Social preferences in the online laboratory: a randomized experiment

Jérôme Hergueux, Nicolas Jacquemet

Social preferences in the online laboratory: a randomized experiment*

Jérôme Hergueux† Nicolas Jacquemet‡

April 2014

Abstract

Internet is a very attractive technology for the implementation of experiments, both in order to obtain larger and more diverse samples and as a field of economic research in its own right. This paper reports on an experiment performed both online and in the laboratory, designed to strengthen the internal validity of decisions elicited over the Internet. We use the same subject pool, the same monetary stakes and the same decision interface, and control the assignment of subjects between the Internet and a traditional university laboratory. We apply the comparison to the elicitation of social preferences in a public good game, a dictator game, an ultimatum bargaining game and a trust game, coupled with an elicitation of risk aversion. This comparison concludes in favor of the reliability of behaviors elicited through the Internet. We moreover find a strong overall parallelism in the preferences elicited in the two settings. The paper also reports some quantitative differences in the point estimates, which always go in the direction of more other-regarding decisions from online subjects. This observation challenges either the predictions of social distance theory or the generally assumed increased social distance in internet interactions.

JEL classification: C90, C93, C70

Keywords: Social Experiment, Field Experiment, Internet, Methodology, Randomized Assignment

* This paper is a revised and augmented version of CES Working Paper n° 2012-70. We are grateful to Anne l'Hôte, Andrews-Junior Kimbembe and Ivan Ouss for their outstanding research assistance, as well as Maxim Frolov and Joyce Sultan for their help in running the laboratory experimental sessions. We are especially indebted to Yann Algan for his help during the development of this project. We gratefully thank the editor, Jacob Goeree, two anonymous referees and Guillaume Fréchette, Olivier L'haridon, Stéphane Luchini, David Margolis, Ken Boum My, Paul Pézanis-Christou, Dave Rand, Al Roth, Antoine Terracol, Laurent Weill and the members of the Berkman Cooperation group for helpful remarks and discussions. We also thank seminar participants at the Berkman Center for Internet & Society at Harvard and the 2012 North American Economic Science Association conference for their comments. We gratefully acknowledge financial support from the European Research Council (ERC Starting Grant). Jacquemet acknowledges the *Institut Universitaire de France*.

† University of Strasbourg, Institute of Political Studies and Sciences Po, Department of Economics. Research fellow, Berkman Center for Internet & Society at Harvard University. 23 Everett Street, 2nd floor, Cambridge, MA, 02138, USA. e-mail: jhergueux@cyber.law.harvard.edu

‡ Université de Lorraine (BETA) and Paris School of Economics. 3 Place Carnot, 54035 Nancy. e-mail: Nicolas.Jacquemet@univ-lorraine.fr

1 Introduction

In the field of experimental economics, it is a long time since researchers called for the development of the "online laboratory" (Bainbridge 2007). The interest in online experimentation has been propelled by the possibility of reaching more diverse samples, recruiting larger subject pools and conducting cross-cultural social experiments in real time at an affordable cost.[1] Besides this methodological concern, the Internet is becoming an increasingly prominent experimental field for social science research in its own right (see, e.g., Resnick et al. 2006; Chesney et al. 2009), as we live more and more of our social and economic lives online. It is thus essential to conduct experiments directly over the Internet if we are to rely on the experimental method to understand the various types of social and economic activities that people engage in online.

Notwithstanding these appealing features, the development of the "online laboratory" still remains in its infancy. The primary goal of this paper is to help fill this gap by conducting a methodological evaluation of an Internet-based experimentation procedure. Horton et al. (2011) underline the difficulty of coming up with procedures for online experiments that ensure their internal validity, *i.e.* the possibility of confidently drawing causal inferences from one's experimental design. A number of confounding factors have been identified that have probably prevented researchers from running experiments online: (i) it is difficult to monitor the identity of subjects participating in the experiment, (ii) subjects may read the experimental instructions too carelessly and/or make decisions too quickly and/or get significantly distracted during the course of the experiment, (iii) subjects may selectively drop out of the experiment in ways that the experimenter does not understand, (iv) subjects may not believe that they interact with other human players and/or that they are going to be paid at the end of the experiment as described in the instructions, and finally (v) the issue of reliably and automatically processing the payment of subjects over the Internet in an anonymous fashion appeared to be a major blocker.

