 2, participants were first shown a brief description of Charity L, and then they made their decision of how much of their € 5 endowment to donate. The two lab games were presented in a random order. In addition to the games, on Day 1, participants also responded to the Interpersonal Reactivity Index and the Big-Five personality questionnaires and provided their demographic information.

At the end of the Day 1 session, one of the game decisions was randomly selected for each participant and they all received the amount resulting from the game. Then subjects went on to participate in Days 2 and 3, which were the same across the three experiments, as described in Section 1.2.2.

Experiment I included a control group of people who did not play the games on Day 1, as indicated in Section c.1. This control group can be used to check if playing the games on Day 1 affected in any way the field behavior of the participants on Days 2 and

3. Instead of the games, on Day 1, people in the control condition responded to a series of hypothetical inter-temporal choices which were unrelated to social preferences, and they were paid a fixed amount of € 3 at the end of the session.

1.2.3.3. Experiment II

On the first day of Experiment II, participants played Game 3, which was like Game 2 (i.e., with Charity L as the recipient) but introducing the element of earned money. To earn their money, subjects did a real effort task which consisted in entering seventeen sequences of letters and numbers on the computer. It usually took the participants 8 to 10 minutes to finish this task. This real effort task has been used in many previous experiments (e.g., Dickinson, 1999; also see Charness et al. (2018) for a review of all types of real effort tasks). All participants received € 5 for completing the real effort task. Then, like in Experiment I, they were shown the description of Charity L and were asked how much of their € 5 they wanted to donate.

As in Experiment I, participants responded to the Interpersonal Reactivity Index and the Big-Five questionnaires and provided demographic information. They also participated in Days 2 and 3, as in the other experiments.

1.2.3.4. Experiment III

On Day 1 of Experiment III, participants played Game 4. This game was like Game 3 but introducing face to face interaction. So, first people earned their money in the same real effort task as in Experiment II. In this case, they earned € 4.9 in this task (instead of € 5) to make it more natural to pay them in coins. After that, they were shown the description of Charity L and they were told that we had an assistant collecting money for this charity next to the outside door of the lab and that they could donate money to the charity if they wanted. Then subjects were paid their € 4.9 and left the lab one by one. They were always paid in the same coin denominations: one € 2 coin, two € 1 coins, one 50 cents coin, one 20 cents coin, and two 10 cents coin. This allowed the

participants to donate any amount they wanted from € 0 to € 4.9 in increments of € 0.1.

We had an assistant (always female) standing next to the outside door of the lab holding a bucket labeled as "donation box". She was instructed to ask each participant that came out of the lab the same question: "Hello, I am collecting donations for charity L, would you like to contribute?" The assistant recorded the amounts donated and was instructed not to ask for any personal or identifying information. Note that this last version of the game (Game 4) is much more similar to our first field situation than the standard dictator game, but everything happens in the lab and people know at all times that they are participating in an experiment.

Also in this experiment, participants filled out the Interpersonal Reactivity Index and the Big-Five questionnaires and we elicited their demographic information. They afterwards participated in Days 2 and 3.

As in Experiment I, we included a control group of people who did not participate in Game 4 on Day 1 of the experiment, which allows us to check for any effect of having played the game on subsequent field behavior. On Day 1, people in the control condition earned € 4 from accomplishing a similar real-effort task and answered the same questionnaires as those in the main group. Note that we included control conditions in Experiments I and III, which incorporate the games that are least and most similar to our first field situation, respectively.

1.2.3.5. Participants

Overall, 440 people participated in our experiments. All of them were recruited from our university subject pool, which consists of about 7,000 people who are mainly current and former students of the university. People could only participate in one of the experiments reported here, and there were no other eligibility or exclusion criteria.

A total of 161 people participated in Experiment I on the first day, 112 in the main group and 49 in the control condition. Due to attrition, we lost some participants on the

second day, so that we ended up with a sample of 148 people who participated in the whole experiment, 102 in the main group and 46 in the control condition. A total of 110 people participated in the first session of Experiment II, and 101 of them participated in both sessions. In Experiment III, we had a total of 169 subjects in the first session, 116 in the main group and 53 in the control condition. Due to attrition, we ended up with 98 participants in the main group and 45 in the control condition.

1.3. Results

We will now present the results, organized in five separate subsections. In subsections 1.3.1 and 1.3.2, we describe the patterns of decisions observed in the two field situations. Then, in subsection 1.3.3, we show the decisions made in the four lab games. In subsection 1.3.4, we present our main analysis of the predictive power of the different lab games in relation to the field behaviors. In subsection 1.3.5, we analyze the correlation between the two different field situations. Finally, in subsection 1.3.6, we explore the explanatory power of the additional psychological measures we elicited in relation to the lab and the field behaviors.

1.3.1. Donations in the field

Figure 1a shows the distributions of participant donations in the main groups (excluding the control conditions) in our first field situation for the three experiments.³ Figure 1b presents the distributions of donations in the control groups of Experiments I and II, separately and together.

As Figure 1a shows, donation behavior in the first field situation was remarkably similar across the main groups of all three experiments. This indicates that the specific games played in the lab on Day 1 (which is the main difference between experiments), did not substantially affect decisions in the first field situation. This conclusion is further reinforced by Figure 1b, which shows that field donations in the control

³ The first field situation has some missing observations mainly due to incidental factors that made it impossible for our assistant to approach some particular participants; we also excluded four observations of zero because the participant told the solicitor that they did not want to donate because they just donated to another similar charity (charity L) recently.

conditions were also very similar to donations in the main groups (Mann-Whitney test comparing main group and control in Experiment I: $z = 1944$, $p = 0.69$; in Experiment III: $z = 1969$, $p = 0.86$).

Figure 1a. Distributions of donations in the main groups (excluding control conditions)

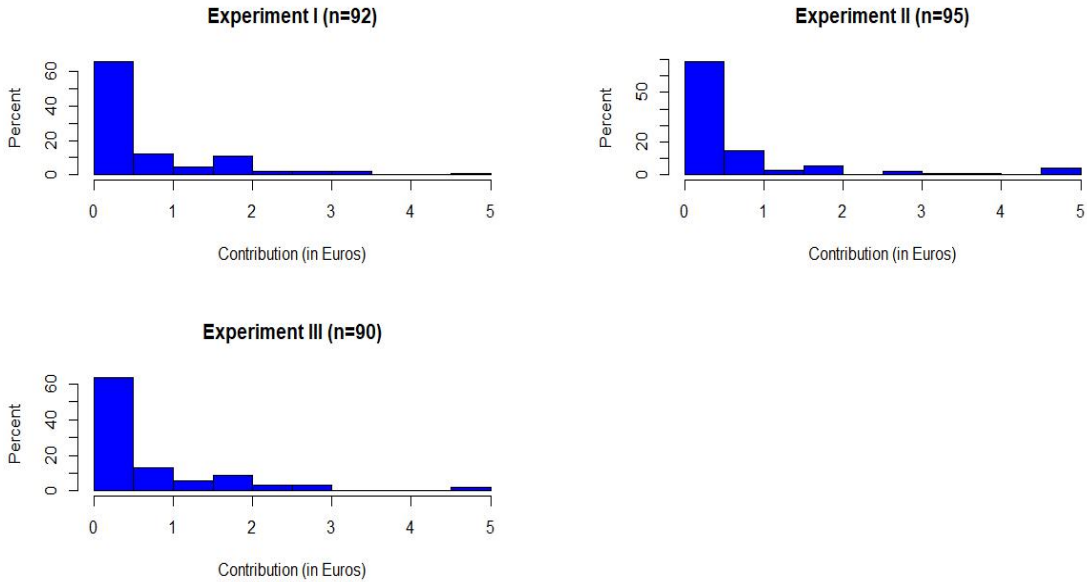
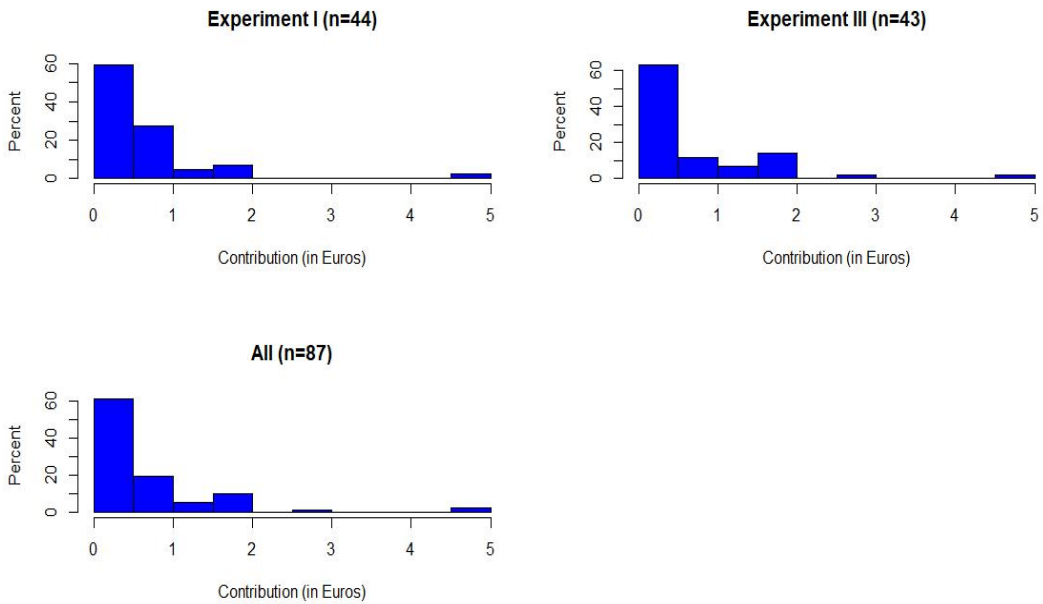


Figure 1b. Distributions of donations in the control conditions



In terms of the shape of the distributions, they are all markedly right skewed, with high spikes of at zero (the percentage of zero in the six distributions are 55%, 59%, 50%, 45%, 53% and 51% respectively; and mean of the distributions are 0.65, 0.69, 0.71, 0.59, 0.71 and 0.65 respectively). This type of distribution broadly matches previous donation data in similar settings (see, e.g., Galizzi and Navarro-Martinez, 2019). All the distributions have maximum values of 5.

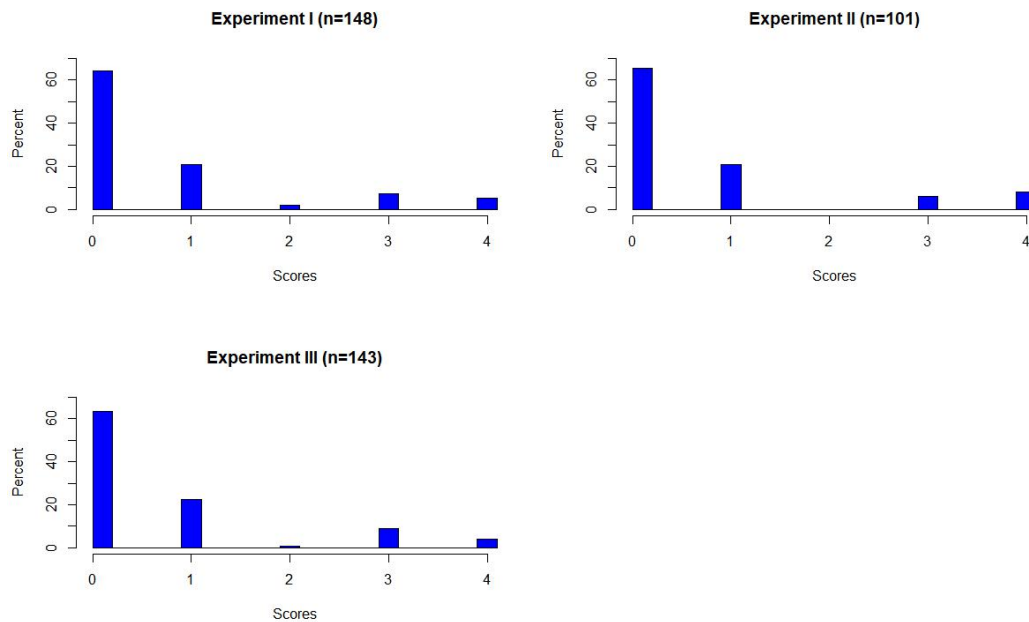
1.3.2. Interests in volunteering

Figure 2 shows the distributions of interest in volunteering scores in the second field situation for the three experiments.

We used individual links to trace the responses of each participant. We coded the responses as numeric values ranging from 0 to 4. If a participant didn't click the link, a value of "0" was recorded; a value of "1" was recorded if a participant clicked the link, but didn't check the information about any volunteering activity; and a value of "2" was recorded if a participant checked one volunteering activity but showed no interest; and a value of "3" was recorded if a participant checked one volunteering activity and also showed interest; and a value of "4" was recorded if a participant was interested in more than one volunteering activities or if he or she clicked one of the links leading to the website of the charities. Just in case that some participants didn't notice the email, we sent another identical email about one month after the first email. The responses were coded in the same way. The largest score a participant got in these two rounds was taken as his or her final score in this field situation.

The distributions of interest in volunteering are very similar across experiments. They are also very positively skewed, with a bit over 60% of the participants showing no interest. About 20% of the participants clicked the link embedded in the email but left the page without checking any volunteering opportunities; about 9% checked the information on one volunteering opportunity and indicated their interest in volunteering for that program; and about 6% were very interested: they checked two or all of the volunteering opportunities and some of them visited the websites of the charities to learn more or sign up.

Figure 2. Distributions of interest in volunteering



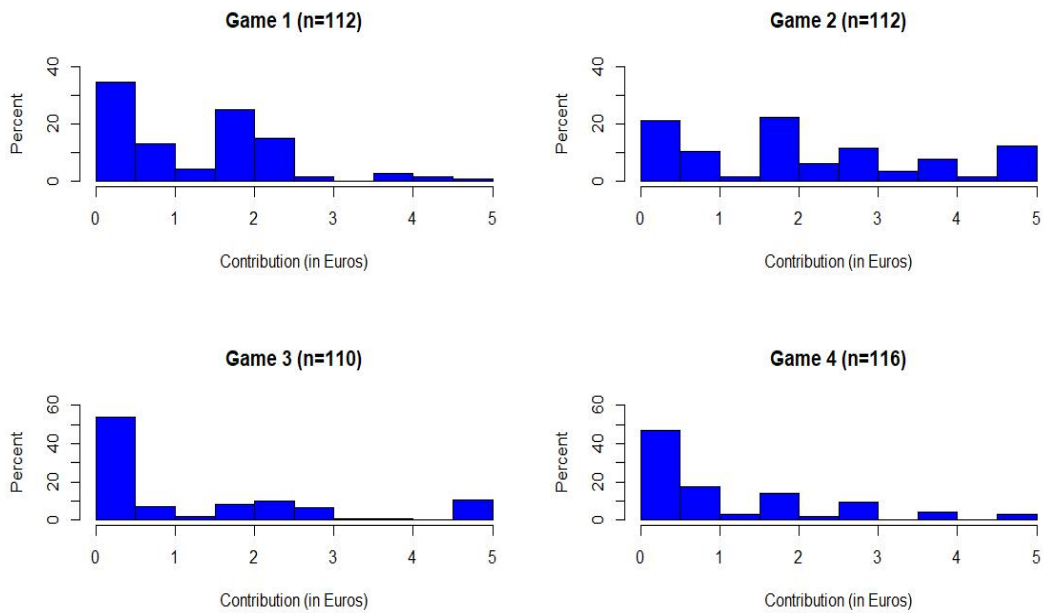
1.3.3. Decisions in the lab games

Figure 3 presents the distributions of decisions in our four lab games. The distribution for the standard dictator game (Game 1) is broadly in line with the results of the literature: positively skewed with a relatively high spike at zero (25%) and another smaller one at about half of the endowment (15%) (see Camerer, 2003). The distribution for Game 2 is notably flatter than the one for Game 1, with less people at zero (15%) and more at the higher end of the distribution, including even 12% of the people at 5 (the full endowment). This means that participants gave on average substantially more when the recipient was a charity (the mean in Game 1 is 1.38, and the mean in Game 2 is 2.24)⁴, which is also consistent with previous literature (see, e.g., Eckel and Grossman, 1996).⁵

⁴ Mann-Whitney comparing the two distributions: $z = 4338$, $p = 0.00$; t-test comparing the means: $t = -4.57$, $p = 0.00$.

⁵ Games 1 and 2 were played in the same session in a randomized order, but the results were not significantly affected by the order (Mann-Whitney comparing Game 1 when it was first and second: $z = 1730$, $p = 0.32$; for Game 2: $z = 1810$, $p = 0.14$).

Figure 3. Distributions of decisions in the four lab games



In Games 3 and 4, the distributions are more positively skewed, with higher spikes at zero (48% in Game 3 and 41% in Game 4). The mean of contribution in these two games are 1.34 and 1.21 respectively. A clear difference between these two games and the dictator game is that they do not show a higher spike around half of the endowment. It is also clear that the distributions in these games look much more similar to the distributions of donations obtained in the first field situation.

1.3.4. Do the lab games predict the field behaviors?

We now turn to our key objective of analyzing the extent to which the different lab games can predict the field behaviors. This will tell us about the predictive power of the standard dictator game and about the improvement when the additional contextual elements from the field are incorporated into the lab games. In the next two subsections, we will analyze our two field behaviors in turn.

1.3.4.1. Donations in the field

To begin with, Table 3 reports the Pearson correlations between the different game decisions and the donations in the field, including 95% confidence intervals. As we can

see in the table, the standard dictator game (Game 1) shows a small to moderate correlation of 0.2 with the field donations, which is only significant at the 10% level in our sample. In the other games, adding contextual elements to make them more similar to our first field situation dramatically increases the correlation, reaching a level of 0.65 in Game 4. By any standards, this is a remarkably high correlation.⁶

Table 3. Correlations between game decisions and field donations (Pearson)

	Point Estimation	Confidence Interval, 95%
Game 1	0.20*	[0,0.39]
Game 2	0.17*	[-0.03,0.36]
Game 3	0.58***	[0.43,0.70]
Game 4	0.65***	[0.51,0.75]

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

Game 2, however, is an exception to this increasing trend. The correlation for Game 2 is slightly lower than, and not significantly different from, the one for Game 1 (see the confidence levels in Table 3, both correlation coefficients fall into the 95% confidence level of the other game). As reflected in Figure 3, the distribution of contributions in Game 2 is ostensibly the least similar to the field distributions (Figures 1a and 1b). This is produced by the fact that, when we introduced Charity L as the recipient (instead of another participant), fewer people gave zero and more people gave higher amounts, resulting in a flatter distribution and higher average contributions. A notable fact is that those participants that didn’t donate any money in the field on average donated 1.98 euros in Game 2. This pattern of donating more to charity in the lab than in the field is consistent with previous literature (see, e.g., Benz and Meier, 2008). Then, in Game 3, introducing the element of earned money, pulled the distribution back towards zero and away from the higher contributions.

⁶ We report here the simplest correlation analysis, but we also performed an analysis using non-parametric Kendall correlations (see Table A1 in the Appendix). The pattern obtained with Kendall correlations is broadly in line with the correlations discussed here.

In Table 4, we present the results of a regression analysis to take a closer look into the proportions of variance in the field behavior explained by the different games. We conducted four regressions using always the field donations as the dependent variable and regressing them on each of the four different games. The pattern of coefficients obtained broadly replicates the one obtained in the correlation analysis. In terms of the proportions of variance explained (R2), we can see that Games 1 and 2 explain a very small proportion (less than 0.05), while Games 3 and 4 explain a much larger proportion, reaching 0.43 in Game 4.⁷

Table 4. Regression analysis (OLS): Donations in the field

	Coefficient (SE in brackets)	Variance Explained (R2)	N
Game 1	0.18* (0.09)	0.04	92
Game 2	0.11* (0.07)	0.03	92
Game 3	0.44*** (0.06)	0.34	95
Game 4	0.55*** (0.07)	0.43	90

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

Overall, our analyses show that the limited predictive power of the standard dictator when it comes to field behavior can be very substantially increased by introducing in the lab contextual elements from the field. In Game 4, we actually reached a level that made the game a very good predictor of the field behavior we wanted to address. In our set-up, it seems that the earned money element is the factor that made most of the work of increasing the predictive power (from Game 2 to Game 3).