In this paper, we seek to compare the behavioral results generated both in a traditional laboratory and over the Internet. To do so, we develop an online platform specifically dedicated to conducting

[1] In a recent paper, Henrich et al. (2010) warned against behavioral scientists' current over-reliance on data overwhelmingly gathered from populations of Western undergraduate students and recommended a major effort to broaden the sample base. The Internet is a promising medium for conducting experiments with large and diverse samples. It is now possible to reach 78.3% of the North American population through the Internet, and while only 11.4% of the African population can currently be reached through this method, the exponential growth of its user base (from 4 million users in 2000 to 118 million users in 2011) could soon make it an attractive tool for conducting experiments in the developing world as well (*source*:www.Internetworldstats.com).

Table 1. In-lab versus online based experiment: overview of experimental results

Paper	Type of experiment	Subject pool	Random allocation of subjects	Main results
Anderhub et al. (2001)	Individual level consumption/saving decisions	47 in lab 50 online	NO	(i) similar economic behavior on average (ii) higher behavioral variance online (iii) shorter decision times online
Shavit et al. (2001)	Individual lotteries evaluation decisions	65 in classroom 70 online	NO	(i) lower risk aversion online (ii) higher behavioral variance online
Charness et al. (2007)	Lost wallet game	178 in classroom 124 online	NO	Very little difference in average economic behavior
Fiedler and Haruvy (2009)	Trust game with pre-play communication	136 in lab 216 online	*NA*	Lower levels of trust and trustworthiness online
Chesney et al. (2009)	Dictator game, Ultimatum game, Public Good game, Minimum Effort game, Guessing game	Respectively 30, 64, 32, 31 and 31 online	*NA*	Behavioral results qualitatively in line with previous laboratory based experiments
Horton et al. (2011)	Watershed experiment, Religiously primed and unprimed versions of the Prisonner's Dilemma	Respectively 213, 189 and 113 online	*NA*	Behavioral results qualitatively in line with previous laboratory based experiments
Amir et al. (2012)	Public Good game, Dictator game, Ultimatum game, Trust game	189 per game online	*NA*	Behavioral results qualitatively in line with previous laboratory based experiments

social experiments over the Internet that is usable as in the laboratory. To account for the effect of self-selection between implementations, we control the allocation of subjects between treatments.

The platform provides controls over many of the above-mentioned confounding factors. In particular we *(i)* control for differences in response times, *(ii)* deal with the issues of selective attrition, concentration and distraction and *(iii)* provide as much control as possible over subjects' beliefs as regards the experimental instructions.

The existing literature has already covered a variety of different games implemented over the Internet (Table 1 summarizes the methodology and main conclusions of this literature). The seminal study of Anderhub et al. (2001) focuses on an individual level decision experiment under uncertainty, both in the laboratory and online. Shavit et al. (2001) compare student bids over buying prices for simple lotteries both in the classroom and online. Charness et al. (2007) also compare classroom experiments with other Internet-based experimental settings to investigate the effect of social distance on trust and reciprocity in a simple lost wallet game. They find that trust and reciprocity both decrease in an Internet-based setting, which they argue is consistent with social distance theory (Akerlof 1997). Fiedler and Haruvy (2009) and Chesney et al. (2009) take an exploratory approach and build a virtual laboratory on the *Second Life* website. Chesney et al. (2009) recruit subjects from the *Second Life* community to perform a series of social experiments and compare the results with those of

the traditional laboratory literature. Similarly, Fiedler and Haruvy (2009) recruit subjects from *Second Life* to perform a Trust game, but directly compare their results with those obtained from traditional laboratory subjects playing in the same virtual environment, but in a physical laboratory. They also find trust and trustworthiness to be lower outside the physical lab. Most recently, Horton et al. (2011) and Amir et al. (2012) have used the online labor market platform *Amazon Mechanical Turk* to conduct a set of classic experiments and replicate qualitatively some general results drawn from the experimental economics literature.