⁷ Also in this regression analysis we report the simplest results here. Table A2 in the Appendix shows the results of using a Tobit regression approach to take into account the censored nature of the data. Again, the patterns obtained in the Tobit regressions are broadly in line with the ones reported in Table 4.

1.3.4.2. Interest in the volunteering

Table 5 reports the Pearson correlations between the game decisions and participants' interest in volunteering scores, including 95% confidence intervals. The standard dictator game (Game 1), shows again a relatively small correlation (0.15) with this second field situation, which is not significant in our sample at any of the usual levels. Game 2 presents a slightly lower, but very similar correlation (0.13). Interestingly, in this case, Games 3 and 4 show an even smaller correlation with interest in volunteering. Both correlations (-0.01 and 0.03, respectively) are very close to zero. None of these correlations are significant in our sample.

Table 5. Correlations between the game decisions with the scores (Pearson)

	Point Estimation	Confidence Interval, 95%
Game 1	0.15	[-0.05,0.33]
Game 2	0.13	[-0.06,0.32]
Game 3	-0.01	[-0.21,0.19]
Game 4	0.03	[-0.17,0.23]

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

In Table 6, we present a regression analysis that is analogous to the one performed in Table 4 for our first field situation. So, we run four separate regressions, using always interest in volunteering scores as our dependent variable and regressing it on the four different games. The pattern obtained in the regression coefficients is in line with the one obtained in the correlations, and the R2s close to zero (0.02, 0.02, 0.00, and 0.00) indicate that none of the games explains any substantial variance in the field behavior.⁸

⁸ As with the first field situation, in this section we reported the simplest analyses, but the appendix includes Kendall correlations (Table A3) and a regression analysis using negative binomial regressions (Table A4) to take into account the count nature of our interest in volunteering data. Both analyses produce results that replicate the patterns reported in Tables 5 and 6.

Table 6. Regression analysis (OLS): Interest in volunteering

	Coefficient (SE in brackets)	Variance Explained (R ²)	N
Game 1	0.14 (0.09)	0.02	102
Game 2	0.09 (0.07)	0.02	102
Game 3	-0.01 (0.07)	0.00	101
Game 4	0.03 (0.08)	0.00	98

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

The results reported in this section show that, when contextual elements from a particular field situation are incorporated into lab games to increase their predictive power in relation to that situation, this does not increase predictive power in relation to field settings that have different characteristics. This is again consistent with the idea that behavior is highly context-dependent.

1.3.5. Relationship between the two field behaviors

In this subsection, we look into the relationship between the two field behaviors to analyze the extent to which behavior in one field situation is generalizable to another situation with different characteristics.

For this, we can use our whole sample, i.e., participants in all three experiments, including also the control conditions. After excluding the missing values, we have in total 364 participants with complete observations. The Pearson correlation between the two behaviors is 0.01 (Kendall correlation: -0.01). So, there is essentially no correlation between these field behaviors.

We also run a regression analysis to look further into the proportion of variance in one situation explained by behavior in the other. In particular, we regressed the interest in volunteering scores on the donations in the field. The R^2 in an OLS regression is 0.00; McFadden's R^2 corresponding to a negative binomial regression is also 0.00. So, no proportion of variance in our second field situation is explained by behaviors in the first one.

In line with the findings reported in the previous subsection, these results show that behavior in one particular situation (in the field in this case) does not translate into behavior in another setting with different characteristics. This highlights again the power of context in determining behavior.

1.3.6. Predictive power of the Big-Five and empathy in the field

In this subsection, we compare the predictive power of our psychological constructs (the Big-Five and empathy) in the field with that of our games.⁹

Now we explore the relationship between the nine psychological constructs and the two pro-social behaviors in the field. Table 7 summarizes the pairwise Pearson correlation between the psychological constructs and the field behaviors.

As the table shows, several measures of the Big-Five and empathy are significantly correlated with the two field behaviors. Agreeableness, Conscientiousness, Neuroticism, and Empathic Concern have some moderate correlation with donation made in the field.

⁹ To see the connection between the psychological constructs and the game decisions, please refer to Table A5 in the appendix.

Extraversion and Conscientiousness correlate moderately with the interest in volunteering.

Table 7. Correlations between the psychological constructs and the field behaviors
(Pearson)

	Donation in the field	Interest in volunteering
Agreeableness	0.14***	0.01
Extraversion	0.03	-0.15***
Conscientiousness	-0.14***	0.09*
Neuroticism	0.10*	-0.03
Openness	-0.05	-0.03
Perspective Taking	0.03	-0.06
Empathic Concern	0.11**	0.06
Fantasy	-0.07	-0.05
Personal Distress	0.02	0.01
N	364	392

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

We run three different regression models to further explore the connection between donation in the field and the nine psychological constructs. In model 1, we only used the Big-Five personality traits as predictors; in model 2, we used the four aspects of empathy; in model three, we integrated all the nine constructs. Table 8 summarizes the regression results.

Table 8. Regression analysis (OLS): Field donation and psychological constructs

	Model 1	Model 2	Model 3
Agreeableness	0.03		0.03*
Extraversion	-0.00		-0.01
Conscientiousness	-0.03*		-0.03*
Neuroticism	0.04**		0.04**
Openness	-0.02		-0.02
Perspective Taking		-0.00	-0.01
Empathic Concern		0.04***	0.03**
Fantasy		-0.03**	-0.03**
Personal Distress		-0.00	0.01
R2	0.04	0.03	0.07
Adjusted R2	0.03	0.02	0.05
N	364	364	364

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

The regression estimates in Table 8 are broadly in line with the correlation coefficients in Table 7. Those with a higher score on Agreeableness, Neuroticism, and Empathic Concern donated more to the charity in the field, while those with a higher score on Conscientiousness and Fantasy donated less. Comparing the regression results in Table 8 with those in Table 4, we can see that the predictive power of the first two lab games is quite similar to that of these psychological constructs, but the predictive power of our last two games is significantly superior to that of those psychological constructs. Our findings suggest that, in relation to predicting behavior in a specific situation,

general personality traits have some moderate predictive power, but the predictive power of behavior in similar situations is much larger.

We also run different regression models to further check the relationship between the nine psychological constructs and interest in volunteering opportunities. Table 9 summarizes the regression results.

Table 9. Regression analysis (OLS): Volunteering and psychological constructs

	Model 1	Model 2	Model 3
Agreeableness	0.04**		0.03
Extraversion	-0.07***		-0.08***
Conscientiousness	0.03*		0.03
Neuroticism	-0.02		-0.02
Openness	-0.01		-0.00
Perspective Taking		-0.03*	-0.02
Empathic Concern		0.04**	0.04**
Fantasy		-0.01	-0.01
Personal Distress		-0.01	-0.01
R2	0.04	0.02	0.05
Adjusted R2	0.02	0.01	0.03
N	392	392	392

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

The regression estimates in Table 9 are consistent with the corresponding correlation coefficients in Table 7. Those with a higher score on Agreeableness, Conscientiousness, and Empathic Concern showed more interest in the volunteering opportunities, while those with a higher score on Extroversion showed robustly less interest¹⁰. Compared to the predictive power of our games showed in Table 6, the predictive power of the psychological constructs is a bit larger.

1.4. Conclusions

Like Levitt and List (2007) noted, the most fundamental question in experimental economics perhaps is whether laboratory findings could provide reliable inferences for the real world. However, more and more studies show that social preference games seem to have the issue of external validity. For example, in a meta-analysis, Galizzi and Navarro-Martinez (2019) concluded that the average lab-field correlation obtained in the literature was 0.14.

To better interpret the lab-field correlations that found in the literature, one question remains that what size of lab-field correlation can be regarded as a satisfying one. Decades of research in psychology and in experimental economics showed that it is very hard for the correlation between behaviors in different situations to break the barrier of 0.3 (e.g., Mischel, 1968; Ross and Nisbett, 1991), and that meanwhile the consistency of behaviors in exactly the same situation over time could be very high. For example, Benz and Meier (2008) found that the correlation between donations from their student participants in different semesters was around 0.8.

Our results in this paper confirm again the low external validity of context-free social preference game. They further suggest that the low lab-field correlation actually comes from some systematic differences between the lab games and the real world in social norm, stakes, scrutiny and other aspects. Furthermore, by identifying and eliminating these differences, not just could we bridge the lab-field gap for distribution of decisions,

¹⁰ Here we found that a significant negative correlation between Extroversion and the interest in volunteering. Please notice that extroverts may like to interact with others more in volunteering, but checking detailed information about volunteering opportunities is very different from volunteering the activity itself.

we could also significantly increase the predictive power of laboratory decisions in relation to behavior in the field: in our case, the lab-field correlation (Pearson's ρ) increased from the initial 0.2 to 0.65.

We are not the first to suspect the role of critical contextual elements in enhancing the external validity of social preference games. Previous research have highlighted the context-dependent nature of human pro-social behavior. For example, among many others, Eckel and Grossman (1996), List (2007), and Andreoni et al. (2017) have explored the impact of social norm, earned money, and face-to-face interaction on contribution of money in dictator game alike situations. Some research, like List (2006, 2007), even have shown that eliminating some important lab-field differences could make the distribution of behavior in experiments closer to the reality and pin down the true motivations behind some observed pro-social behaviors. Furthermore, Levitt and List (2007) have warned that some systematic lab-field differences could result in low external validity of laboratory experiments studying pro-social behaviors. We are just surprised that very few systematic investigation of the external validity of context-free games have been conducted and that even fewer work has been done to test whether eliminating critical lab-field differences could enhance the external validity of lab studies.

Alekseev et al. (2017) presents a literature review of experimental studies that used context-rich instructions instead of abstract ones. They suggested that context-rich language could increase the understanding of laboratory tasks, and that when it changes behavior of the participants, the change is usually beneficial for the research purpose (for example, in the case that certain emotions or social norms need to be evoked). Our results suggest that experimental economics should go one step further: not only using context-rich instructions is often very necessary but also is incorporating critical contextual elements that matter for the field behavior to be addressed.

In this paper, we focused on dictator games and donations. But we believe that the insights we revealed are not only limited to these specific cases. Our methodology could be applied in a broader range of experimental studies of social behavior. Some potential interesting research questions could be bridging the gap between cooperating in public

goods games and cooperating in real social dilemmas,¹¹ the gap between reciprocating in gift-exchange games and reciprocating in the workplace,¹² the gap between cheating in lab games and cheating in various situations in the real life, etc.

Decades of research in psychology and experimental economics have accumulated tons of evidences suggesting that when critical elements of the situation change, behavior will also change, usually in predictable ways. All of these evidences suggest that experimental method is a useful and powerful tool for studying behaviors. Meanwhile, these evidences also suggest that laboratory games should incorporate more contextual elements when behavior in some specific field domains are to be addressed.

¹¹ Stoop et al. (2012) tried to bridge the gap between cooperating in public goods game and cooperating in a real social dilemma. But their methodology is different from ours: they focused on distributions of behaviors at group level and on different aspects of lab-field differences such as subject pool, laboratory setting and contextualization.

¹² For example, Gneezy and List (2006) found that the correlation between the offered wage and workers' productivity in their two field experiments is much lower than the typical ones in gift-exchange games.

Appendix

Table A1. Correlations Between the Game Decisions and Donations in the Field
(Kendall)

Game 1	0.23***
Game 2	0.15*
Game 3	0.35***
Game 4	0.46***

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

Table A2. Donation in the Field: Regression Analysis (Tobit)

	Coefficient (SE in brackets)	Variance Explained (R2)	N
Game 1	0.47** (0.19)	0.04	92
Game 2	0.24* (0.14)	0.03	92
Game 3	0.70*** (0.14)	0.34	95
Game 4	0.78*** (0.13)	0.40	90

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

Table A3. Correlations Between the Game Decisions and Interest in Volunteering
(Kendall)

Game 1	0.13
Game 2	0.10
Game 3	0.02
Game 4	0.02

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

Table A4. Interests in Volunteering: Regression Analysis (Negative Binomial)

	Coefficient (SE in brackets)	McFadden Pseudo R^2	N
Game 1	0.21 (0.14)	0.01	102
Game 2	0.14 (0.11)	0.01	102
Game 3	-0.01 (0.10)	0.00	101
Game 4	0.04 (0.14)	0.00	98

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

Table A5. Correlations between the psychological constructs and the game decisions (Pearson)

	Game 1	Game 2	Game 3	Game 4
Agreeableness	0.17*	0.03	0.05	0.07
Extraversion	0.02	0.12	-0.02	0.04
Conscientiousness	-0.06	0.08	-0.09	0.02
Neuroticism	0.14	0.22**	0.05	0.11
Openness	0.06	-0.02	0.08	-0.03
Perspective Taking	0.14	0.23**	0.20**	-0.01
Empathic Concern	0.05	0.29***	0.14	0.00
Fantasy	0.15	0.16*	0.26***	-0.04

Personal Distress	-0.15	0.03	-0.07	0.04
N	112	112	110	98

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

2. COMPARING PRO-SOCIAL BEHAVIOR IN THE LAB AND THE FIELD OVER TIME

Joint with Daniel Navarro Martinez

“To know the capability of a horse, ride it for a long journey; to know the character of a person, interact with him for a long time.”

----- A Chinese Proverb

2.1. Introduction

Research on “social preferences” has been developing dramatically in the last decades and has been one of the main research topics in behavioral science. The benchmark research method in this line of research is studying behavior in laboratory games designed to capture different aspects of social behavior, such as altruism, cooperation, inequity aversion, reciprocity, trust, etc. Some of the most often studied games include the dictator game, the ultimatum game, the trust game, and the public good game, among others (see Camerer, 2003, for a review).

However, an increasing number of papers are questioning the predictive power of this games in relation to pro-social behavior outside of the lab (i.e., their external validity). For example, in a recent meta-analysis conducted by Galizzi and Navarro-Martinez (2019), the authors found that only 39.7% of the lab-field correlations reported in the papers they analyzed were statistically significant, and the average lab-field correlation obtained was 0.14.

Why would social preference games have problems of external validity? One possible reason is that human social behavior depends a lot on context but social preference games are designed to be context-free. The games thus fail to capture some important contextual elements that determine behavior in the field and cannot correlate well with it. For example, Levitt and List (2007) discussed some systematic differences between social preference games and the real world that could undermine their external validity.

In addition to Levitt and List's theoretical discussion, there is a substantial amount of evidence showing that social behavior depends on contextual elements (e.g., Eckel and Grossman, 1996; Konow, 2001; Dana et al., 2006; List, 2007; DellaVigna et al. 2012; Andreoni et al., 2017). In a recent paper analysing the external validity of the dictator game, Wang and Navarro-Martinez (2020) further showed that the correlation between laboratory decisions and donations in the field increased dramatically when crucial contextual differences between the lab game and the field situation were eliminated.

Another possibility is that although these games have low predictive power in relation to one-shot field behaviors, they have better predictive power for aggregated social behavior over extended time periods. This could be the case if social preference games capture some fundamental aspect of social preferences that can explain behavior across different settings (in combination to the particular contextual elements). A closely related finding in personality psychology is that personality traits do not predict well single behaviors, but predict behavior much better over longer periods of time (see, e.g., Fleeson, 2001, 2004).

There is very little research addressing this idea in behavioral and experimental economics, and the few existing papers provide mixed evidence. Benz and Meier (2008) had their participants play some donation games in the lab; they also tracked their participants naturally occurring donations in the four semesters before the experiment and in the four semesters after the experiment. They found that donations in the lab have a much stronger predictive power in relation to average donations over the four semesters than in relation to donations in one particular semester. Gurven and Winking (2008) examined the relationship between Amazonian forager-horticulturalists' decisions in a variety of social preference games and some daily social behaviors in some discrete time episodes that embedded in several months. These behaviors included food-sharing, social visitation, beer production and consumption, labor participation, and contributions to a feast. However, the authors did not find any robust correlations between the game decisions and the frequency of engaging in these behaviors.¹³

¹³ Gurven and Winking (2008) did not aggregate their field behaviors, a difference between their research method and ours.

In this paper, we systematically investigate whether social preference games can predict pro-social behavior in daily life over an extended period of time. To this end, we compare social preferences elicited in several lab games (dictator game, dictator game with taking option, and public good game) with naturally occurring pro-social behaviors elicited daily over a two-week (Study 1) or a one-week (Study 2) period. In Study 1, we elicited the daily behaviors using a modified version of the day reconstruction method (Kahneman et al., 2004), which is well-established in the area of subjective well-being. In Study 2, we used an experience sampling approach (Andrews and Withey, 1976; Larson and Csikszentmihalyi, 2014), adapted to focus on pro-social behaviors. To the best of our knowledge, this is the first paper to use this type of methods to compare economic games with field behaviors. In addition, participants in both studies responded to the Self-Report Altruism (SRA) Scale (Rushton et al., 1981), in which they were asked about the frequency of various pro-social behaviors in the past. We also included personality and empathy measures used in psychology, in order to compare their predictive power with that of the games.

Both of our studies consistently show that the context-free games achieve a significant (albeit moderate) predictive power when the daily pro-social behaviors are aggregated over time. This finding is further supported by the significant correlation between the game decisions and SRA scores. Moreover, our results show that the predictive power of the games is larger than that of the psychological measures of personality and empathy.

In our second study, we have an additional layer of analysis in the form of two naturalistic field situations that the same participants faced (some time before the elicitation of daily the behaviors) and in which they could behave pro-socially. These two situations were being asked by a solicitor to make a donation to a charity and having the chance to check information about volunteering opportunities. We found that our two dictator games significantly (but moderately) predict donation in the field, with which they bare some similarities, but none of our three games predicts to any extent interest in volunteering.

Overall, our findings show that social preference games have low predictive power in relation to one-shot pro-social behaviors that take place in field settings characterized by

contextual elements that do not resemble the games. However, this predictive power can be substantially increased by aggregating pro-social behaviors over time or by matching games to particular real-world contexts (see also Wang and Navarro-Martinez, 2020). This suggests that context-free games might capture some fundamental aspects of social behavior that are relevant across situations in combination with specific contextual elements. These results may be seen as reinforcing the usefulness of context-free games, and they open new avenues for social preference research that aggregates behaviors over time.

The rest of the paper is organized as follows: Section 2.2 presents our Study 1, Section 2.3 explains Study 2, and Section 2.4 discusses our findings and concludes.

2.2. Study 1

2.2.1. Methods

Participants in our study were presented with the following five elements: (1) a set of social preference games; (2) daily reports of pro-social behaviors during the 15 days of our study; (3) a questionnaire that inquired about their altruistic behaviors in the past; (4) measures of the big-five personality traits and empathy; (5) demographic questions.

These elements were organized in the following way. On the first day of the study, participants came to our laboratory to play the social preference games and fill in some questionnaires (including elements 3, 4 and 5 listed above). In this introductory session, the structure of the whole study was explained and all the questions raised were answered. Then, in the following two weeks, participants completed a daily report of pro-social behavior every night. During the two weeks of the study, participants also played the games another six times and reported their big-five personality and empathy six times as well. We will now explain of the five different elements in turn.

2.2.1.1. Social preference games

In this study, we focused on three games: the dictator game, dictator game with taking option, and the public goods game. To facilitate the implementation of these games, they were split into the following five specific games.¹⁴

(1) *Dictator Game 1 (DG1)*: In this game, two players were paired up. Player 1 decided how to divide 5 euros between the self and Player 2. Player 2 simply received the amount indicated by Player 1.

(2) *Dictator Game 2 (DG2)*: Like Dictator Game 1, but the roles were switched.

(3) *Dictator game with taking option 1 (DGT1)*: This game is very similar to the dictator game. The only difference is that Player 1 in this game also has the option to take some money from Player 2. The biggest amount allowed to take is 2 euros.

(4) *Dictator game with taking option 2 (DGT2)*: Like dictator game with taking option 1, but the roles were switched.

(5) *Public Goods Game (PGG)*: In this game, three players were grouped together. Each player had an endowment of 5 euros and had to decide simultaneously how much of it to contribute to a common group fund. The overall money in the group fund was then multiplied by two and split evenly among the three players.

Participants in this study first played these games in the introductory session, and then played the same games another six times in the following two weeks. Each time the order of these games was (partially) randomized: DG1 and DG2 always appeared together, DGT1 and DGT2 also went together, but the order of the DG and DGT blocks was randomized. For each participant, the order of roles in these two games was also randomized. No feedback about the payoffs in the games was provided to the participants during the study. Then, after the whole study had finished, one of the games

¹⁴ All the games were implemented using Qualtrics.

that participants had played was randomly selected, his/her payoff in which was paid as extra bonus for attending our study. Players were grouped randomly when the payoff of the selected game was calculated.

2.2.1.2. Pro-social behaviors in daily life

During the 14 days of our study, our participants reported on their pro-social behaviors at the end of each day, by responding to a questionnaire that was sent to them at night. We used a modified version of the day reconstruction method (Kahneman et al., 2004), adapted to focus on pro-social behavior, to help the participants recall all the details of their behaviors. Specifically, the participants were first asked to create a list of all the pro-social behaviors they had done during the day, organized by the time that the behaviors took place. To help the participants understand better which kinds of behaviors to report, we gave the following definitions:

“The word ‘pro-social’ refers to being helpful with somebody or doing something for a good cause (for example, something that benefits society, a particular group or community, the university, etc.). Helping behavior could involve helping anyone, such as family members, your partner, friends, teachers, classmates, strangers, etc. A behavior for a good cause could be anything as long as it benefits society or any particular group or community. Some examples of this kind of behaviors are donating money, volunteering to support a particular group or community, cleaning or picking up trash or fixing things in public places, etc.”

They were asked to write a brief description of each reported behavior. Then, for each behavior, we asked for more detailed information. This information included: what the participant was doing before the behavior took place, who he/she was interacting with before the behavior, the relationship between the participant and the recipient of the behavior, the amount of time and money spent on the behavior, and the participant's subjective rating of size of the favor from 1 (a very small favor) to 5 (a very big favor). This produced, for each participants and each day, a detailed account of pro-social behavior, based on self-reports.

After the study had finished, we hired two assistants to rate the reported behaviors independently as impartial judges. The two judges were instructed to do three things. They were first showed our definition of pro-social behavior (the same definition given to the participants). Then, for each behavior (based on all the information we had about it) each judge decided whether it should be counted as a pro-social behavior or not. If the judge thought it should be counted, then he further put the behavior in one of the eight categories: helping family members, helping boyfriend/girlfriend, helping friends, helping roommates, helping acquaintances (classmates, neighbors, etc.), and behavior for a good cause. Finally, the judge rated the size of the favor on a scale from 1 (a very small favor) to 5 (a very big favor). It is important to note that the subjective ratings of favor size by the participants were not shown to the judges.

2.2.1.3. Self-reported altruistic behavior in the past

In our introductory session, the participants completed the Self-Report Altruism (SRA) scale (Rushton et al. 1981), which inquired about the frequency of a variety of pro-social behaviors in the past. The scale consists of 20 items, mainly helping behaviors towards acquaintances or strangers, and behaviors for a good cause (e.g., making donations or volunteering). Three of the items are monetary. See Appendix A for a complete list of the scale items.

2.2.1.4. Big-Five personality traits and empathy

We also included in our study measures of the Big-Five personality traits and of empathy. These psychological measures provide an interesting benchmark to compare with the predictive power of the social preference games. The Big Five model is arguably the most influential account of human personality in psychology and it has played a central role in studies and discussions of individual differences (e.g., Barrick and Mount, 1991; Paunonen, 2003; Giluk and Postlethwaite, 2015). As indicated in the introduction, personality measures have been found to have a low correlation with specific behaviors, but a much stronger explanatory power once various behaviors are

aggregated. This is precisely what we are testing in this paper in relation to social preference games, so it will be interesting to compare the predictive power of the Big Five with that of the games once daily behaviors are aggregated. The Big Five, however, are not focused on pro-social behavior, so we also used part of the Interpersonal Reactivity Index (IRI) (Davis, 1980), which focuses on empathy and has been related to pro-social behavior (e.g., Dovidio and Penner, 2001; Borman et al., 2001; Penner, 2002).

In the introductory session, the participants answered a 44-item Big-Five questionnaire (John, 1991) and the 14 IRI items that designed to measure empathic concern and perspective taking. The full IRI questionnaire includes the dimensions of empathic concern, perspective taking, fantasy, and personal distress. Previous research on pro-social personality (e.g., Penner et al., 1995; Penner, 2002) conceptualized the dimensions of empathic concern and perspective taking as other-oriented and distinguished them from self-oriented empathy, like the dimension of personal distress. Other-oriented empathy is ostensibly more relevant in relation to social preferences. We used a Spanish version of the Big-Five questionnaire created by Benet-Martínez and John (1998) and a Spanish IRI based on Pérez-Albénez et al. (2003). See Appendix B for the relevant IRI items.

Then, in the following two weeks, the participants reported their big-five personality and empathy six times. The big-fives were measured using a brief 10-item scale (Gosling et al., 2003); empathy was measured by a 9-item scale developed from the IRI items included in the Pro-social Personality Battery (Penner, 2002). These two scales were slightly modified so that they asked about the behaviors of each particular day. See Appendix C and D for the two scales.

2.2.1.5. Participants, demographics and payments

88 participants were recruited from our university subject pool, which includes mostly undergraduate students. Their ages range from 18 to 23, and about 67% of them are female. Five participants dropped out of the study in the middle. In the introductory session, we asked the participants a series of demographic questions about their age,

gender, study year, romantic relationship status, who they lived with and whether they had kids or other family members to take care of at home.

In addition to payoffs resulting from playing the games, participants earned 1.5 euros for each survey they completed. The participants' payments did not depend on the number of behaviors reported in the daily reports. On average, the participants finished 93% of the 20 surveys they received and earned 28.1 euros from attending our study.

2.2.2. Results

2.2.2.1. Game decisions

Figure 1a below shows the distribution of game decisions in the introductory session. In this study, the participants not only played the three games in the introductory session, but also another six times over the two weeks. Figure 1b depicts the distribution of the game decisions averaged across the seven times they played each game.

Figure 1a. Distribution of game decisions in the introductory session

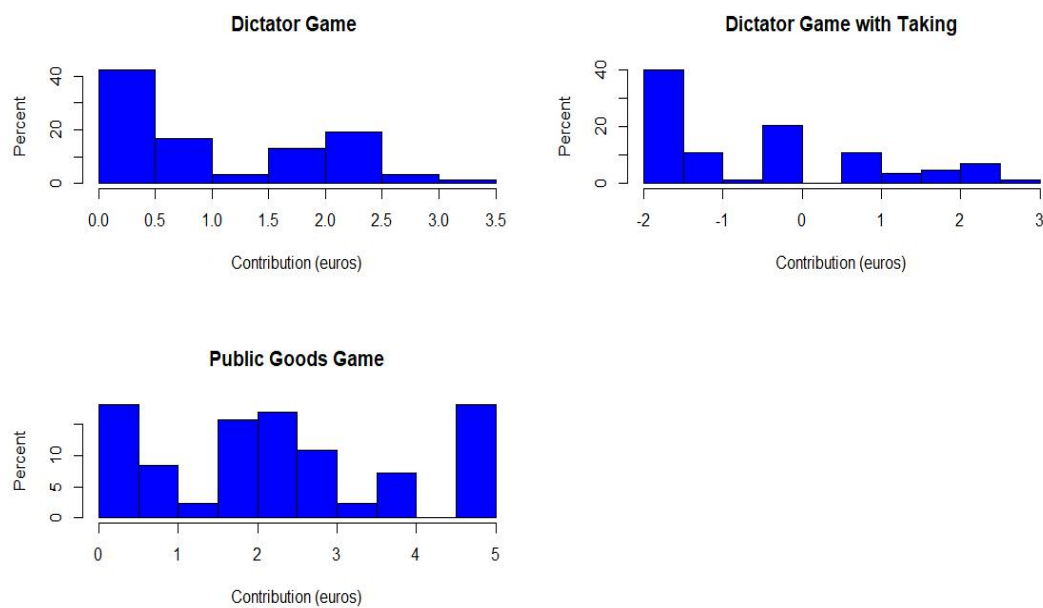
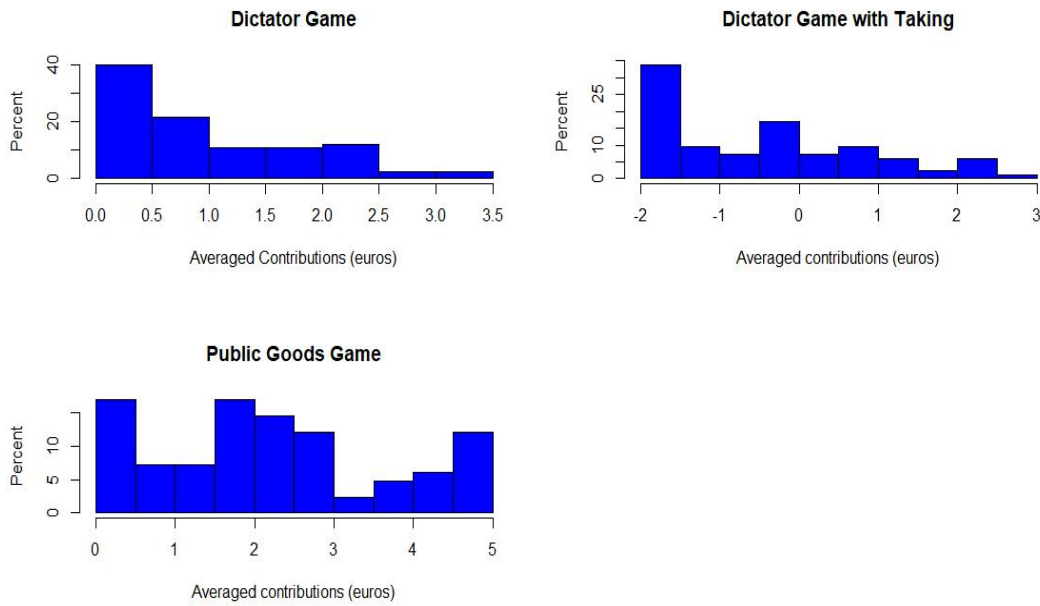


Figure 1b. Distribution of averaged game decisions



The distributions displayed in Figure 1a and 1b are broadly in line with those reported in the literature, for example, with those reported in Galizzi and Navarro-Martinez (2019) and List (2007). As Figure 1a and 1b show, the distribution of these game decisions are quite stable over time. In the dictator game, about 40% of participants did not share any money with their matched partners, and about 20% of participants shared around half of the endowment. When the taking option was added to the dictator game, the players were much less generous. In this setting, many players who did not share money with their partners in the DG started to take money from their partners; and many who shared some money in the DG now shared a smaller amount. This findings are similar to what List (2007) found in this game. The distribution of decisions in the PGG had three peaks: one at 0 (about 15%), another one at half of the endowment (about 25%), and one more at the full endowment (about 12%).

Table 1 presents pairwise Pearson correlations between different game decisions (the one-shot game decisions in the initial session and the averaged game decisions over time). As the table shows, decisions in the three games are quite stable over time, and that there is also a high degree of internal consistency among decisions in different games.

Table 1. Pairwise correlations between game decisions

	DG	DGT	PGG	DG Average	DGT Average	PGG Average
DG	1					
DGT	0.60***	1				
PGG	0.42***	0.44***	1			
DG Average	0.82***	0.72***	0.43***	1		
DGT Average	0.66***	0.83***	0.34***	0.81***	1	
PGG Average	0.34***	0.34***	0.82***	0.43***	0.31***	1

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

Two measures of test-retest reliability further confirm the high stability of game decisions over time. First, for each game, we calculated the Pearson correlation between the game decision in the introductory session and the game decision in each of the later six rounds. These correlations are between 0.54 and 0.78, and all of them are statistically significant at the 1% level. Second, the ratio of within-person variation to between-person variation in the three games is between 43% and 58%. The within-person variation of a participant in a game is captured by the standard deviation of all of his/her decisions in that game; overall within-person variation in a game is calculated as the mean of the standard deviations of all the participants. Between-person variation in a game is simply the standard deviation measuring the individual differences in the averaged game decision over time. The ratios are much less than one, suggesting that the within-person variation in the game is much smaller than the between-person variation.

2.2.2.2. Daily pro-social behaviors

In the 15 days, 1967 pro-social behaviors were reported. After reviewing these behaviors, the two judges agreed on excluding 124 reported behaviors as not pro-social. For the remaining behaviors, we compared the subjective rating of favor size from the participants and the ratings from the two judges. The Pearson correlation between the ratings of the two judges is 0.45; the correlation between the participant ratings and the

two judges ratings are 0.19 and 0.29 respectively. All these correlations are statistically significant at the 1% level.

The two judges assigned the 1843 behaviors to 7 categories. Table 2 shows detailed information on the categorization.

Table 2. Reported pro-social behaviors

	Number of behaviors	Percentage
Helping family members	453	25%
Helping boyfriend/girlfriend	95	5%
Helping friends	510	28%
Helping roommates	82	4%
Helping acquaintances	210	11%
Helping strangers	374	20%
Behaviors for a good cause	119	6%

2.2.2.3. Exploring the correlations between the lab and the field

We now turn to the main research question of this study: whether social preference games predict daily pro-social behaviors over an extended period of time. To analyze this, we explored the relationship between the different game decisions and various measures constructed from the reported daily pro-social behaviors.

We focused our analysis on six game decision variables: the three one-shot game decisions made in the introductory session and the same three game decisions averaged over time (with seven observations for each of them). From the daily behaviors, we constructed five different measures of pro-sociality that assign scores to the participants by categories: frequency of pro-social behaviors, total size of the favors as rated by the judges, total size as rated by the participants, total amount of time spent, and total amount of money spent. In constructing these measures, we controlled for how many daily surveys each participant completed in the two weeks, so that the variables can be interpreted as average "per survey" measures.

To illustrate and clarify how these measures are constructed, let us give a specific example. Suppose participant K responded to 11 daily surveys in the two weeks, in which he reported two helping behaviors towards strangers, behavior x and behavior y . Participant K spent 10 minutes and 1 euro on x , and rated the size of the favor as 4; he spent 20 minutes and 0.5 euro on y , and rated the favor as 3. Assume that after having reviewed these two behaviors, judge A assigned values (ratings of the size favor) of 5 and 2 to x and y respectively; judge B assigned values of 5 and 4. Then participant K's score on frequency in the category of helping strangers is 2 (the number of behaviors) divided by 11 (the number of daily surveys responded). The rating from the judges is the sum of the average ratings divided by the number of daily surveys completed: $[(5+5)/2 + (2+4)/2]/11$. Similarly, the rating by the participant is $(4+3)/11$, the time spent is $(10+20)/11$, and the money spent is $(1+0.5)/11$.

The motivations behind helping behaviors towards family members or friends could be very different from those behind helping strangers or engaging in behaviors for a good cause. Also, whether participants had the chance to help their romantic partners or roommates depended on their relationship status and who they were living with. So, to make our lab-field comparisons more precise and informative, we also organized the seven types of reported behaviors into three big categories. Category A includes helping acquaintances, helping strangers, and behaviors for a good cause; category B includes helping family members and friends; and category C includes helping boyfriend/girlfriend and roommates. Our main focus is on behaviors in category A, because these behaviors are ostensibly the closest to the social preference games and to the type of behaviors discussed in the literature on social preferences; they are also closest to the behaviors listed in the SRA scale.

Figure 2 shows the distribution of the participants' scores for behaviors in category A. As the figure shows, the distributions of the first three variables are concentrated around the center with a slight positive skew. The distribution of participants' rating shows higher scores than the one of judges ratings, which suggests that participants' own pro-sociality ratings were somewhat inflated. The distribution of time spent on pro-social behaviors is more positive skewed, with most behaviors corresponding to short durations. Most reported behaviors involved no money (54%) or small amounts, which results in a distribution that is highly positive skewed.

Figure 2. Distributions of scores for the daily pro-social behaviors

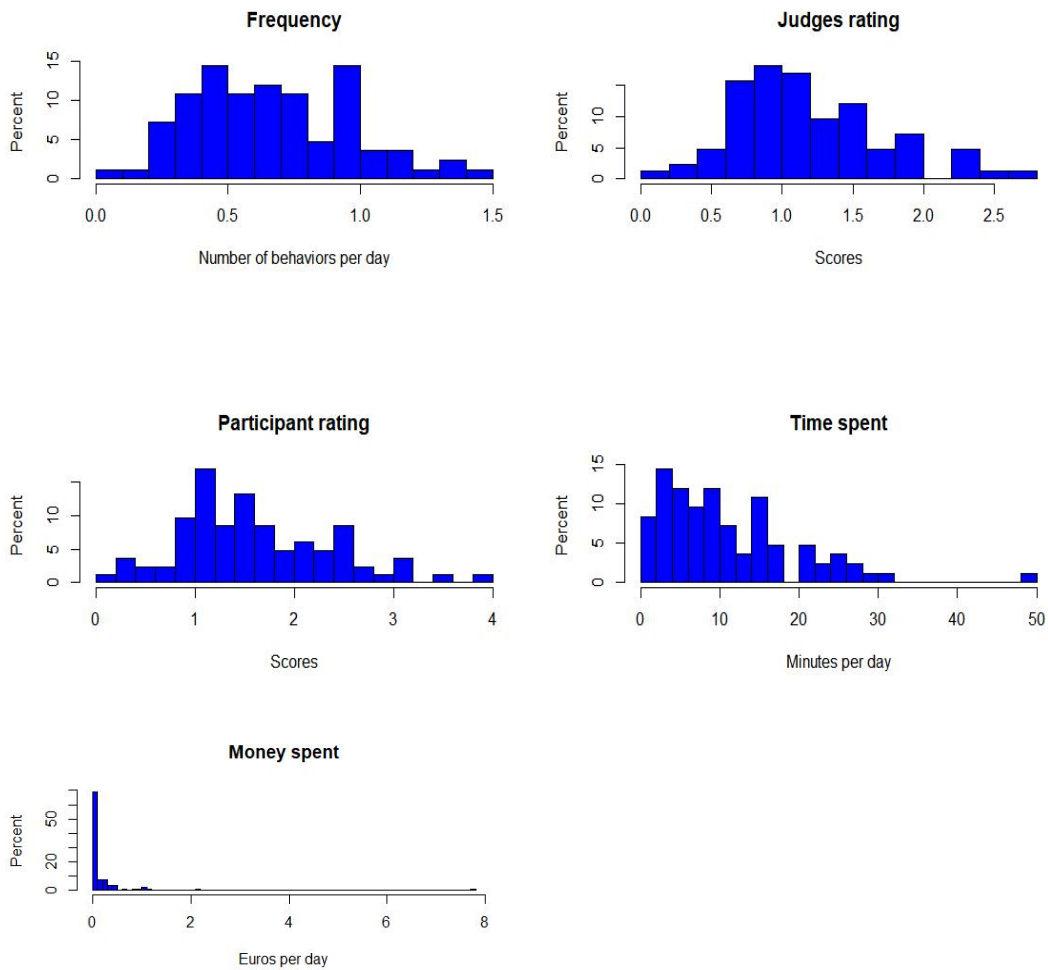


Table 3 below shows pairwise Pearson correlations between the six game variables and the five measures constructed from the behaviors reported over the two weeks in category A.¹⁵ As the table shows, among the game variables, decisions in the dictator game and the dictator game with the taking option significantly correlate with the frequency of pro-social behaviors in category A, and also with the judges ratings of those behaviors. Decisions in the public goods game do not correlate with frequency and judges ratings, but they shows a moderate correlation with the amount of time spent and the amount of money spent on these behaviors.

¹⁵ We also performed an analysis using non-parametric Kendall correlations (see Table E1 in the Appendix). The pattern obtained with Kendall correlations is broadly in line with the correlations reported here.