We contribute to this burgeoning literature by looking at social preferences and by providing a rigorous comparison of the Internet-based experimentation with traditional lab experiments. We apply our methodology to the measurement of social preferences – combined with a risk aversion task – through a Public Good game, a Trust game, a Dictator game and an Ultimatum game (using a within-subjects design). The main conclusions that we draw from this comparison are twofold. First, the social preferences elicited in the lab and online are qualitatively very similar – all common inferences on social preferences that we replicate in the laboratory would also be obtained based on online data. Second, we do, however, observe some differences in the point estimates between treatments. Social distance theory (Akerlof 1997) predicts that the stronger anonymity that prevails in Internet-based interactions should drive social preferences down as compared with the laboratory setting, where people can (i) see each other before and after the experiment, (ii) recognize that they often come from the same socio-economic background and (iii) know that they are going to be matched with one another during the experiment. On the contrary, we find robust and significant evidence that subjects allocated to the Internet treatment behave more altruistically and, when insignificant, the differences in social preferences always go in the direction of more other-regarding decisions online. We suggest an explanation for our results based on the nature of the social and economic interactions in which individuals tend to engage online, which they are likely to bring to the experiment through its contextual implementation.

Our results are important to the community of researchers wishing to develop the online laboratory as a medium for running social experiments over the Internet and to relate their results to the established laboratory literature. They are also important for social scientists wishing to use social experiments to research the Internet as a field: given the observed parallelism between fields, it makes sense for researchers to bring their experimental tools directly to the field, *i.e.* over the Internet, if they want to learn from subjects' behavior in this context, rather than sticking with the more difficult approach of trying to bring a subsample of those subjects into a traditional university laboratory.

The rest of the paper proceeds as follows. Section 2 documents the design of the experiment, reports on the development of our online experimental economics platform and explains our

experimental procedures. Section 3 reports the main results of the experiment. We then move to additional evidence on the reliability of the comparison based on an analysis of the internal validity of the online experiment, secondary outcomes and robustness treatments. We discuss the main outcomes of this comparison in Section 4, and conclude in Section 5.

2 Design of the experiment

Social isolation and greater anonymity are well-recognized distinctive features of online interactions. In order to provide a rather conservative testbed comparison between online and lab experiments, we focus on the elicitation of social preferences. Shavit et al. (2001) have also shown that subjects tend to be less risk-averse when making decisions online rather than in a classroom. We thus complement our preference measures with a risk aversion task. Our main methodological contribution is to build an Internet-based experimental environment which can be implemented both online and in the laboratory. We conclude this section with a detailed description of the procedures and decision interface we used.

2. 1 The decision problems

At the beginning of the experiment, each subject is attributed a role: either participant A or participant B. The assigned role remains the same during the whole experiment. The experiment is divided into two different parts. First, we elicit decisions in five different games. The first four games are taken from the social preferences literature (see, e.g., Fehr & Camerer 2004) while the last one elicits individual risk aversion. At the end of each game, subjects are asked to answer non-incentivized questions about their beliefs and intentions in the game they have just played. In the second part of the experiment, subjects are asked to answer some standard demographic and social preference-related questions, along with some questions eliciting their beliefs about the study.

Public Good Game. Subjects play in groups of four with an initial endowment of 10€ per player. Each euro invested in the common project by a member of the group yields a return of 0.4 euro to each group member. Following Fischbacher et al. (2001), we elicit both unconditional and conditional contributions, asking subjects to make two contribution decisions in turn. They first decide on how much of their 10€ they want to invest in the common project. They then provide their intended contribution for each possible value (on the scale of integers from 0 to 10) of the average contribution

of the three other members.[2] One of the two decisions is randomly drawn to be binding and determines the individual earnings for this game according to the following payoff function:

$$\pi_i = 10 - \text{contrib}_i + 0{,}4 \sum_{j=1}^{4} \text{contrib}_j \quad (1)$$

Right after the decision screen, we ask subjects about (i) their normative opinions about how much people *should* contribute to the Public Good (ii) whether they had an idea about how much the other members of their group would contribute to the Public Good when they made their decision, and if so (iii) their beliefs about how much the other members of their group actually contributed on average.

Dictator Game. Each participant A is matched with a participant B and plays the role of dictator. The dictator receives a 10€ endowment, of which he must decide how much is transferred to participant B. The difference is participant A's earning for this game.