Table 3. Correlations between game variables and the daily behaviors (category A)

	Frequency	Judges rating	Participant Rating	Time spent	Money spent
DG	0.19*	0.22*	0.14	0.05	0.10
DGT	0.24**	0.22**	0.16	0.10	0.16
PGG	0.02	0.07	0.09	0.19*	0.22**
DG.Average	0.14	0.18	0.07	0.04	0.13
DGT.Average	0.20*	0.22**	0.10	0.12	0.07
PGG.Average	-0.04	0.00	-0.03	0.12	0.21*

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

Table 4 reports the Pearson correlations between the game variables and the participants’ reported behaviors in category B (helping behaviors towards family members or friends). The low correlations in Table 4 confirm our earlier conjecture about pro-social behavior in different categories being driven by different motivations. They also suggest that the observed significant correlations between the game decisions and the daily behaviors in category A was not generated by self-presentation concerns or individual differences in committing to the study.

Table 4. Correlations between game decisions and daily behaviors (category B)

	Frequency	Judges rating	Participant Rating	Time spent	Money spent
DG	-0.01	-0.04	-0.02	-0.09	-0.04
DGT	-0.02	-0.03	-0.05	-0.06	0.01
PGG	0.10	0.08	0.02	0.10	-0.11
DG.Average	-0.09	-0.09	-0.07	-0.10	-0.01
DGT.Average	-0.01	-0.04	0.01	-0.10	0.02
PGG.Average	0.01	0.03	-0.01	0.14	0.1

To explore whether having a longer time period of observation has indeed increased the correlation between the games and the field behaviors, we also analyzed the correlations by week. Table 5 below shows the pairwise Pearson correlations between the game decisions and our two most important measures for the in category A: the frequency of behaviors and the judges ratings.

Table 5. Correlations between game decisions and daily behaviors by week

	Frequency (Week 1)	Judges rating (Week 1)	Frequency (Week 2)	Judges rating (Week 2)
DG	0.10	0.12	0.11	0.12
DGT	0.11	0.10	0.15	0.13
PGG	-0.02	0.02	-0.04	0.01
DG.Average	-0.02	0.03	0.08	0.08
DGT.Average	0.05	0.08	0.10	0.10
PGG.Average	-0.10	-0.02	-0.04	-0.03

Table 5 shows that none of the game variables significantly correlates with the frequency of behaviors or the judges ratings if only one week is considered. This confirms that aggregating field behaviors over longer time periods significantly increases the correlation with social preference games.

2.2.2.4. Correlations between the game decisions and the SRA scores

We now explore the relationship between the game decisions and the self-reported altruistic behaviors in the past. Table 6 below shows the pairwise Pearson correlation between the game decisions and the SRA scores. As the table shows, decisions in the two dictator games have statistically significant correlations with the SRA scores; and decisions in the PGG have a marginally significant correlation with the SRA scores. This finding provides some further support for the earlier results that our social preference games have some predictive power for daily pro-social behaviors in a long time period.

Table 6. Pairwise correlation between game decisions and SRA scores

	SRA	SRA Money
DG	0.22**	0.12
DGT	0.15	0.09
PGG	0.16	0.07
DG.Average	0.30***	0.22**
DGT.Average	0.22**	0.14
PGG.Average	0.17	0.05
Observations	83	83

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

2.2.2.5. The predictive power of the Big-Five and empathy in the field

We will now investigate the predictive power of the big-five and empathy in relation to the daily pro-social behaviors. This will provide a benchmark to compare with the predictive power of the social preference games. We start our analysis with the big-five personality traits (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism), and the IRI measures of perspective taking and empathic concern collected in the introductory session. Table 7 shows pairwise Pearson correlation between these psychological measures and the five variables constructed from the daily behaviors reported in category A.

As Table 7 shows, no correlation is significant at the usual 5% level (and only one is significant at the 10% level). We also conducted a regression analysis including several different specifications to further investigate the predictive power of the psychological measures and to obtain estimates of the proportion of variance they explain. The results confirm that the psychological measures have very little predictive power in relation to the daily field behaviors.¹⁶

¹⁶ Please refer to Table E2 in the appendix for the regression analysis.

Table 7. Correlations between psychological measures and daily behaviors

		Frequency	Judges rating	Participant rating	Time spent	Money spent
Big-Five	Openness	0.07	0.08	0.14	0.11	0.04
	Conscientiousness	-0.07	-0.06	-0.10	-0.02	-0.17
	Extraversion	0.10	0.06	0.12	-0.02	0.08
	Agreeableness	0.12	0.12	0.19*	0.06	0.11
	Neuroticism	0.00	0.00	0.02	0.08	-0.08
Empathy	Perspective taking	0.06	0.05	0.06	-0.01	0.10
	Empathic concern	-0.10	-0.08	-0.01	0.02	0.13

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

During this study, the participants reported their big-five personality and empathic inclinations six times. This allows us to have distributions of these measures. We found that the degree of within-person variation is very substantial, and is no smaller than the between-person variation. This finding is in line with previous findings in personality psychology (e.g., Fleeson 2001). One relevant question then is whether the mean of these distributions can better predict pro-social behaviors in the daily life. Table 8 below shows the pairwise Pearson correlations between the means of these distributions and the five daily behavior variables for the behaviors in category A.

Table 8. Correlations between means of psychological measures and daily behaviors

		Frequency	Judges rating	Participant rating	Time spent	Money spent
Big-Five	Openness	0.00	-0.03	-0.00	-0.23**	0.06
	Conscientiousness	0.01	-0.04	-0.06	-0.13	0.04
	Extraversion	0.01	0.03	-0.01	-0.17	0.13
	Agreeableness	0.13	0.19*	0.12	0.16	0.08
	Neuroticism	0.16	0.18*	0.07	0.10	0.10
Empathy	Perspective taking	0.05	0.09	-0.02	0.09	0.05
	Empathic concern	-0.02	0.00	-0.10	-0.02	0.01

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

Only one correlation is significant at the usual 5% level in Table 8. Specifically, a negative correlation between openness to experience and time spent on pro-social behaviors. Two more correlations appear as marginally significant at the 10% level. We again conducted a regression analysis using several different specifications. The numbers are somewhat higher in this analysis using the mean of the distributions, but again these psychological measures show very limited predictive power in relation to the daily field behaviors.¹⁷ Crucially, the psychological measures used here have a lower predictive power than the social preference games.

2.3. Study 2

Our second study includes similar elements and has a similar structure to our first study, but there are two important differences between them. First, in Study 2, we tracked the participants’ daily pro-social behaviors using an experience-sampling method (Andrews and Withey, 1976; Larson and Csikszentmihalyi, 2014). This method required participants to report pro-social behaviors right after they occurred (instead of constructing a report at the end of the day). Using a different method to track daily pro-social behaviors allowed us to check the robustness of the findings we obtained in Study 1. The experience-sampling methodology has both potential advantages and disadvantages compared to the day reconstruction method. For instance, it relies less on memory and therefore avoids (at least to some extent) bias related to imperfect or distorted recollection of the pro-social episodes. It also seems less prone to the creation of fictitious behaviors, because during the study the participants are not prompted to fill in questionnaires in which they have to report lists of behaviors. On the other hand, reporting behaviors seems more costly with experience sampling, given that people are expected to fill in a questionnaire after every pro-social episode. So, we predicted that we would obtain fewer behaviors in this study. Overall, obtaining the same pattern of results in both studies would substantially add to the robustness of our findings.

¹⁷ Please refer to Table E3 in the appendix for a summary of the regression results.

The second main difference is that for the participants in this study we also have data on their behavior in two naturalistic field situations in which they could behave pro-socially, which they faced (several months before) as part of a previous study. These situations were encountering a solicitor who asked them if they wanted to contribute money to a charity, and receiving an email through which they could show interest in volunteering by checking information. This gives us an important additional element to analyze people's pro-social behavior in the field.

2.3.1. Methods

Similar as in Study 1, the participants were faced with five elements: (1) a set of social preference games; (2) daily pro-social behaviors reported during our study; (3) a questionnaire that inquires about their altruistic behaviors in the past; (4) measures of the big-five personality traits and empathy; (5) demographic information. We included the same social preference games and used the same materials as in Study 1 to collect information about the big-five, empathy, and demographics.

These elements were also organized in a similar way as in Study 1. Participants came to our laboratory on the first day to play some social preference games and to finish some questionnaires. In the introductory session, they were instructed to create an icon on their phones, leading to a form that would be used to report daily pro-social behaviors that occur in the following week. Then in that week, in addition to reporting behaviors, they completed a questionnaire each night, which asked them some simple questions and reminded them to report pro-social behaviors right after they happened. The participants also played the games and reported their big-five personality states and empathy states twice in that week.

As before, we collected detailed information of each reported behavior. After the study has finished, we hired two native Spanish master students to rate the reported behaviors independently as impartial judges. The two judges were instructed to do the same job as in Study 1.

106 students at Pompeu Fabra University participated in this study. 91 of them were faced with the two naturalistic field situations in a previous study. The first situation took place on a large square inside a university campus. A research assistant solicited donations for a well-known charity that devoted to helping children in developing countries. She was wearing a university T-shirt, with a professional charity bucket that had a large sticker with the logo of the charity. She also had a laminated color-printed leaflet that introduces the charity and its projects. The participants walked out our laboratory after finishing an experiment¹⁸ and encountered this solicitor on the square. They were asked explicitly if they wanted to make a contribution. All the donations were recorded and sent to the charity later.

For the second field situation, we sent our participants an email with a link to a website where they could check information about volunteering opportunities at three different charities. If participants clicked the link, they were directed to a website where they could see the names of three real volunteering projects managed by three real charities. Participants were able to click the volunteering project they wanted and go to a page where they could see a brief description of the project. At the bottom of that page, people were asked “Are you interested in volunteering for this program?” They could answer “yes” or “no” to this question. This was followed by a page where participants could click a link to go to the charity’s website if they wanted to learn more about that particular volunteering project or to sign up for it. After having checked the information about one volunteering project, participants were automatically redirected back to the initial page where they could click on the links for the other volunteering projects if they wanted. This set-up allowed us to measure people's interest in volunteering by tracking their information search behavior (i.e., the links they clicked).

Six of the 106 participants dropped out of the study in the middle, among which four of them were faced earlier with the field situations. Plus the payoff in the games, the participants earned a fixed amount of 5 euros and 2 euros for each of the nine survey they finished (seven daily surveys received at night and two surveys that comprised of games and psychological measures). On average, the participants finished 95% of the

¹⁸ All the participants earned a payoff of 15 euros from attending the study and were paid using exactly the same bill and coin denominations.

surveys they received and earned a total payoff of 22.7 euros. The same as in Study 1, the participants' payoff didn't depend on the number of behaviors they reported.

2.3.2. Results

2.3.2.1. Game Decisions

Figure 3a shows distribution of the three game decisions in the introductory session. Figure 3b then describes distribution of decisions in the three games that averaged over the three rounds.

Figure 3a. Distribution of the three game decisions in the introductory session

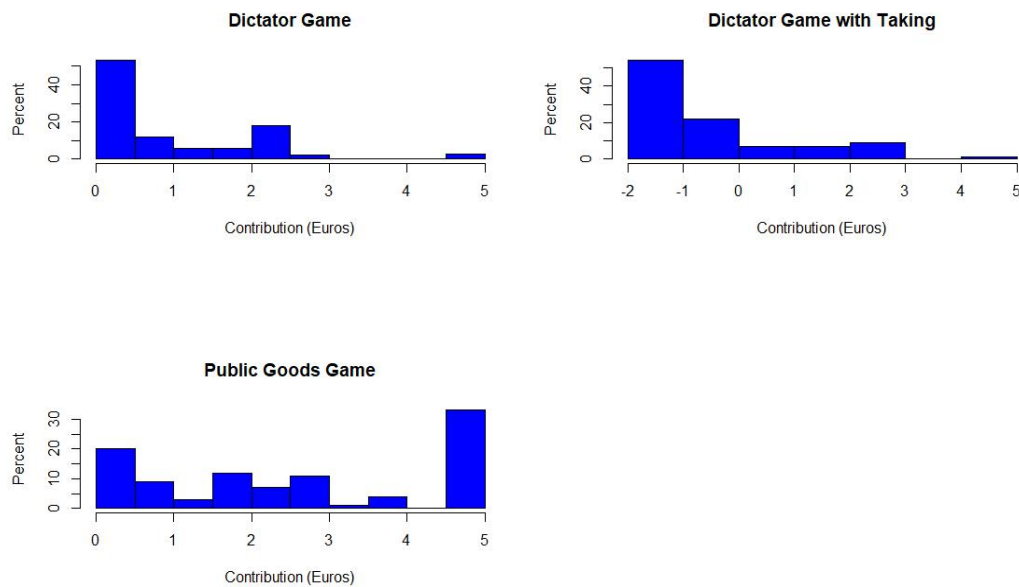
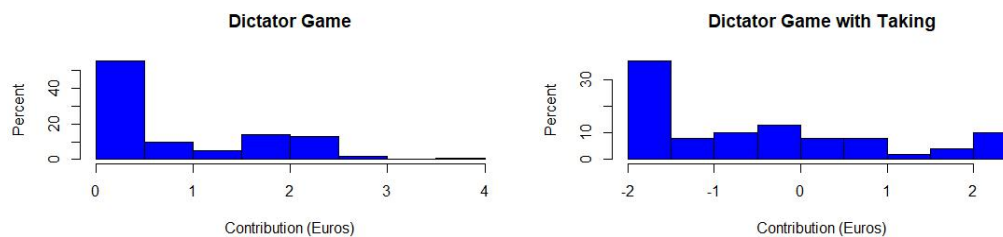
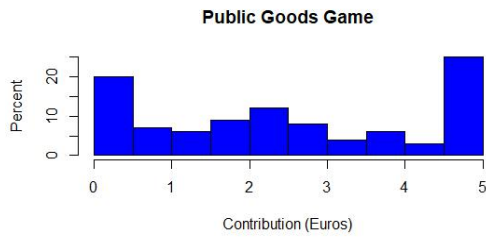


Figure 3b. Distribution of the averaged game decisions over time





The distributions showed in Figure 3a and 3b are quite similar to those obtained in Study 1 (see Figure 1a and 1b). The three one-shot game decisions are also quite similar to the averaged game decisions over time. This suggests responses in the games are quite stable over time. In the dictator game, about 50% of participants didn't share money with their matched partners, and about 20% of participants shared an amount around half of the endowment. When taking options were added to the game, about 50% of participants took the largest amount allowed (2 euros) from their partners, and now only about 10% of participants were willing to share half of the endowment with their partners. In the public goods game, as before, the distribution of decisions spikes at three points: 0 (about 20%), a half of the endowment (about 20%), and the whole endowment (about 30%).

2.3.2.2. Daily pro-social behaviors reported

During this study, the 100 participants reported 709 behaviors. After reviewing these behaviors, the two judges agreed to exclude 49 behaviors as pro-social behavior. The two judges assigned the kept 660 behaviors to 7 categories. See Table 9 for the details.

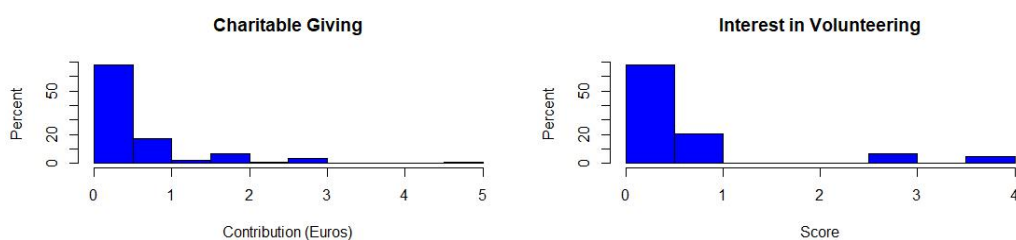
Table 9. The reported pro-social behaviors

	Number of behaviors	Percentage
Helping family members	119	18%
Helping boyfriend/girlfriend	37	6%
Helping friends	144	22%
Helping roommates	24	4%
Helping acquaintances	103	16%
Helping strangers	174	26%
Behaviors for a good cause	59	9%

2.3.2.3. Behaviors in the two naturalistic field situations

Figure 4 below shows the distribution of the 91 participants' responses in the two field situations. The panel on the left describes distribution of the amount donated to a charity by our participants, while the panel on the right shows the distribution of our participants' scores that reflect their interest in volunteering. The participants' scores were constructed from their responses in the following way. If a participant didn't click the link embedded in the email he received, a value of "0" was recorded; a value of "1" was recorded if a participant clicked the link, but didn't check the information about any volunteering activity; and a value of "2" was recorded if a participant checked one volunteering activity but showed no interest; and a value of "3" was recorded if a participant checked one volunteering activity and also showed interest; and a value of "4" was recorded if a participant was interested in more than one volunteering activities or if he or she clicked one of the links leading to the website of the charities. Just in case that some participants didn't notice the email, we sent another identical email about one month after the first email. The responses were coded in the same way. The largest score a participant got in these two rounds was taken as his or her final score in this field situation.

Figure 4. Distribution of behaviors in the two field situations



The two distributions are quite similar to those obtained in Wang and Navarro (2020). As the panel on the left shows, about 60% of participants didn't donate any amount, and the other participants donated an amount between 0.02 euro and 5 euros. In our second field situation, about 67% of participants didn't show any interest in the volunteering opportunities, about 17% of participants went to our website that introduces

volunteering opportunities but left quickly, and only about 16% of participants checked some of these opportunities in detail and said that they want to sign up for them.

2.3.2.4. Correlations between the game decisions and daily pro-social behaviors

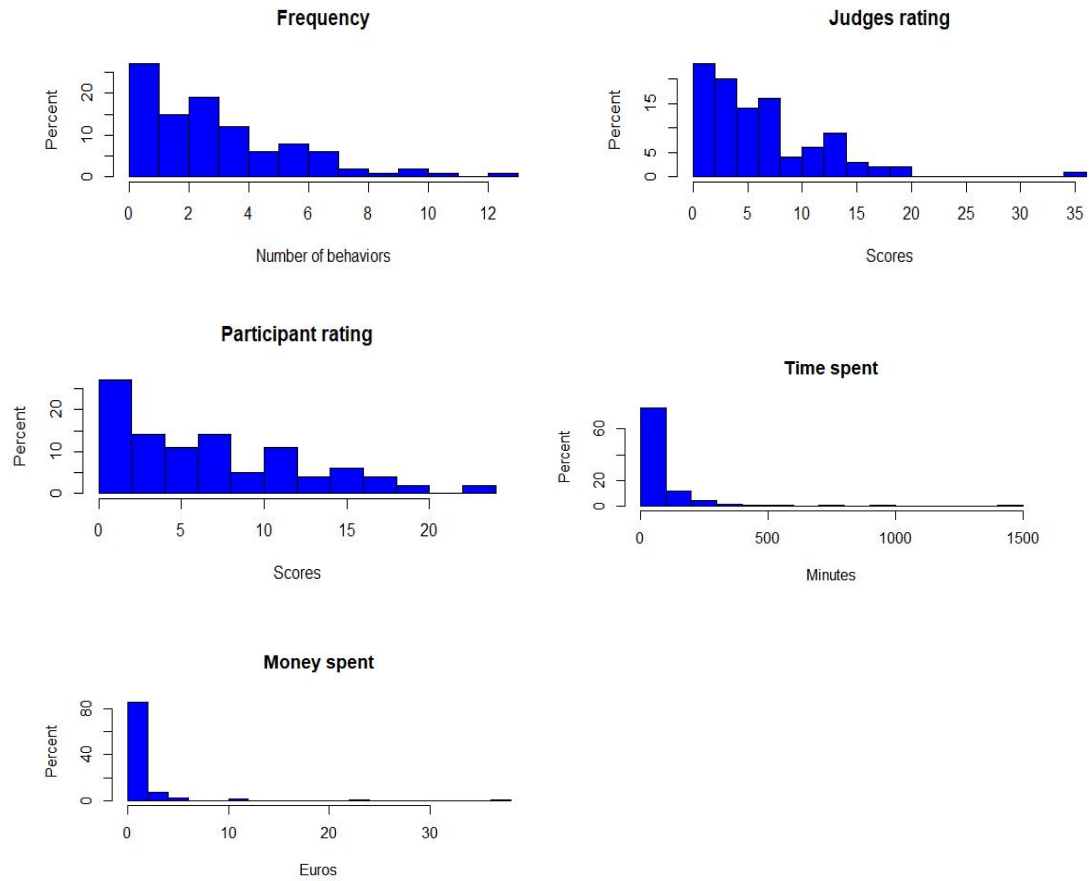
We now explore the relationship between our game decisions and the reported daily pro-social behaviors. The majority of measures used here are the same as those used in Study 1. We focused on six game decisions: the three one-shot game decisions in the introductory session and the three game decisions averaged over time. As before, with the reported daily pro-social behaviors, we constructed five measures of pro-sociality that assign values to different participants by categories. The five measures are frequency of pro-social behaviors, total size of the favors as rated by the judges, total size as rated by the participants, total amount of time spent, and total amount of money spent. The way we constructed these five measures is a bit different from how we did in Study 1. That is, in this study, the five measures only count the total number of reported behaviors or the total amount of resources spent on behaviors in certain categories, no longer controlling for the number of daily survey that each participant completed during the study. The reason is that participants in study 1 reported their behaviors only at night by filling out daily surveys, while participants in study 2 reported pro-social behaviors right after they occurred, thus could be at any time during the study.¹⁹ The same as in study 1, the main daily pro-social behaviors we focus on are those in category A, which is the collection of helping behaviors towards acquaintances or strangers and behaviors for a good cause. We use behaviors in category B (i.e., helping behaviors towards family members or friends) as a reference group.

Figure 5 below shows the distribution of the five scores that constructed from the reported behaviors in category A. All the five distributions in the figure are positive skewed, with less probability density on larger values. Most participants only reported a

¹⁹ We also conducted data analyses controlling for the number of daily surveys completed. All the results are quite similar to those presented here.

few pro-social behaviors in category A, and a large fraction (about 68%) of participants didn't spend any money on these behaviors.

Figure 5. Distribution of the five scores



We now first present the relationship between the six game decisions and the five scores that constructed using the reported behaviors in category A. Table 10 below shows the pairwise Pearson correlation between the game decisions and the five scores.²⁰

²⁰ Table E4 in the appendix shows the Kendall correlations between the game decisions and the scores. The Kendall correlations are broadly in line with the correlations presented in Table 10.

Table 10. Correlations between the game decisions and the five scores

	Frequency	Judges rating	Participant rating	Time spent	Money spent
DG	0.14	0.13	-0.02	0.16	0.04
DGT	0.22**	0.22**	0.04	0.16	0.07
PGG	-0.02	0.00	-0.05	0.03	0.04
DG.Average	0.20*	0.17*	0.00	0.21**	0.04
DGT.Average	0.20**	0.19*	0.03	0.17*	0.08
PGG.Average	-0.00	0.02	-0.07	0.03	0.03

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

As Table 10 shows, decisions in the dictator game and dictator game with taking option have some moderate predictive power for the frequency of engaging in, the objective rating of, and the amount of time spent on behaviors in category A; but the public goods game has no significant predictive power for any of the scores that constructed from these behaviors. This finding is broadly in line with our finding in Study 1.

We also checked the relationship between the game decisions and those reported behaviors in category B. Table 11 shows the pairwise Pearson correlation between the six game decisions and the the two scores that constructed from the reported behaviors in category B.

Table 11. Correlations between game decisions and reported behaviors in category B

	Frequency	Judges rating
DG	-0.09	-0.04
DGT	0.00	0.04
PGG	-0.06	-0.03
DG.Average	-0.04	0.00
DGT.Average	0.03	0.08
PGG.Average	-0.06	-0.03

As Table 11 shows, the six game decisions have no significant correlation with the two scores. As before, this suggests that the significant correlations between the game decisions and those reported behaviors in category A was not generated by self-presentation concern or the individual differences in committing to the study.

We now turn to the relationship between the game decisions and the self-reported altruistic behaviors in the past. Table 12 below shows the pairwise Pearson correlation between the game decisions and the SRA scores obtained in Study 2. In the table, we also present the correlations using the merged data from Study 1 and Study 2.

Table 12. Pairwise correlation between game decisions and SRA scores

	SRA	SRA Money	SRA	SRA Money
DG	0.17*	0.19*	0.18**	0.16**
DGT	0.16	0.15	0.15**	0.12
PGG	0.19*	0.21**	0.18**	0.15**
DG.Average	0.20**	0.21**	0.24***	0.21***
DGT.Average	0.19*	0.19*	0.20***	0.16**
PGG.Average	0.20**	0.23**	0.20***	0.15**

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

As Table 12 shows, all of the three games have some moderate correlations with scores on self-reported altruistic behaviors in the past. This finding provides some further support for the earlier results that our social preference games have some predictive power for daily pro-social behaviors in a long time period.

2.3.2.5. Correlations between the game decisions and behaviors in the field situations

All of our results so far have confirmed the existence of a moderate correlation between the laboratory decisions and daily pro-social behaviors in a long time period. One question remains, however, is whether the observed lab-field correlation could be seen

as a strong one. Or in other words, what size of lab-field correlation we should expect in studies on the external validity of social preference games.

Table 13 shows the pairwise Pearson correlation between the game decisions and the participants' one-shot behaviors in two naturalistic field situations, which could serve as a reference to compare the earlier observed lab-field correlations with.

Table 13. Pairwise Person correlation between game decisions and field behaviors

	Charitable Giving	Interest in Volunteering
DG	0.20*	0.06
DGT	0.32***	0.04
PGG	0.03	-0.07
DG.Average	0.18	0.01
DGT.Average	0.30***	-0.02
PGG.Average	0.04	-0.11

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

As the table shows, the Pearson correlation between decision in the dictator game and charitable giving in the field is 0.20. The correlation between decision in the dictator game with taking option and charitable giving is stronger, around 0.30. But there is no significant correlation between contribution in the public goods game and charitable giving in the field. None of the three games have predictive power for the interest in volunteering. These findings are broadly in line with the reported results in Wang and Navarro-Martinez (2020).

We found that correlations between the game decisions and daily pro-social behaviors reported in a long time period are no larger than that between these game decisions and one-shot donation in our field situation, but much larger than that between the game decisions and the displayed interest in volunteering. It seems that this pattern appears not just because donation behavior was recorded in a well-controlled environment, in which every participant was faced with the same situation. Otherwise, we should have also seen a significant correlation between these game decisions and interest in volunteering. Another reason, we believe, is that most of the reported daily

pro-social behaviors (like helping strangers fix a small issue and recycling) and checking information about volunteering opportunities are very different from sacrificing one’s monetary payoff for others’ interest in social preference games in terms of contexts (e.g., stakes involved). Some support for this was provided by Wang and Navarro-Martinez (2020), which showed that the predictive power of laboratory decision for charitable giving in the field increased dramatically when some crucial lab-field contextual differences were eliminated.

We also found that pro-social behaviors in context-specific situations (in our case, the two field behaviors) have a lower predictive power for daily pro-social behaviors than the dictator games have. Table 14 showed the pairwise Pearson correlation between the two field behaviors and the five scores that constructed from the reported daily pro-social behaviors in category A.

Table 14. Correlations between the field behaviors and the five scores

	Charitable Giving	Interest in volunteering
Frequency	-0.02	0.05
Judges rating	0.06	0.02
Participant rating	-0.08	-0.03
Time spent	-0.02	-0.08
Money spent	-0.05	-0.06

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

The above results together suggest that less contextual differences between a laboratory setting and a field setting leads to higher correlation between behaviors in these two settings, but behavior in a more specific context may have a lower predictive power in relation to settings with different characteristics. This observation highlights the context-dependent nature of social behavior.

2.3.2.6. The predictive power of the Big-Five and empathy in the field

Another reference to compare the predictive power of the game decisions with is that of the psychological measures of the big-five and empathy. The same as in Study 1, we checked the predictive power of these psychological measures for the reported daily

pro-social behaviors. Table 15 shows the pairwise Pearson correlation between the seven psychological measures and the five scores that constructed from the reported behaviors in category A.

Table 15. Pairwise correlation between the psychological measures and the five scores

		Frequency	Judges rating	Participant rating	Time spent	Money spent
Big-Five	Openness	-0.03	-0.08	-0.03	-0.13	-0.14
	Conscientiousness	0.03	0.01	0.04	-0.00	-0.08
	Extraversion	-0.03	-0.03	0.02	-0.02	0.08
	Agreeableness	-0.02	0.03	-0.03	-0.05	0.06
	Neuroticism	-0.02	-0.03	-0.06	-0.17*	-0.11
Empathy	Perspective taking	-0.03	-0.02	-0.11	-0.05	-0.09
	Empathic concern	0.15	0.16	0.10	0.05	0.17*

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

As Table 15 shows, only two out of the 35 correlations are statistically significant at 10% level. Neuroticism correlates negatively with the amount of time spent on behaviors in category A, and Empathic Concern correlates positively with the amount of money spent. The low correlations showed in the table further confirm the low predictive power of these psychological measures for the reported daily pro-social behaviors.²¹

2.4. Conclusions

In this paper, we set out to investigate the power of social preference games to predict daily pro-social behaviors in the field when those behaviors are aggregated over time. Our findings show that, when the period of aggregation is long enough, the games

²¹ Table E5 in the appendix summarizes some of our regression analyses, in which the psychological measures were used to predict the five scores. These results are broadly in line with the correlations showed in the table.

significantly (although moderately) correlate with daily behaviors, even though those behaviors are diverse and do not share obvious characteristics with the games. This suggests that social preference games capture some underlying component of pro-sociality that is relevant across situations. This is analogous to the “person” side of the longstanding “person-situation” debate in psychology, which argues that personality traits do not predict well single behaviors, but predict behavior much better over longer periods of time (see, e.g., Fleeson, 2001, 2004).

On the other hand, the highest correlation between the game decisions and the daily pro-social behaviors that we obtained was 0.24. This is clearly higher than the correlations with single situations that do not resemble the games, but we have seen that it is not higher than the correlation with a naturalistic field situation (giving money to a charity solicitor) that bears some similarities with the games. This pattern is even clearer in the results of Wang and Navarro-Martinez (2020), in which making a lab game contextually similar to one specific field situation achieved a correlation of 0.65 between the two. This shows that pro-social behavior highly depends on the context and cannot be completely captured by context-free games. This is reminiscent of the “situation” side of “person-situation” debate, which argues that contextual elements are more important than personality in predicting behavior in specific situations (Fleeson, 2001, 2004).

Some previous research (e.g., Benz and Meier, 2008; Galizzi and Navarro-Martinez, 2019) has proposed that the person-situation debate in psychology should be taken more seriously in experimental economics. Our results suggest again that a person-situation framework is very useful to understand external validity in the domain of economic games. This dichotomy not only helps us in interpreting the low lab-field correlations widely documented in the literature, but also in suggesting interesting research avenues. We believe that, on the one hand, future research on social preferences needs to further investigate how well and under which circumstances social preference games predict social behaviors aggregated over long time periods. This will tell us how far context-free games can go in explaining real world behaviors. On the other hand, if social preference games want to address behavior in specific field domains, they should incorporate more context that brings them closer to the real-world situations they want to address.

Appendix

Appendix A. The Self-Report Altruism (SRA) Scale

	<i>Never</i>	<i>Once</i>	<i>More than once</i>	<i>Often</i>	<i>Very often</i>
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.					

Notes: Items 4, 13 and 15 involves spending money. Together they are used as a subscale, SRA money.

Appendix B. Interpersonal Reactivity Index

The following statements inquire about your thoughts and feelings in a variety of situations. For each item, indicate how well it describes you by choosing a number from 1 (does not describe me at all) to 5 (describes me very well).

1. I often have tender, concerned feelings for people less fortunate than me. (EC)
2. I sometimes find it difficult to see things from the "other guy's" point of view. (PT) (R)
3. Sometimes I don't feel very sorry for other people when they are having problems. (EC) (R)
4. I try to look at everybody's side of a disagreement before I make a decision. (PT)
5. When I see someone being taken advantage of, I feel kind of protective towards them. (EC)
6. I sometimes try to understand my friends better by imagining how things look from their perspective. (PT)
7. Other people's misfortunes do not usually disturb me a great deal. (EC) (R)
8. If I'm sure I'm right about something, I don't waste much time listening to other people's arguments. (PT) (R)
9. When I see someone being treated unfairly, I sometimes don't feel very much pity for them. (EC) (R)
10. I am often quite touched by things that I see happen. (EC)
11. I believe that there are two sides to every question and try to look at them both. (PT)

12. I would describe myself as a pretty soft-hearted person. (EC)
13. When I'm upset at someone, I usually try to "put myself in his shoes" for a while. (PT)
14. Before criticizing somebody, I try to imagine how I would feel if I were in their place. (PT)

Notes: Statements with R mark are reverse statements.

Appendix C. Survey inquiring about the big-five personality states

Here are a number of traits that may or may not describe well your behaviors and feelings today. Please indicate below how well each pair of traits describe you today. You should rate how well the pair of traits describe you, even if one characteristic describe you much better than the other. Please use the scale given below.

Strongly disagree, 1; Moderately disagree, 2; Neither agree nor disagree, 3; Moderately agree, 4; Strongly agree, 5.

- _____ Extraverted, enthusiastic.
- _____ Critical, quarrelsome.
- _____ Dependable, self-disciplined.
- _____ Anxious, easily upset.
- _____ Open to new experiences, complex.
- _____ Reserved, quiet.
- _____ Sympathetic, warm.
- _____ Disorganized, careless.
- _____ Calm, emotionally stable.
- _____ Conventional, uncreative.

Appendix D. Survey inquiring about the empathic concern and perspective taking personality states

The following statements inquire about your thoughts and feelings today. For each item, indicate how well it describes your thoughts and feelings today by choosing the appropriate number from: 1, 2, 3, 4, and 5. “1” indicates the statement doesn't describe you well at all, and “5” indicates the statement describes you extremely well.

1. Today, I would find it difficult to see things from the "other person's" point of view. PT (R)
2. Today, if I saw someone being taken advantage of, I would feel kind of protective towards them. EC
3. Today, I feel like trying to understand my friends better by imagining how things look from their perspective. PT
4. Today other people's misfortunes would not disturb me a great deal. EC (R)
5. Today, if I were sure I am right about something, I wouldn't waste much time listening to other people's arguments. PT (R)
6. Today, if I saw someone being treated unfairly, I wouldn't feel very much pity for them. EC (R)
7. Today I am quite touched by things that I see happen. EC
8. Today I believe that there are two sides to every question and try to look at them both. PT
9. Today, if I were upset at someone, I would try to "put myself in their shoes" for a while. PT

Notes: Statements with R mark are reverse statements.

Appendix E.

Table E1. Correlations between the game decisions and the five scores (Kendall, Study 1)

	Frequency	Judges rating	Participant rating	Time spent	Money spent
DG	0.14*	0.16**	0.13	0.10	0.10
DGT	0.19**	0.20**	0.18**	0.14*	0.02
PGG	0.01	0.06	0.09	0.18**	0.08
DG.Average	0.08	0.11	0.07	0.10	0.04
DGT.Average	0.14*	0.17**	0.11	0.18**	0.00
PGG.Average	-0.05	-0.01	-0.01	0.15**	0.01

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

Table E2. Regression analysis (OLS, Study 1): The predictive power of psychological variables

		Frequency	Judges rating	Participant rating	Time spent	Money spent
Big-Five	Openness	0.00	0.01	0.02	0.21	0.02
	Conscientiousness	0.00	0.00	0.01	-0.07	-0.03
	Extraversion	0.02	0.02	0.04	-0.14	-0.02
	Agreeableness	0.01	0.02	0.04	0.12	0.00
	Neuroticism	0.01	0.01	0.01	0.15	-0.03
Empathy	Perspective taking	0.01	0.01	0.01	0.03	0.01
	Empathic concern	-0.02*	-0.02	-0.02	0.06	0.03
R2		0.06	0.04	0.07	0.02	0.05
Adjusted R2		-0.02	-0.04	-0.01	-0.07	-0.04
Observations		83	83	83	83	83

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

Table E3. Regression analysis (OLS, Study 1): Means of distributions of psychological variables

		Frequency	Judges rating	Participant rating	Time spent	Money spent
Big-Five	Openness	-0.00	-0.03	0.02	-1.55*	-0.00
	Conscientiousness	-0.00	-0.04	-0.03	-1.20*	0.02
	Extroversion	-0.01	0.00	-0.02	-0.81	0.07
	Agreeableness	0.04	0.11	0.19	1.88	0.03
	Neuroticism	0.03	0.06	0.02	0.78	0.04
Empathy	Perspective taking	-0.00	-0.00	-0.02	0.22	0.00
	Empathic concern	-0.01	-0.01	-0.05	-0.12	-0.02
R2		0.04	0.07	0.05	0.15	0.02
Adjusted R2		-0.05	-0.02	-0.04	0.07	-0.07
Observations		83	83	83	83	83

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

Table E4. Correlations between the game decisions and the five scores (Kendall, Study 2)

	Frequency	Judges rating	Participant rating	Time spent	Money spent
DG	0.11	0.11	0.03	0.12	0.09
DGT	0.18**	0.17**	0.09	0.13*	0.13
PGG	-0.04	-0.03	-0.05	0.08	0.02
DG.Average	0.10	0.11	0.03	0.13*	0.12
DGT.Average	0.12	0.12*	0.07	0.15**	0.16*
PGG.Average	-0.03	-0.03	-0.05	0.09	0.00

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

Table E5. Regression analysis (OLS, Study 2): The predictive power of general traits

		Frequency	Judges rating	Participant rating	Time spent	Money spent
Big-Five	Openness	-0.04	-0.16	-0.11	-7.76	-0.24*
	Conscientiousness	0.04	0.10	0.07	-1.06	-0.07
	Extroversion	-0.10	-0.29	-0.03	-5.63	-0.02
	Agreeableness	0.05	0.24	0.09	2.13	0.17
	Neuroticism	-0.04	-0.13	-0.14	-12.4	-0.19
Empathy	Perspective taking	-0.08	-0.19	-0.29	-6.57	-0.32**
	Empathic concern	0.14*	0.31**	0.25	5.32	0.28**
R2		0.04	0.06	0.05	0.06	0.12
Adjusted R2		-0.03	-0.01	-0.03	-0.01	0.06
Observations		100	100	100	100	100

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

3. SOCIAL PREFERENCE GAMES GO TO WORK: TESTING FOR EXTERNAL VALIDITY IN THE WORKPLACE

Joint with Pinghan Liang and Daniel Navarro-Martinez

3.1. Introduction

The last few decades have witnessed an unforeseen surge of research interest in “social preferences”. An important feature of this line of research is that it evolved out of developing and running so called "social preference games", experimental games designed to target certain non-pecuniary motives, e.g., altruism, cooperation and reciprocity. These games share some prominent features: context-free instructions, clear game-theoretic structures and real monetary incentives. The most often studied social preference games are perhaps dictator game (e.g., Kahneman et al. (1986)), ultimatum game (e.g., Guth et al. (1982)), trust game (e.g., Berg et al. (1995)), and public goods game (e.g., Ledyard (1995)). These games have become one of the main building blocks in social preference studies in experimental economics.

A question here, however, is that to what extent these games tap into the the principles that determine social behavior outside the laboratory. This is a fundamental question to be answered, because what behavioral science really wants to study is natural social behavior in the real world, not just how participants play these artificial games in the lab. We can draw reliable conclusions about social behavior from decisions in these games only if the game decisions represent well natural social behavior.

So far, however, very few systematic investigation of the external validity of these games have been done, and the accumulated evidence is very sparse and mixed. In a meta analysis conducted by Galizzi and Navarro-Martinez (2019), the authors only found 39 within-subject studies that investigated the predictive power of social preference games for pro-social behaviors in the field. More importantly, only 39.7% of the lab-field correlations reported in these studies were statistically significant, and the average correlation between game decisions and field behaviors was 0.14. The authors also conducted a systematic experiment comparing various lab games against several

pro-social behaviors in the field and found no significant correlation between them. These findings suggest that social preference games seem to have the issue of external validity. However, to reach final conclusions about the external validity of all these games, more systematic lab-field investigations, especially those with different subject pools and different sets of behaviors, need to be conducted.