Ultimatum Bargaining Game. Each participant A is matched with a participant B. Participant A is the proposer and must decide on how much of an initial endowment of 10€ is transferred to participant B – the responder. The responder is simultaneously asked for the threshold level of transfer below which the offer will be refused. The earnings of each player in this game are computed according to the proposal if participant A's transfer is higher or equal to the threshold. Otherwise, both players' earnings are set equal to 0.

Trust Game. Each participant A is matched with a participant B, and both players receive a 10€ initial endowment. Participant A is the trustor and chooses how much of his endowment is transferred to participant B – the trustee. The trustee receives three times the amount sent by the trustor, and chooses how much is sent back to the trustor. We elicit this decision through the strategy method: for each possible transfer from the trustor (from 1 to 10) the trustee chooses how much will be returned without knowing the trustor's actual choice. Right after the decision screen, we ask trustors about (i)

[2] The second decision is a variant of the "strategy method" (Selten 1967), introduced by Fischbacher et al. (2001) to elicit conditional cooperation. As in the original strategy method, subjects are asked decisions for each possible state of the world, but these states are reduced to average contributions of other subjects instead of all possible combinations of their individual decisions. In order to give subjects a monetary incentive to take both decisions seriously, we applied the same compensation rule as in Fischbacher et al. (2001): for one randomly chosen subject, the table of unconditional decisions is binding; for the other three the relevant decisions are the unconditional ones. These realizations of the draw are the monetary outcomes of this stage for each subject.

whether they had an idea about how much the trustee would return to them when they made their decisions, and if so (ii) their beliefs about the amount that the trustee would return.

Risk aversion elicitation. Each participant faces a menu of ten choices between lottery pairs, adapted from Holt & Laury (2002). The probability of getting the higher amount is always the same between the two lottery pairs, but the safe option pays either 20€ or 16€ while the risky option pays either 38.5€ or 1€. The probability that subjects get the higher amount in both options steadily increases from 10% in the first decision problem to 100% in the last one. Thus, in decision 10, subjects actually choose to earn either 20€ or 38.5€ with certainty. One of the ten decisions is randomly drawn to determine the binding lottery choice. Earnings for this game are then derived from a random draw according to the probability of the corresponding lottery.

Social values survey. After all games have been played, subjects are asked to fill in a questionnaire with some standard demographic questions followed by social preference-related questions. This set of questions has been taken from the *World Values Survey* (WVS), the *General Social Survey* (GSS) and the *German Socio-Economic Panel* (GSEP) – the three most commonly used sources in the empirical literature. Specifically, we ask subjects:

(i) to what extent they consider it justifiable to free-ride on state benefits (cooperation variable; WVS question);

(ii) whether they think that people are mostly looking out for themselves as opposed to trying to help each other (altruism variable; WVS question);

(iii) whether they think that people would try to take advantage of them if they got a chance as opposed to trying to be fair (fairness variable; WVS question);

(vi) whether they think that most people can be trusted or that one needs to be very careful when dealing with people (trust variable; WVS and GSS question);

(v) how trusting they generally are of people (trust variable; GSEP question);

(vi) how trusting they are of people they have just met (trust variable; GSEP question);

(vii) whether they generally see themselves as fully prepared to take risks or as trying to avoid them (a question taken from Dohmen et al. 2011). All questions are mandatory and none is remunerated.

Debriefing questionnaire. As demonstrated by Eckel and Wilson (2006), the internal validity of online experiments can be challenged by subjects' skepticism about whether they actually interact with other human subjects and whether they will actually be paid according to the rules described in the

instructions. To get some control over these dimensions, we ask subjects to rate their level of confidence in those two critical features of the study. As a complement, we end the survey by asking subjects to report on how carefully they read the experimental instructions, on how calm their environment was when they performed the experiment and on whether they had participated in any similar studies in the past.

2. 2 Procedures common to both implementations

All five games, followed by the survey, are played successively in each experimental session. As we seek to elicit social preferences in isolation from learning effects and strategic concerns, each game is only played once. To neutralize reputation effects, we match subjects in each game according to a perfect stranger procedure. Last, in order to further break any possible correlation between games, only one game out of the whole session is randomly drawn as binding to compute each subject's earnings. Final payoffs equal the earnings from the corresponding decision plus a 5€ show-up fee. Subjects are only informed of their earnings in each game at the very end of the experiment.