In this project, we systematically assess the external validity of a variety of social preference games, using a comprehensive data set we collected at the workplace. We first bring the lab to the field: Employees at five high-end hotels in China played a variety of social preferences games, which are supposed to capture altruism, cooperation, reciprocity, trust and trustworthiness. These games include the dictator game, the ultimatum game, the public goods game, a conditional public goods game, the trust game, and the gift exchange game. We also collected data from the hotels about these employees' performance in their daily work. These employees' supervisors assessed their work performance using a standard Organizational Citizenship Behavior scale. Organizational citizenship behavior (OCB) is a concept that we borrowed from the industrial and organizational psychology, referring to employees' pro-social behavior that is directed at coworkers or the organization but not directly rewarded by the organizations. We focus on OCB in three aspects: (1) Identification with the company, such as protecting reputation of the company and making constructive suggestions; (2) Altruism, mainly helping co-workers with work-related issues; (3) Conscientiousness, such as doing work that goes beyond the minimum requirement of one's job. We also inquired the employees about their perceived respect and fairness inside the organization. We then compare these employees' social preferences elicited from the games with their OCB scores, controlling for perceived respect and fairness in the organization.

We are collaborating with the hotels on cleaning some detailed records of the employees' work performance in the last two years, including all the rewards, honors, bonuses, or punishment given to the employees for their work or behavior, the frequency of being late or absent from work, and the frequency of helping co-workers cover their work or working for extra hours voluntarily. These detailed records will provide us some objective measures of these employees' work performance. But this part is still work in process, thus won't be included in the data analysis of this chapter.

Our results suggest that social preference games do capture some characteristics of the players at group level. For example, we found that senior managers, compared to ordinary employees, displayed more trust toward their matched partners in the trust game and in the gift exchange game. However, social preferences elicited from these games do not predict well individual employee's organizational behaviors. Our findings question again the external validity of social preference games and suggest that some crucial elements for organizational behavior are missing in these games.

The rest of the paper is organized as follows: Section 3.2 describes our research methods, Section 3.3 presents the current results we have, Section 3.4 discusses our findings.

3.2. Methods

To test whether game decisions in the laboratory predict well work performance in the real world, we combine three types of data that collected from or for the same sample of participants: (1) social preferences elicited in a variety of laboratory games; (2) field measures of work performance; (3) perceived fairness and respect inside the companies and personality traits.

3.2.1. The social preference games

Participants in our study played the same set of twelve social preference games, which were further separated into two sessions, with six games in each session. The games were played in the order specified below:

- 1) Dictator Game 1 (DG1): In this game, two players were paired up. Player 1 decides how to divide RMB30 between the self and Player 2; Player 2 simply receives the amount indicated by Player 1.

- 2) Dictator Game 2 (DG2): Like Dictator Game 1, but the roles were switched and players were matched with new partners.

- 3) Ultimatum Game 1 (UG1): This is a two-player game in which Player 1 proposed how to divide RMB30 between the self and Player 2. Player 1 can only choose an amount to offer from $\{0,2,5,8,10,13,15,17,20,25,30\}$. Player 2 indicates whether to accept the allocation or not for each possible allocation proposed by Player 1. If the actual proposed allocation is rejected, both players get nothing; if the actual proposed allocation is accepted, the two players share the RMB30 accordingly.
- 4) Ultimatum Game 2 (UG2): Like Ultimatum Game 1, but the roles were switched and players were matched with new partners.
- 5) Public Goods Game (PGG): In this game, three players were grouped together. Each player has an endowment of RMB30 and has to decide simultaneously how much of it to contribute to a common group fund. The choice set in this game is set to be $\{0,6,12,18,24,30\}$. The overall money in the group fund is then multiplied by two and split evenly among the three players.
- 6) Conditional Public Goods Game (CPGG): This is a three-player game in which all the players have an endowment of RMB30. All the players were matched with their teammates in the previous Public Goods Game. They were required to indicate the amount they would like to contribute to the group fund, for each possible averaged contributions of their two teammates in the previous game. After all the players have made decisions, the actual contribution of each player is calculated by matching the indicated strategies and their teammates' actual contribution in the previous game. The overall money in the group fund is then multiplied by two and split among the three players.
- 7) Trust Game 1 (TG1): In this game, two players were paired up. Player 1 has an endowment of RMB30 and make a decision about the amount to send to Player 2. The choice set for Player 1 is $\{0,5,10,15,20,25,30\}$. The amount Player 2 actually receives is always the amount sent over multiplied by three. Player 2 needs to indicate the amount to send back to Player 1 for each possible amount sent over.
- 8) Trust Game 2 (TG2): Like Trust Game 1, but the roles were switched and players were matched with new partners.

- 9) Trust Game 1 with Punishment (TGP1): This is a two-player game in which Player 1 has an endowment of RMB30 and decides how much of it to send over to Player 2. When sending money, Player 1 is also asked to tell Player 2 what minimum amount she requires, and whether she would like to impose a conditional punishment on Player 2. The conditional punishment of reducing RMB12 would be enforced, if Player 2 returns less money than Player 1 required and if Player 1 imposed a conditional punishment. The amount sent over is multiplied by three and given to Player 2, who then decides the amount to send back to Player 1. Before making decision, Player 2 is informed about the minimum amount required and whether a conditional punishment is imposed.
- 10) Trust Game 2 with Punishment (TGP2): Like Trust Game with Punishment 1, but the roles were switched and players were matched with new partners.
- 11) Gift Exchange Game 1 (GE1): In this game, the two players were randomly assigned a role, either a principal or an agent. Principal has an endowment of RMB 30 and chooses a wage to pay the agent from the set {0,5,10,15,20,25,30}. The agent indicates her effort level for each possible wage the principal pays. Table 1 below summarizes information of the payoff. Half of the participants were Player 1 and the other half Player 2.
- 12) Gift Exchange Game 2 (GE2): Like Gift Exchange Game 1, but the roles were switched and players were matched with new partners.

Table 1. Payoffs in the gift exchange game

Effort Level	0	1	2	3	4	5
Cost (RMB)	0	2	7	10	15	20
Revenue (RMB)	0	5	12	20	30	50

These games target a comprehensive category of social preferences, including altruism, reciprocity, cooperation, trust and trustworthiness. In the literature of

experimental economics, giving in Dictator Game 1 and 2 is usually interpreted as altruism; rejecting small offers in the Ultimatum Game 1 and 2 is understood as punishment of unfairness (negative reciprocity) or as inequity aversion; contributing in the Public Goods Game is seen as cooperation; the strategy of contributing in the Conditional Public Goods Game reveals a participant's type, either as unconditional contributor, conditional contributor or as free rider (see Fischbacher et al. 2001; Fischbacher and Gächter 2010); decisions of Player 1 in the Trust Game and Trust Game with Punishment reveal participants' trust in others, while decisions of Player 2 reflects the players' trustworthiness (or positive reciprocity); the amount of wage offered by the principle in the Gift Exchange Game 1 and 2 indicates a player's trust while effort put by the Player 2 reflects trustworthiness (or positive reciprocity).

At the beginning of the experiment, the participants were assigned a random ID codes. Specific and detailed instructions were displayed at the beginning of each task. All the instructions was translated in Chinese. To make sure the participants understand well the tasks, before the participants proceeded, the experimenter read aloud the instructions for each task and answered all the questions that raised about the task rules. In all of these games, players were matched anonymously. The software we used to implement the lab games was oTree (Chen et al., 2016).

A fixed amount of RMB70 was promised for finishing all the two sessions. We didn't provide any information or feedback about the results of the different games to the participants in the middle of the experiment. At the end of session 2, for each participant, one of the twelve played tasks was randomly selected, the participant's payoff in which was paid as an extra bonus. All the participants were asked to sign a receipt after they get paid. In the process of making payment, we matched secretly each participant's name with his (her) lab ID code.

3.2.2. Organizational citizenship behavior

Organizational citizenship behavior (OCB) is a concept intensively studied in industrial and organizational psychology and management, the definition of which is employees' pro-social behavior that is directed at coworkers and the organization but not directly

rewarded by the organizations. OCB is crucial for an organization to operate smoothly and to succeed, but the employees' engaging in these behaviors are usually voluntary and hard to monitor.

We used a standard OCB scale that developed by Farh et al. (1997) and Zhao et al. (2012), which is designed for the Chinese context. This scale contains thirteen questions for OCB in three aspects: (1) Identification with the company, such as protecting reputation of the company and making constructive suggestions; (2) Altruism, mainly helping co-workers with work-related issues; (3) Conscientiousness, such as doing work that goes beyond the minimum requirement of one's job. All the participants (both ordinary employees and managers) in our study were rated by their direct supervisors using the scale. The whole scale consists of 13 items: 4 about Identification, 4 about Altruism, and 5 about Conscientiousness. Each item is rated on a scale from 1 to 5, with 1 refers to a very low level of engaging, and 5 a very high level. See Appendix A for the scale and instructions.

3.2.3. Survey data: other variables

During our laboratory study, the participants fill out surveys that inquire about their perceived respect and fairness inside the company, big-five personality traits and some demographic information.

3.2.3.1. Perceived respect and fairness

Some previous studies suggest that perceived respect and fairness are associated with employees' organizational citizenship behavior (e.g., Organ, 1988; Organ and Ryan, 1995; Borman and Penner, 2001; Rioux and Penner, 2001; Blader and Tyler, 2009; Zhang et al., 2017). To assess how important perceived respect and fairness is for our participants' OCB in their daily work, we inquired the participants about four specific aspects of respect and fairness inside their company using an adapted survey that designed by Blader and Tyler (2009).

The first aspect is perceived respect, that is, the degree to which supervisors value employees' unique contribution to the team. The second aspect is procedural fairness, i.e., the degree to which fair procedure is used in the company to deal with issues and make decisions. The third aspect is distributive justice, i.e. how fair benefits are distributed among employees. The last aspect is incentives, that is, how strongly work effort is associated with economic rewards inside the organization. See Appendix B for the whole survey and instructions.

3.2.3.2. Personality traits

The participants also responded to a questionnaire that measures the Big-Five personality dimensions (John, 1991): Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These five aspects of personality are fundamental constructs in psychology and they have been shown to explain a wide variety of behaviors (e.g., Barrick and Mount, 1991; Paunonen, 2003; Giluk and Postlethwaite, 2015). We thus included them to control for their effect on the employees' organizational behavior.

3.2.3.3. Demographic Information

We also included survey questions that asked about the participants' job position, education level, income, gender, age, and the length of time working in the company.

3.2.3.4. Participants

We worked with five high-end hotels in Sichuan province, China. These hotels have similar organizational structure and adopt similar management practices and rewarding system. All of these hotels have eight departments that in charge of different functions. The departments are guest room department, reception center, restaurant, marketing, accounting, human resource department, maintenance and security. The hotels evaluate their employees' work performance monthly and yearly. They also established some single rewards for employees with excellent work performance, such as rewards for

those that return found lost items of customers, or those that suggest a constructive idea to the hotel.

In total, 306 employees (including 36 senior managers, 82 junior managers, and 188 ordinary employees) working at the hotels attended our study. One participant dropped out of the study in the middle. Here senior managers are mainly heads of one department while junior managers are those supervisors that monitor a small group in the department. Among these participants, 69% are female. About 30% of them are below 25 years, about 47% of them have an age that between 25 years and 40 years, and 21% of them are above 40 years but below 50 years. The average age is 32.3 years. Regarding education level, about 31% of them completed vocational college or university education, 33% of them finished senior high, 30% of them finished junior high, and the rest of them only finished elementary school. The average income of our participants is about RMB 3500. Hotel industry typically has a high turnover rate. The mean tenure of our participants with the hotels is 2.9 years when they attended our study. At that moment, 83% of participants have been working in the hotels more than one year, and 90% more than half a year.²²

3.3. Results

3.3.1. Game decisions

We constructed nine parameters from decisions in the social preference games²³, each of which reflects some aspects of social preferences in the lab. A general pattern is that, compared to distributions of these parameters that generated by college students' responses (see e.g., Galizzi and Navarro-Martinez, 2019), our participants behaved more pro-socially. In the rest part of this section, we will explain each parameter in detail and presents its distribution among the participants.

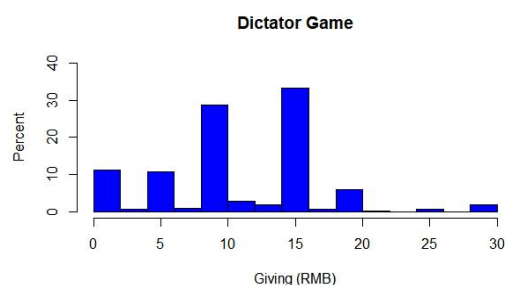
²² 22 participants worked less than a quarter inside the hotels. We will drop these 22 participants' responses in some of our later analyses when including them could affect the reliability of the results.

²³ Decisions in the trust game with punishment can vary on several dimensions. It is hard to construct simple but precise parameters from these decisions. We thus doesn't include decisions in that game in the current data analysis.

1. Dictator game

The first parameter we have is the amount of endowment that shared in the dictator game. Figure 1 below shows the distribution of this parameter. As the figure shows, most of the participants shared a positive amount with their matched partner in this game, with a high 35% spike at RMB15 (half of the endowment); and only about 11% of participants didn't share any money with Player 2.

Figure 1. Distribution of giving in the dictator game

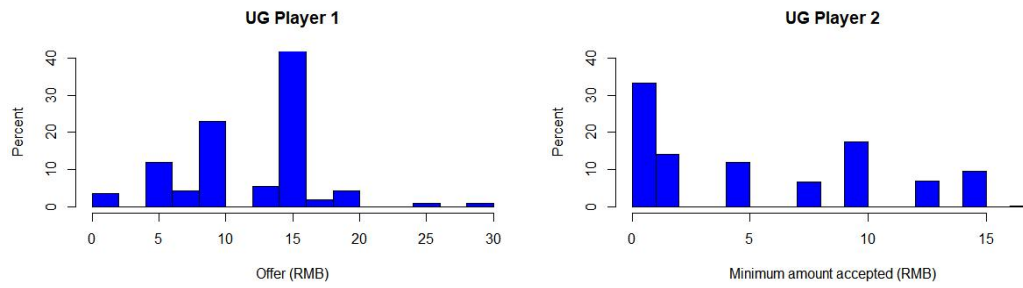


2. Ultimatum game

We constructed two parameters from the ultimatum game. One is the amount proposed by Player 1, and the other one is the minimum of amount that Player 2 would accept.

The left panel of Figure 2 describes the distribution of the proposed offer, while the right panel shows the distribution of the minimum amount that Player 2 would accept. Player 1 in this game was more generous than its counterpart was in the dictator game: now about 40% of Player 1 proposed an amount around RMB15, and only 3% at RMB0. Player 2 in this game showed a strong negative reciprocity: about 34% of participants would reject small offers that below RMB10.

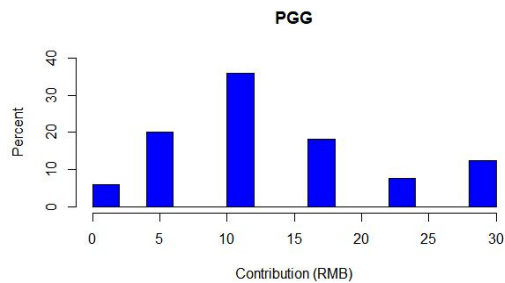
Figure 2. Distribution of decisions in the ultimatum game



3. Public goods game

Figure 3 shows the distribution of contribution in the public goods game. Remember that in this game, the participants can only choose an amount from the choice set $\{0,6,12,18,24,30\}$. As the figure shows, most of players contributed a positive amount, with a spike 35% at RMB12, and only about 6% of participants contributed RMB0.

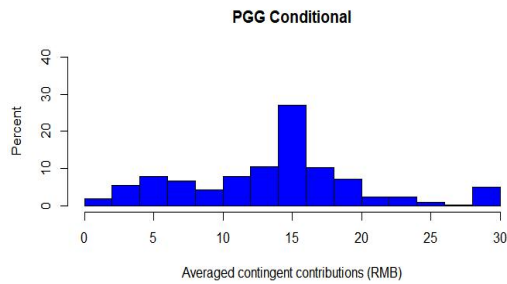
Figure 3. Distribution of decisions in the PGG



4. Conditional public goods game

In this game, each player needs to indicate his contribution for each possible level of contribution of his teammates in the PGG. Thus, the parameter we constructed from responses in this game is the averaged contingent contributions. Figure 4 shows the distribution of this parameter. We can see that most participants are conditional cooperators in this game, i.e., contributing more if others contributed more, and there is only about 2% of free riders (i.e., those that always contribute 0) and about 5% of unconditional cooperators (i.e., those that always contribute all of the endowment).

Figure 4. Distribution of decisions in the Conditional PGG



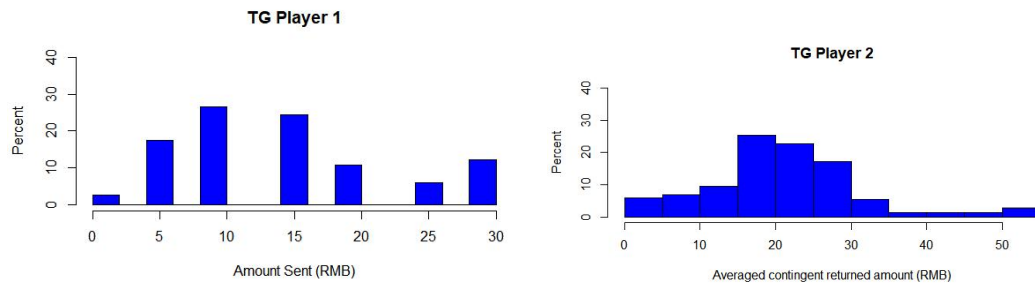
5. Trust game

We constructed two parameters from the responses in the trust game. For Player 1, it is simply the amount sent to Player 2. The left panel of Figure 5 shows the distribution of this parameter. We can see that most Player 1 showed some level of trust for Player 2 by sending a positive amount, and that only about 2% of players sent RMB0.

Since Player 2 needs to decide the amount of sending back for each possible amount received, the averaged contingent transfers is used as a parameter measuring social preferences. The right panel of the figure describes the distribution of this parameter. As the panel shows, most Player 2 returned a positive amount to Player 1, and only about 6% of participants always sent back nothing. A pattern is that a bigger amount is sent back for a larger investment: the Pearson correlation between those two is 0.62 if we put together all the contingent responses of all the Player 2.²⁴

²⁴ If we only consider the actual sent amounts and sent back amounts, the Pearson correlation between these two is 0.67.

Figure 5. Distribution of decisions in the trust game



6. Gift exchange game

Decisions in the gift exchange game generated two interesting parameters. One is the wage paid by the Principle, and the other is the averaged contingent effort that exerted by the Agent. The two panels of Figure 6 show the distribution of the two parameters. As the left panel of Figure 6 shows, most principles paid the agents a positive wage, with a spike 30% at RMB20; and only about 3% of principles paid a wage of RMB0. From the right panel, we see that most agents exerted a positive level of effort, and only about 2% of agents indicated an effort level of 0 for any possible wage. As in the trust game, we found a strong evidence of positive reciprocity: the Pearson correlation between wage and effort is 0.69 if we put together all the contingent responses of the agents.²⁵

Figure 6. The distributions of decisions in the gift exchange game

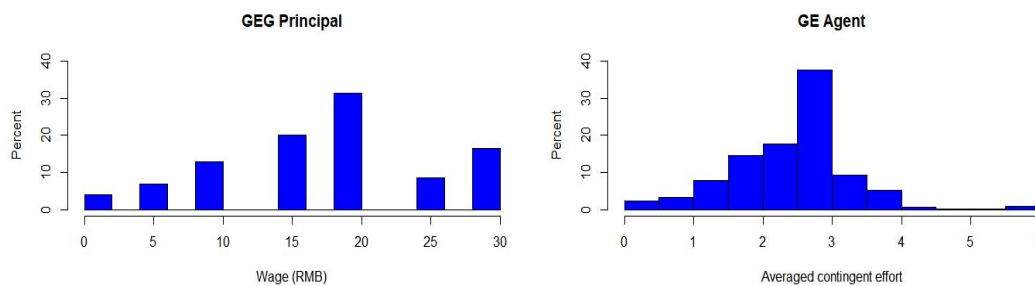


Table 2 summarizes the pairwise Pearson correlation between the nine parameters we constructed. The majority of correlations in the table are statistically significant at the 1% level. Only behavior of Player 2 in the Ultimatum Game has no correlation with the

²⁵ We observed a same level of Pearson correlation between the paid wage and the enforced effort level.

other game decisions. This suggests that rejecting unfair offers is very different from sacrificing one's own payoff for others' interest. In general, we observe a very high level of internal consistency among decisions in the social preference games.

Table 2. Pairwise correlations between game decisions (Pearson)

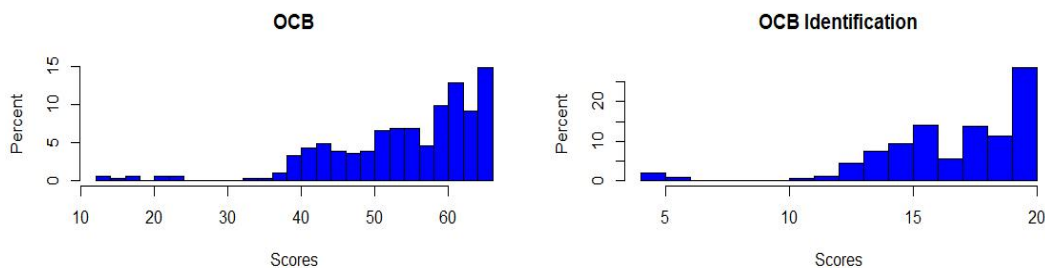
	<i>DG</i> <i>P.1</i>	<i>UG</i> <i>P.1</i>	<i>UG</i> <i>P.2</i>	<i>PGG</i>	<i>CPGG</i>	<i>TG</i> <i>P.1</i>	<i>TG</i> <i>P.2</i>	<i>GEG</i> <i>P.1</i>	<i>GEG</i> <i>P.2</i>
<i>DG</i> <i>P.1</i>	1.00***	0.47***	-0.00	0.22***	0.2***	0.20***	0.32***	0.11**	0.14**
<i>UG</i> <i>P.1</i>	0.47***	1.00***	0.09	0.28***	0.22***	0.22***	0.25***	0.17***	0.06
<i>UG</i> <i>P.2</i>	-0.00	0.09	1.00***	-0.01	0.00	0.09	0.09	-0.04	-0.06
<i>PGG</i>	0.22***	0.28***	-0.01	1.00***	0.55***	0.50***	0.36***	0.35***	0.20***
<i>CPGG</i>	0.2***	0.22***	0.00	0.55***	1.00***	0.48***	0.49***	0.27***	0.29***
<i>TG</i> <i>P.1</i>	0.20***	0.22***	0.09	0.50***	0.48***	1.00***	0.46***	0.35***	0.25***
<i>TG</i> <i>P.2</i>	0.32***	0.25***	0.09	0.36***	0.49***	0.46***	1.00***	0.29***	0.20***
<i>GEG</i> <i>P.1</i>	0.11**	0.17***	-0.04	0.35***	0.27***	0.35***	0.29***	1.00***	0.25***
<i>GEG</i> <i>P.2</i>	0.14**	0.06	-0.06	0.20***	0.29***	0.25***	0.20***	0.25***	1.00***

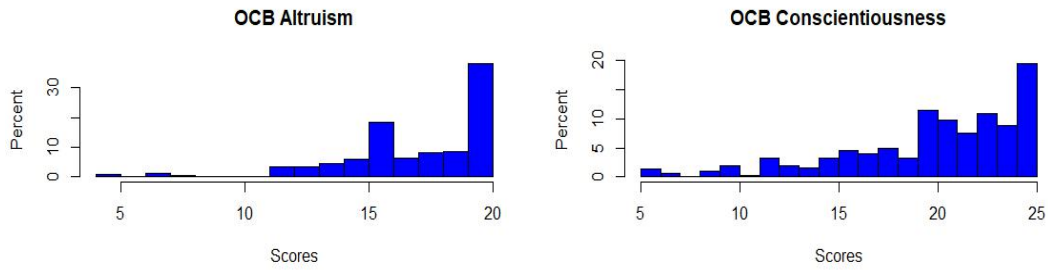
Notes: “*”, “**” and “***” stand for statistical significance at the 10%, 5% and 1% level respectively.

3.3.2. Organizational citizenship behavior

Figure 7 below describes distribution of the OCB scores, both those employees' total scores on the whole OCB scale and their scores on each of the three sub-categories of OCB.

Figure 7. Distribution of the OCB scores



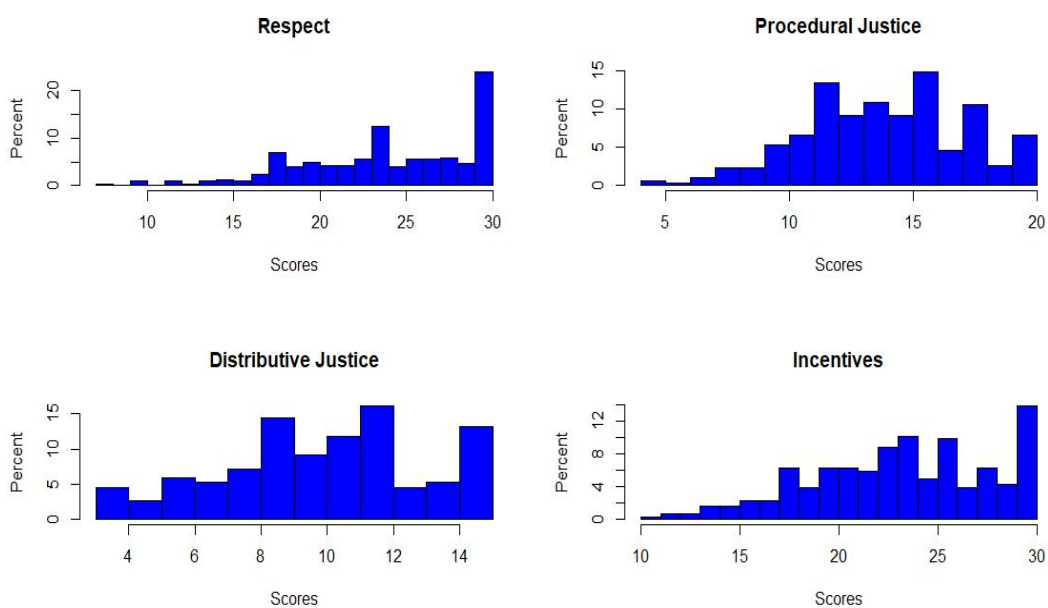


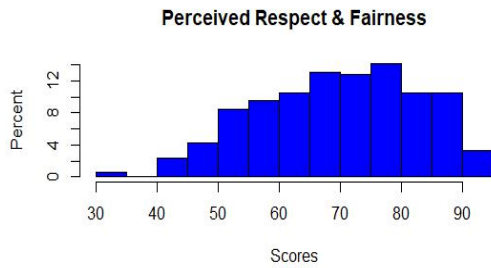
As the figure shows, all the four distributions have a very similar shape. The distributions have a big dispersion, with the majority of scores on values that refer to high level of engaging. This is probably because the hotels we work with are high-end ones and those employees having a poor work performance are replaced frequently.

3.3.3. Perceived respect and fairness

Figure 8 below consists of five panels, which together show distribution of the employees' perceived respect and fairness inside the organizations. The first four panels correspond to the four aspects of these perception. They are respect, procedural justice, distributive justice, and incentives. The last panel presents the distribution of these perception as a whole.

Figure 8. The distributions of perceived respect and justice





As Figure 8 shows, there is a wide variety in the perceived respect and fairness, with the majority of participants being satisfied but still a big fraction of employees complained about the organizations’ practices on these aspects. More specifically, about 8% of employees reported a low level of perceived respect (a score less than 18); about 19% of employees reported a low level of perceived procedural justice (a score less than 12); about 26% of employees reported a low level of distributive justice (a score less than 9); and about 9% of employees think their effort is not well rewarded in the company (giving a score less than 18). It seems that those employees were in general satisfied with their relationship with their supervisors and the economic incentives provided by the organizations, but were less happier with the procedural justice and distributive justice inside the organizations. The observed wide variety of the scores suggests that those employees seem not having concern for reporting their true feelings when responding to our surveys.

Table 3 below reported the Cronbach’s alphas of these four scales (the numbers on the diagonal) and the pairwise Pearson correlations among them.

Table 3. Cronbach’s alphas and cross-factor correlations

	Respect	Procedural Justice	Distribution Justice	Incentives
Respect	0.93			
Procedural Justice	0.45***	0.76		
Distributive Justice	0.40***	0.67***	0.90	
Incentives	0.54***	0.53***	0.60***	0.76

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

The Cronbach's alphas in the table are very high, suggesting that these scales possess a high level of internal consistency. We also observed very high correlations among these scales, suggesting that an employee tends to give a low rating of the other aspects of respect and fairness inside the organization if her perception of one aspect is low.

We also explored the basis of the employees' perceived respect and fairness. First, we observed that all the four aspects of perceived respect and fairness correlate significantly with the personality trait Agreeableness: all the Pearson correlations are between 0.22 and 0.28, statistically significant at 1% level. Second, for the perceived distributive fairness, we found that income starts to have a statistically significant predictive power (p-value is 0.03) for the perceived distributive fairness after controlling for education level and job title. This supports the idea that distributive fairness should not be only about one's income, but also about how much income one thinks she deserves.

3.3.4. Do game decisions predict performance in the workplace?

Now we turn to the central research question of whether the game decisions predict performance in the workplace. To begin with, we check the group level differences in game decisions among the ordinary employees, junior managers and senior managers. Table 4 shows the differences in game decisions between the junior managers and the ordinary employees, and the differences between senior managers and the ordinary employees.

As the table shows, senior managers, compared to ordinary employees and junior managers, showed much more trust in others both in the trust game and the gift exchange game. This result is consistent with the finding in Fehr and List (2004), that is, the CEO participants displayed more trust than the student participants did in the trust games. We also found that senior managers rejected more often small offers in the ultimatum game; they were more trustworthy in the trust game; and they also cooperated much more in the conditional public goods game (only a bit more in the public goods game). We didn't find that senior managers performed differently from

others with respect to the other aspects of elicited social preferences. The junior managers didn't display any difference from the ordinary employees in any aspect.

Table 4. The group level differences between managers and ordinary employees

<i>Game Decisions</i>	<i>Junior Manager (Dummy)</i>	<i>Senior Manager (Dummy)</i>
<i>DG</i>	0.09	-0.53
<i>P.1</i>		
<i>UG</i>	-0.42	-0.45
<i>P.1</i>		
<i>UG</i>	0.89	2.00**
<i>P.2</i>		
<i>PGG</i>	1.01	2.48*
<i>CPGG</i>	0.91	3.56***
<i>TG</i>	0.13	5.50***
<i>P.1</i>		
<i>TG</i>	0.06	4.61**
<i>P.2</i>		
<i>GEG</i>	0.84	4.38***
<i>P.1</i>		
<i>GEG</i>	0.10	0.05
<i>P.2</i>		

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

We now explore whether the nine parameters that constructed from the game decisions predict the OCB scores. Table 5 below shows the pairwise Pearson correlation between the nine parameters and the OCB scores.

As the table shows, only 6 out of 36 correlations are statistically significant. The amount offered in the ultimatum game has some moderate correlation with the total OCB scores and scores on the three aspects of OCB. The correlation between it and Altruism (helping co-workers with job-related issues) is the only correlation coefficient that is statistically significant at 1% level. The minimum amount to accept in the ultimatum game has a correlation of 0.1 with Identification with the organization, and the averaged contingent contribution in the conditional PGG has a correlation of 0.1 with Altruism, both of which are marginally significant at 10% level. Some conceptual connection between social preferences elicited in the games and OCB scores are not

validated, including the connection between giving in the dictator game and Altruism, the link between cooperation in the public goods games and OCB, and the connection between effort exerted in the gift exchange game and OCB Conscientiousness.

Table 5. Correlations between the game decisions and the OCB scores (Pearson)

	<i>Identification</i>	<i>Altruism</i>	<i>Conscientiousness</i>	<i>OCB</i>
<i>DG</i>				
<i>P.1</i>	0.06	0.04	0.00	0.03
<i>UG</i>				
<i>P.1</i>	0.13**	0.17***	0.11*	0.14**
<i>UG</i>				
<i>P.2</i>	0.10*	0.07	0.00	0.06
<i>PGG</i>	0.03	0.03	-0.03	0.00
<i>CPGG</i>	0.08	0.10*	0.08	0.09
<i>TG</i>				
<i>P.1</i>	0.06	0.09	0.06	0.07
<i>TG</i>				
<i>P.2</i>	0.04	0.09	0.06	0.07
<i>GEG</i>				
<i>P.1</i>	0.03	0.03	0.03	0.03
<i>GEG</i>				
<i>P.2</i>	0.09	0.06	0.08	0.08
<i>N</i>	305	305	305	305

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

To better assess the external validity of social preference games, we also conducted some regression analyses, controlling for the effect of three factors: demographic variables, perceived respect and fairness, and personality traits. We included five

demographic variables: education level, age, gender, tenure with the hotel and income. The four aspects of perceived respect and fairness are also incorporated in the data analyses. For personality traits, we included the big-five: Agreeableness, Extroversion, Conscientiousness, Openness to experience, and Neuroticism. Table 6 summarizes results of the regression model, in which the total OCB scores was regressed on each of the nine game decisions, controlling for the effect of the three factors. We did the same analysis with scores on each of the three aspects of OCB. These results are showed in Table 7, 8, and 9.

Table 6. Regression analysis: the total OCB scores

<i>Independent Variables</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>
DG P.1	0.10								
UG P.1		0.32**							
UG P.2			0.11						
PGG				-0.02					
CPGG					0.09				
TG P.1						0.06			
TG P.2							0.04		
GEG P.1								0.02	
GEG P.2									0.83
Respect	0.04	0.07	0.05	0.04	0.05	0.04	0.05	0.05	0.04
Procedural Justice	-0.15	-0.16	-0.14	-0.14	-0.15	-0.16	-0.15	-0.14	-0.15
Distributive Justice	0.18	0.24	0.20	0.20	0.21	0.22	0.21	0.21	0.24
Incentives	-0.12	-0.13	-0.13	-0.14	-0.15	-0.13	-0.13	-0.15	-0.18
Education (Year)	.75***	.68***	.76***	.79***	.75***	.77***	.76***	.78***	.78***
Female	2.96**	3.21**	3.14**	2.91**	3.13**	3.10**	3.05**	3.04**	3.15**
Age	0.76	0.67	0.74	0.74	0.69	0.75	0.73	0.72	0.72
Age^2	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Tenure	0.02	-0.05	-0.00	0.02	0.01	-0.00	0.02	0.02	0.04
Income (RMB 1000)	0.59	0.74	0.52	0.55	0.54	0.52	0.54	0.54	0.54
Agreeableness	0.25	0.23	0.25	0.25	0.23	0.24	0.24	0.25	0.23
Conscientiousness	-0.06	-0.07	-0.06	-0.05	-0.04	-0.04	-0.05	-0.06	-0.03
Extraversion	-0.10	-0.10	-0.10	-0.12	-0.11	-0.11	-0.11	-0.11	-0.12
Openness	0.11	0.11	0.11	0.12	0.12	0.11	0.13	0.12	0.11
Neuroticism	0.13	0.11	0.12	0.12	0.13	0.13	0.13	0.12	0.12
Fixed Effect (Hotel)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	283	283	283	283	283	283	283	283	283

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

Table 7. Regression analysis: OCB Identification

<i>Independent Variables</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>
DG P.1	0.05								
UG P.1		0.09**							
UG P.2			0.05						
PGG				0.01					
CPGG					0.03				
TG P.1						0.01			
TG P.2							0.01		
GEG P.1								0.00	
GEG P.2									0.36
Respect	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Procedural Justice	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
Distributive Justice	0.10	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.12
Incentives	-0.07	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.08	-0.10
Education (Year)	0.19**	0.17**	0.20**	0.20**	0.19**	0.20**	0.20**	0.20**	0.20**
Female	0.69	0.77*	0.77*	0.71	0.75*	0.71	0.71	0.70	0.77*
Age	0.19	0.16	0.18	0.17	0.16	0.18	0.18	0.18	0.17
Age^2	-0.003	-0.002	-0.003	-0.002	-0.002	-0.003	-0.002	-0.002	-0.002
Tenure	0.02	0.00	0.01	0.02	0.02	0.02	0.02	0.02	0.03
Income (RMB 1000)	0.39**	0.42***	0.36**	0.37**	0.36**	0.36**	0.36**	0.37**	0.36**
Agreeableness	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05
Conscientiousness	-0.01	-0.00	-0.00	0.00	0.01	0.01	0.00	-0.00	0.01
Extraversion	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.07	-0.07
Openness	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Neuroticism	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Fixed Effect (Hotel)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	283	283	283	283	283	283	283	283	283

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

Table 8. Regression analyses: OCB Altruism

<i>Independent Variables</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>
DG P.1	0.03								
UG P.1		0.12***							
UG P.2			0.06						
PGG				-0.00					
CPGG					0.03				
TG P.1						0.02			
TG P.2							0.02		
GEG P.1								0.00	
GEG P.2									0.20
Respect	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01
Procedural Justice	-0.05	-0.06	-0.05	-0.05	-0.05	-0.06	-0.05	-0.05	-0.05
Distributive Justice	0.06	0.08	0.06	0.06	0.06	0.07	0.06	0.06	0.07
Incentives	-0.03	-0.03	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.04
Education (Year)	0.24***	0.22***	0.24***	0.25***	0.24***	0.25***	0.24***	0.25***	0.25***
Female	0.76*	0.86*	0.86*	0.76*	0.81*	0.82*	0.81*	0.78*	0.81*
Age	0.19	0.16	0.19	0.18	0.17	0.19	0.18	0.18	0.18
Age^2	-0.003	-0.002	-0.003	-0.003	-0.002	-0.003	-0.003	-0.003	-0.002
Tenure	-0.01	-0.03	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01
Income (RMB 1000)	0.09	0.15	0.06	0.08	0.07	0.07	0.07	0.07	0.08
Agreeableness	0.08*	0.07	0.09*	0.08	0.08	0.08	0.08	0.08	0.08
Conscientiousness	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	-0.02	-0.02	-0.01
Extraversion	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03
Openness	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01	0.00	-0.00	-0.00
Neuroticism	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Fixed Effect (Hotel)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	283	283	283	283	283	283	283	283	283

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

Table 9. Regression analyses: OCB Conscientiousness

<i>Independent Variables</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>	<i>Estimates</i>
DG P.1	0.02								
UG P.1		0.11*							
UG P.2			0.01						
PGG				-0.02					
CPGG					0.04				
TG P.1						0.03			
TG P.2							0.01		
GEG P.1								0.02	
GEG P.2									0.28
Respect	0.02	0.03	0.02	0.01	0.02	0.01	0.02	0.02	0.01
Procedural Justice	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.04
Distributive Justice	0.03	0.05	0.03	0.03	0.04	0.04	0.04	0.04	0.05
Incentives	-0.02	-0.02	-0.03	-0.02	-0.03	-0.02	-0.02	-0.03	-0.04
Education (Year)	0.32***	0.29**	0.32***	0.33***	0.31***	0.32***	0.32***	0.32***	0.32***
Female	1.50**	1.59**	1.51**	1.44**	1.57**	1.57**	1.53**	1.56**	1.56**
Age	0.38*	0.35	0.37	0.39*	0.36	0.38*	0.37	0.37	0.37
Age^2	-0.005	-0.004	-0.005	-0.005	-0.004	-0.005	-0.005	-0.005	-0.005
Tenure	0.01	-0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.01
Income (RMB 1000)	0.12	0.17	0.11	0.11	0.10	0.09	0.10	0.10	0.10
Agreeableness	0.11	0.10	0.11	0.11	0.10	0.10	0.10	0.10	0.10
Conscientiousness	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.04	-0.04	-0.03
Extraversion	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02	-0.03
Openness	0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.06	0.05
Neuroticism	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.05
Fixed Effect (Hotel)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	283	283	283	283	283	283	283	283	283

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

From these tables, we can see that the only game decision that predicts well organizational citizenship behavior is still the amount offered in the ultimatum game (the estimate is statistically significant at 5% level). This game decision predicts well OCB falling in all the three categories, with the strongest predictive power for helping coworkers in daily work (which is statistically significant at 1% level).