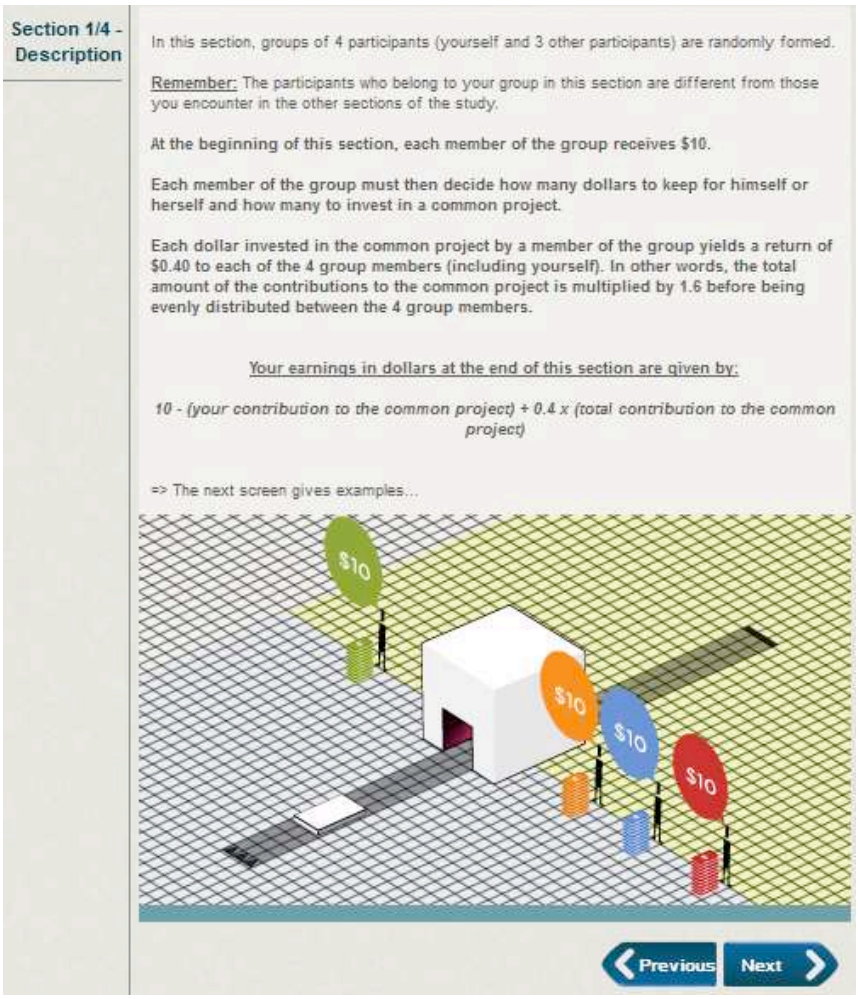
As all games are played one after the other, order effects could influence the preferences we elicit. This led us to implement three different orderings. The Public Good game is the most cognitively demanding, so we start all sessions with this game. The Dictator, Ultimatum and Trust games all appear afterwards in varying orders. As we mainly use the risk aversion task for purposes of replication and as a control variable, we maintain this decision problem as the last in all sequences.

- Order 1: *Public Good – Dictator – Ultimatum – Trust – Risk Aversion*
- Order 2: *Public Good – Trust – Ultimatum – Dictator – Risk Aversion*
- Order 3: *Public Good – Ultimatum – Dictator – Trust – Risk Aversion*

Subjects face the exact same decision interface both in the lab and online. The online implementation of the experiment requires a fully self-contained interface, so that every communication between the subjects and the experimenter has to proceed through the screen.[3] The first screen of the decision interface provides subjects with general information about the experiment, including the number of sections and how their earnings will be computed. Each game is then performed in turn, following a given sequence of screens.

[3] The interface has been developed under *Lime Survey* (http://www.limesurvey.org/), a highly customizable open-source survey tool.