We found that several characteristics of the employees predict well the OCB scores. Among them, the strongest predictor of OCB is education level (statistically significant at 1%): those with a higher education level have a stronger identification with the organization (statistically significant at 5%); they are more helpful with coworkers (statistically significant at 1%); and they commit more to their job (statistically significant at 1%). Compared to male peers, female workers have higher scores on OCB (the difference is statistically significant at 5%), mainly in two aspects of OCB: Altruism and Conscientiousness. We also found that those employees that have a higher income (mainly the senior managers) developed a stronger identification with the organization (statistically significant at 5%).

In our data set, the perceived respect and fairness don't predict OCB well. This finding contrasts the results that reported in some earlier OCB studies, especially those that conducted with employees in the Western cultures (e.g., see Colquitt et al., 2001; Colquitt et al., 2013 for meta-analyses). But some studies on the OCB in the Eastern cultures suggest that cultural differences could weaken the typical links that found in the Western cultures. First, some studies showed that cultural differences affect the definition of OCB. For example, Lam et al. (1999) showed that employees in Eastern cultures understand some categories of OCB as part of their job duty. Moreover, Farh et al. (1997), which developed a OCB scale specifically for Chinese society (the scale used in this project), found a positive and significant correlation between distributive justice and the OCB, but this correlation was much weaker for those employees that endorse more Chinese traditional values, and disappeared for their female sample. Some studies (e.g., Chen et al., 2010; Cheng et al., 2003) also found that personal relationship with one's supervisor matters more for OCB than organizational attitudes in the Chinese context. These studies thus suggest that some characteristics of our data set may explain why the perceived respect and fairness not predicting the OCB scores, e.g., the

significant correlations between job position and OCB, about 70% of participants are female, and in general being in a good relationship with the supervisors.

The big-five personality traits in general have very small predictive power for the OCB scores. Only Agreeableness has some marginally significant correlation with helping co-workers with job-related issues.

3.4. Conclusions

This paper presents a systematic investigation of the external validity of social preference games by comparing game behavior with organizational behaviors in a group of hotels. We found that our social preference games capture some characteristics of the participants. For example, senior managers, compared to junior managers and ordinary employees, displayed more trust in others when playing the games. However, among the nine parameters that constructed from our game decisions, only the amount that offered in the ultimatum game has a robust and moderate correlation with organizational behavior. We didn't find strong connections between giving in the dictator game and helping behavior in the workplace, between cooperation in the public goods game and real work performance, and between reciprocity measured in the games and reciprocity that constructed from OCB. Though our project has no intention to establish clear theoretical mapping from one game decision to some specific organizational behaviors, our results suggest that social preferences games in general have a low predictive power for organizational behaviors.

As Levitt and List (2007) discussed, many systematic differences between the games and the real world could lead to the lab behavior not that representative of social behavior in the field. In our case, some of these differences are detected. First, the stakes in all of our games are (small) monetary payoff, but the cost of engaging in most of OCB is real effort. Second, the game decisions were made in a short time window, but the OCB scale measures the employees' behavior in a long time period.²⁶ Third,

²⁶ One related finding is provided by Gneezy and List (2006), which found that the time period of performing a task matters a lot for the profitability of the gift-exchange management practice.

there is no deep interaction between the players in the games, but repeated interaction with one's peer and supervisors is an indispensable part of most jobs.

A common belief is that behavioral patterns observed in social preference games can be well generalized to the real world. For example, the implications of positive reciprocity that observed in the lab for the labor market have been intensively discussed (e.g., Fehr et al., 1993; Fehr and Falk, 1999). Our results suggest that there exists a gap between game behavior and organizational behaviors: some quantitative results, even qualitative results, that generated by the lab decisions cannot be directly applied to the field. A potentially interesting avenue for the future research is to design lab games that have a better predictive power for organizational behaviors. Wang and Navarro-Martinez (2020) showed that it is possible to make contribution in the lab much more predictive in relation to charitable giving in the field by bringing the right contextual elements from the field into the lab. This suggests that incorporating contextual elements that matter for organizational behaviors into lab games may have the potential to bridge the gap.

Appendix

Appendix A. Organizational Citizenship Behavior Scale

Instructions. Here are a number of statements that may describe this employee very well or not. For each statement, please indicate the extent to which you agree with it (1=Strongly disagree, 5=Strongly agree).

This employee ...

1. Is willing to stand up to protect the reputation of the company;
2. Is eager to tell outsiders good news about the company and clarify their misunderstandings;
3. Makes constructive suggestions that can improve the operation of the company;
4. Actively attends company meetings;
5. Is willing to assist new colleagues to adjust to the work environment;
6. Is willing to help colleagues solve work-related problems;
7. Is willing to cover work assignments for colleagues when needed;
8. Is willing to coordinate and communicate with colleagues;
9. Complies with company rules and procedures even when nobody watches and no evidence can be traced;
10. Takes one's job seriously and rarely makes mistakes;
11. Tries hard to self-study to increase the quality of work outputs;
12. Does not mind taking on new or challenging assignments;
13. Often arrives early and starts to work immediately.

Appendix B. Self-reported perceived feelings questionnaire

Section 1. Respect

Please indicate to what extent that you agree with the following statements (1=Strongly disagree, 5=Strongly agree).

Would you agree or disagree that the manager of your work group

1. Values you as a member of your work group?
2. Respects your work-related ideas?

3. Values what you contribute at work?
4. Respects the work you do?
5. Appreciates your unique contributions on the job?
6. Approves of how you do your job?

Section 2. Procedural Justice

Please answer the following questions with respect to your current work group and the job you held within that work group at your company.

1. How would you rate the overall fairness with which issues and decisions that come up at work are handled? (1 =not fair at all, 5 =very fair);
2. Overall, how fair would you say decisions and processes are where you work? (1 =not fair at all, 5= very fair);
3. Is there a general sense among employees that things are handled in fair ways where you work? (1=not at all, 5 = definitely);
4. How much of an effort is made to be fair to employees when decisions are being made? (1= none, 5=a lot).

Section 3. Distributive Fairness

Please basing on your experience in the company, answer the three questions below.

1. How fairly are resources (e.g., salary, bonuses, etc.) allocated among employees where you work? (1= not at all fairly , 5=very fairly);
2. Overall, how fair is the salary you receive? (1=not fair at all, 5=very fair);
3. Would you say that there is an emphasis where you work on distributing things fairly? (1 = not at all, 5= definitely).

Section 4. Incentives

Please indicate the degree to which you agree with the following statements (1=strongly disagree, 5= strongly agree, except as noted):

1. The benefits I get are directly linked to the effort I put into my job;
2. If I do my job well, I will improve may pay and benefits a lot.

3. Rewards are distributed among employees according to the effort that they put into their jobs;
4. Employees are rewarded based on how well they do their jobs, not how well they maneuver through office politics;
5. Being good at one' s job is more important to one' s success than “looking” good;
6. Management is not impressed by employees who only try to make it seem like they are working hard.

Bibliography

Aleksandr, A., Charness, G., Gneezy, U., 2017. "Experimental methods: When and why contextual instructions are important." *Journal of Economic Behavior & Organization*, vol.134, 48–59.

Andrews, F. M., Whithey, S. B., 1976. "Social Indicators of Well-Being: Americans' Perceptions of Life Quality". *Plenum, New York*.

Andreoni, J., 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *Journal of Political Economy*, vol. 97, 1447-1458.

Andreoni, J. and Rao, J., 2011. "The power of asking: How communication affects selfishness, empathy, and altruism." *Journal of Public Economics*, vol.95, 513-520.

Andreoni, J. and Payne, A., 2013. "Charitable giving." *Handbook of Public Economics*. vol. 5, 1-50.

Andreoni, J., Rao, J., Trachtman, H., 2017. "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." *Journal of Political Economy*, vol. 125, 625-653.

Ariely, D., Loewenstein, G., and Prelec, D., 2003. "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences. *The Quarterly Journal of Economics*, Volume 118, Issue 1, 73–106.

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

Auten, G.E., Sieg, H., Clotfelter, C.T., 2002. "Charitable giving, income, and taxes: An analysis of panel data." *American Economic Review*, vol.92 , 371-82.

Bardsley, N., 2008. "Dictator game giving: altruism or artefact?" *Experimental Economics*, vol.11, 122--133.

Barrick, M. R., Mount, M. K., 1991. "The Big Five personality dimensions and job performance: A meta-analysis." *Personnel Psychology*, vol. 44, 1-26.

Benet-Martínez, V. and John, O. P., 1998. "Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English." *Journal of Personality and Social Psychology*, vol. 75, 729–750.

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*, July 10(1):122-42.

Blader, S.L., Tyler, T.R., 2009. "Testing and Extending the Group Engagement Model: Linkages Between Social Identity, Procedural Justice, Economic Outcomes, and Extrarole Behavior." *Journal of Applied Psychology*, Vol. 94, No. 2, 445-464.

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S., 2001. "Personality predictors of citizenship performance." *International Journal of Selection and Assessment*, vol. 9, 52-69.

Borman, W. C., & Penner, L. A., 2001. "Citizenship performance: Its nature, antecedents, and motives." In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace* (pp. 45-61). Washington, DC: American Psychological Association.

Branas-Garza, P., 2006. "Poverty in dictator games: Awakening solidarity." *Journal of Economic Behavior and Organization*, 60, 306-320.

Camerer, C., 2003. "Behavioral Game Theory: Experiments in Strategic Interaction." *Princeton University Press, Princeton*.

Camerer, C., 2011. "The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List." *Working paper, California Institute of Technology, Pasadena*.

Carpenter, J., Connolly, C. & Myers, C. K., 2008. "Altruistic behavior in a representative dictator experiment." *Experimental Economics*, vol.11, 282-98.

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

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

Charness, G., Gneezy, U., and Henderson, A., 2018. "Experimental methods: Measuring effort in economics experiments." *Journal of Economic Behavior and Organization*, 149, 74-87.

Chen, D.L., Schonger, M., Wickens, C., 2016. "oTree - An open-source platform for laboratory, online and field experiments." *Journal of Behavioral and Experimental Finance*, vol 9, 88-97.

Chen ZX, Tsui AS, Farh J. 2010. "Loyalty to supervisor vs. organizational commitment: relationships to employee performance in China." *Journal of Occupational and Organizational Psychology*, 75:339-56.

Cheng B, Jiang D, Riley JH., 2003. "Organizational commitment, supervisory commitment, and employee outcomes in the Chinese context: proximal hypothesis or global hypothesis?" *Journal of Organizational Behavior*, 24:313-34.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y., 2001. "Justice at the millennium: A meta-analytic review of 25 years of organizational justice research." *Journal of Applied Psychology*, 86(3), 425-445.

- Colquitt, J. A., Scott, B. A., Rodell, J. B., Long, D. M., Zapata, C. P., Conlon, D. E., & Wesson, M. J., 2013. "Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives." *Journal of Applied Psychology*, 98(2), 199–236.
- Dana, J., Cain, D.M., Dawes, R.M., 2006. "What you don't know won't hurt me: Costly (but quiet) exit in dictator games." *Organizational Behavior and Human Decision Processes*, Volume 100, Issue 2, 193-201.
- Davis, M. H., 1980. "A multidimensional approach to individual differences in empathy." *JSAS Catalog of Selected Documents in Psychology*, 10, 85.
- DellaVigna, S., List J.A., and Malmendier, U., 2012. "Testing for altruism and social pressure in charitable giving." *The Quarterly Journal of Economics* Vol. 127, 1–56.
- Dickinson, D., 1999. "An Experimental Examination of Labor Supply and Work Intensities." *Journal of Labor Economics* , Vol. 17, No. 4, 638-670.
- Dovidio, J. F., & Penner, L. A., 2001. "Helping and altruism." In M. Brewer & M. Hewstone (Eds.), *Blackwell international handbook of social psychology: Interpersonal processes*, 162–195. Cambridge, MA: Blackwell.
- Eisenberg, N., and Miller, P., 1987. "The relation of empathy to prosocial and related behaviors." *Psychological Bulletin*, vol.101, No. 1, 91-119.
- Ellingsen, T. and Johannesson, M., 2008. "Anticipated verbal feedback induces altruistic behavior." *Evolution and Human Behavior* 29, 100--105.
- Eckel, C.C. and Grossman, P.J., 1996. "Altruism in anonymous dictator games." *Games and Economic Behavior*, 16, 181-191.
- Farh, J.-L., Earley, P. C., & Lin, S.-C., 1997. "Impetus for action: A cultural analysis of justice and organizational citizenship behavior in Chinese society." *Administrative Science Quarterly*, 42(3), 421-444.
- Fehr, E., Falk, A., 1999. "Wage rigidity in a competitive incomplete contract market." *Journal of Political Economy*, 107, 106–134.
- Fehr, E., Kirchsteiger, G., Riedl, A., 1993. "Does fairness prevent market clearing? An experimental investigation." *Quarterly Journal of Economics*, 108 (2), 437–460.
- Fehr, E., List, J. A., 2004. "The hidden costs and returns of incentives—trust and trustworthiness among CEOs." *Journal of the European Economic Association*, 2(5):743–771.
- Fischbacher, U. , Gächter, S., Fehr E., 2001. "Are people conditionally cooperative? Evidence from a public goods experiment." *Economics Letters*, 71, 397–404.

Fischbacher, U. , Gächter, S., 2010. "Social preferences, beliefs, and the dynamics of free riding in public goods experiments." *American Economic Review*, Vol. 100, No. 1, pp. 541-556.

Fleeson, W., 2001. "Towards a structure- and process-integrated view of personality: Traits as density distributions of states." *Journal of Personality and Social Psychology*, 80, 1011-1027.

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, 83-87.

Franzen, A. and Pointner S., 2012. "Anonymity in the dictator game revisited." *Journal of Economic Behavior & Organization*, 81, 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.

Galizzi, M. and Navarro-Martinez, D., 2019. "On the external validity of social preference games: a systematic lab-field study." *Management Science*, 65, 976-1002.

Glazer, A. and Konrad, K.A., 1996. "A signaling explanation for the charity." *The American Economic Review*. Vol. 86, No. 4, 1019-1028.

Gneezy, U. and List J., 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica*, vol. 74, No. 5, 1365-1384.

"Giving USA: The Annual Report on Philanthropy for the Year 2018." *Chicago: Giving USA Foundation, 2019.*

Giluk, T. L., Postlethwaite, B. E., 2015. "Big Five personality and academic dishonesty: A meta-analytic review." *Personality and Individual Differences*, 72, 59 – 67.

Gosling, S., Rentfrow, P., Swann Jr., W., 2003. "A very brief measure of the Big-Five personality domains." *Journal of Research in Personality*, 37, 504–528.

Guth, W., Schmittberger, R., and Schwarze, B., 1982. "An experimental analysis of ultimatum bargaining." *Journal of Economic Behavior and Organization*, 3(2): 367.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., et al., 2006. "Costly punishment across human societies." *Science*, 312, 1767–1770.

Hoolwerf B., Schuyt T. (eds.), 2017. "Giving in Europe. The State of Research on Giving in 20 European Countries." *Lenthe Publishers, Amsterdam.*

John, O.P., Donahue, E. M., & Kentle, R. L., 1991. "The "Big Five" Inventory - Versions 4a and 54." *Berkeley: University of California, Berkeley, Institute of Personality and Social Research.*

Kahneman, D., Knetsch, J. and Thaler, R., 1986. "Fairness and the Assumptions of Economics." *The Journal of Business*, Vol. 59, No. 4, pp.S285-S300.

Kahneman, D., Krueger, A., Schkade, D., and Schwarz, N., Stone, A., 2004. "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method." *Science*, Vol. 306, 1776-1780.

Karlan, D. and List, J.A., 2007. "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment." *American Economic Review*, 97, 1774-1793.

Kahneman, D., Knetsch, J. and Thaler, R., 1986. "Fairness and the Assumptions of Economics." *The Journal of Business*, Vol. 59, No. 4, S285-S300.

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.

Konow, J., 2001. "Fair and Square: The Four Sides of Distributive Justice." *Journal of Economic Behavior and Organization* 46(2): 137-64.

Konow, J., 2010. "Mixed feelings: Theories of and evidence on giving." *Journal of Public Economics*, 94, 279-29.

Lam SSK, Hui C, Law KS. 1999. "Organizational citizenship behavior: comparing perspectives of supervisors and subordinates across four international samples." *Journal of Applied Psychology*, 84:594– 601.

Larson R., Csikszentmihalyi M., 2014. "The Experience Sampling Method." In: *Flow and the Foundations of Positive Psychology*. Springer, Dordrecht.

Ledyard, J., 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, ed. J. Kagel and A. Roth, chap. 2. NJ: Princeton University Press.

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

Levitt, S., and List, J.A., 2017. "What do laboratory experiments measuring social preferences reveal about the real world." *Journal of Economic Perspectives*, 21, 153-174.

List, J.A., 2006. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions." *Journal of Political Economy*, Vol. 114, No. 1, 1-37.

List, J.A., 2007. "On the interpretation of giving in dictator games." *Journal of Political Economy*, vol. 115, No.3, 482-493.

- List, J.A., 2011. "The Market for Charitable Giving." *Journal of Economic Perspectives*, Vol. 25, No. 2, 157-180.
- Loewenstein, G., 1999. "Experimental economics from the vantage-point of behavioural economics." *Economics Journal*, 109, 453, 25–34.
- Mischel, W., 1968. "Personality and Assessment." *New York; Wiley*.
- Okten, C. and Weisbrod, B., 2000. "Determinants of donations in private nonprofit markets." *Journal of Public Economics*. Volume 75, Issue 2, Pages 255-272.
- Organ, D. W., 1988. "Organizational citizenship behavior: The good soldier syndrome." *Lexington, MA: Lexington Books*.
- Organ, D. W., & Ryan, K., 1995. "A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior." *Personnel Psychology*, 48, 775-802.
- Pardo, D., Cotrina, J., 2016. "Validation of the Self-report altruism scale test in Colombian University Students." *Ánfora*, 23(41), 17- 35.
- Paunonen, S., 2003. "Big Five Factors of Personality and Replicated Predictions of Behavior." *Journal of Personality and Social Psychology*, Vol. 84, No. 2, 411–424.
- Penner, L.A., Fritzsche, B.A., Craiger, J.P., & Freifeld, T.S., 1995. "Measuring the prosocial personality." In *J.N. Butcher & C.D. Spielberger (Eds.), Advances in Personality Assessment*, Vol. 10 (pp. 147-163). Hillsdale, NJ: Erlbaum.
- Penner, L., 2002. "Dispositional and Organizational Influences on Sustained Volunteerism: An Interactionist Perspective." *Journal of Social Issues*, Vol. 58, No. 3, 447-467.
- Pérez-Albéniz, A., Paúl, J., Etxeberria, J. Montes, M.P. y Torr, E. 2003. "Adaptación del Interpersonal Reactivity Index (IRI) al español." *Psicothema*, 15, 267-272.
- Rioux, S.M., Penner L.A., 2001. "The causes of organizational citizenship behavior: a motivational analysis." *Journal of Applied Psychology*, vol. 86, No. 6, 1306-1314.
- Ross, L. and Nisbett, R.E., 1991. "The Person and the Situation: Perspectives of Social Psychology." *New York: McGraw-Hill*.
- Slovic, P., 1995. "The construction of preference." *American psychologist*, 50(5), 364-371.
- 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 Soest, D., 2012. "From the Lab to the Field: Cooperation among Fishermen." *Journal of Political Economy*, Vol. 120, No. 6, 1027-1056.

Stoop, J., 2014. From the lab to the field: Envelopes, dictators and manners. *Experimental Economics, Volume 17, Issue 2, 304–313.*

Wang, X., Navarro-Martinez, D., 2020. “Bridging the gap between the lab and the field: Dictator games and donations.” Working paper.

Zhang, Y., Guo, Y., Newman A., 2017. “Identity judgements, work engagement and organizational citizenship behavior: The mediating effects based on group engagement model.” *Tourism Management, vol. 61, pp 190-197.*

Zhao, H., Wu, J., Sun, J.-M., & Chen, C. W., 2012. “Organizational citizenship behavior in Chinese society: A reexamination.” *The International Journal of Human Resource Management, 23(19), 4145-4165.*