Figure 1. The description screen of the Trust game

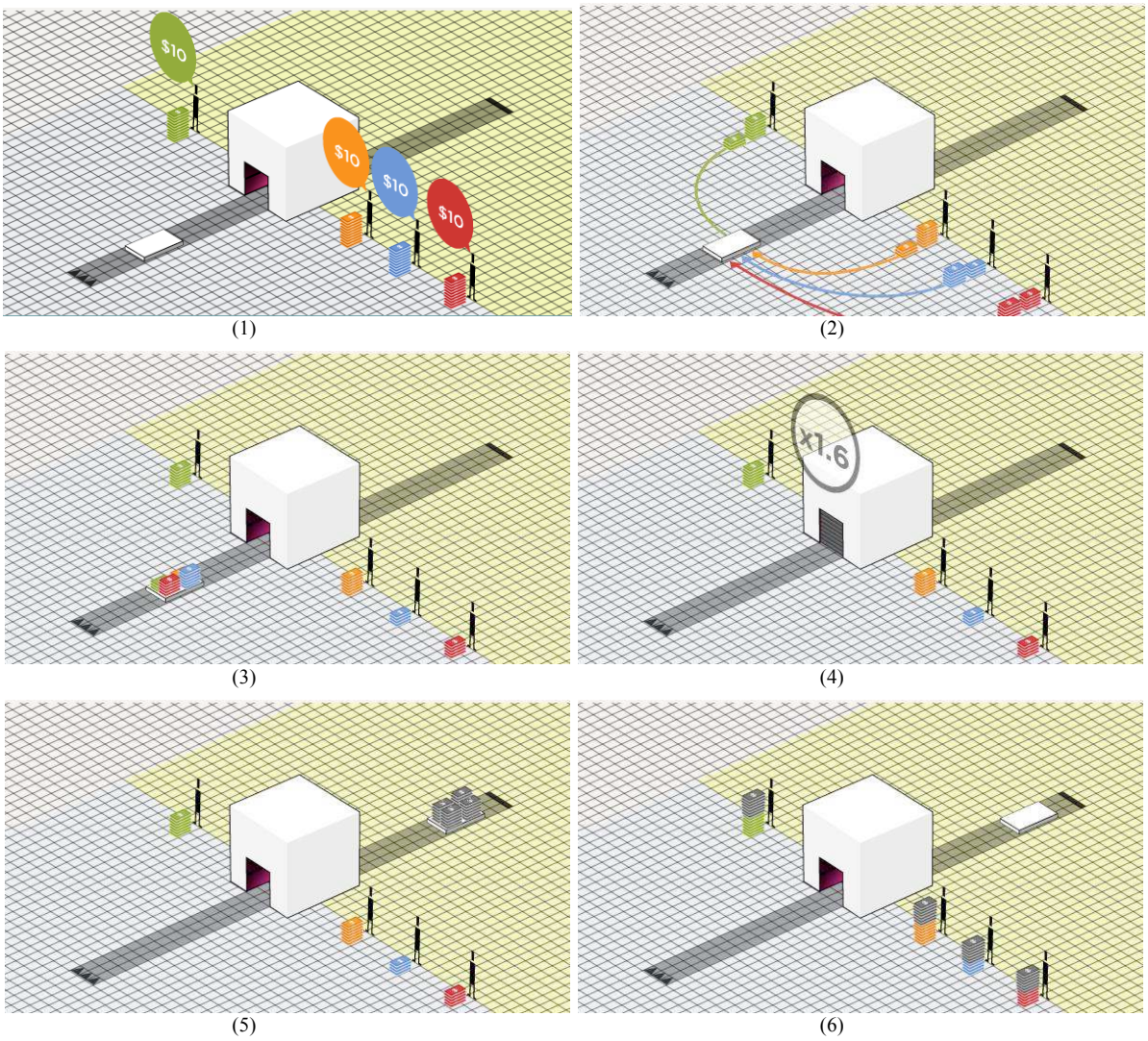


The first screen of each section describes the instructions for the game that subjects are about to play (Figure 1 provides an English translation of the original instructions in French for the Trust game).

One important methodological concern with online experiments is to guarantee an appropriate understanding of the decision problems when no interaction with the experimenter is possible, which makes it difficult, for instance, to rely on the standard post-instructions questionnaire coupled with oral questions. We address this issue through several distinctive features of the interface. First, we include suggestive flash animations illustrating the written experimental instructions at the bottom of each instruction screen (the animation appears at the bottom of the first screen, as shown in Figure 1; the animation is illustrated in Figure 2 by step-by-step screen captures).

Displaying a purely random sequence of flash animations would introduce uncontrolled and subject specific noise – through, e.g., anchoring on a particular behavior or sequence of events.

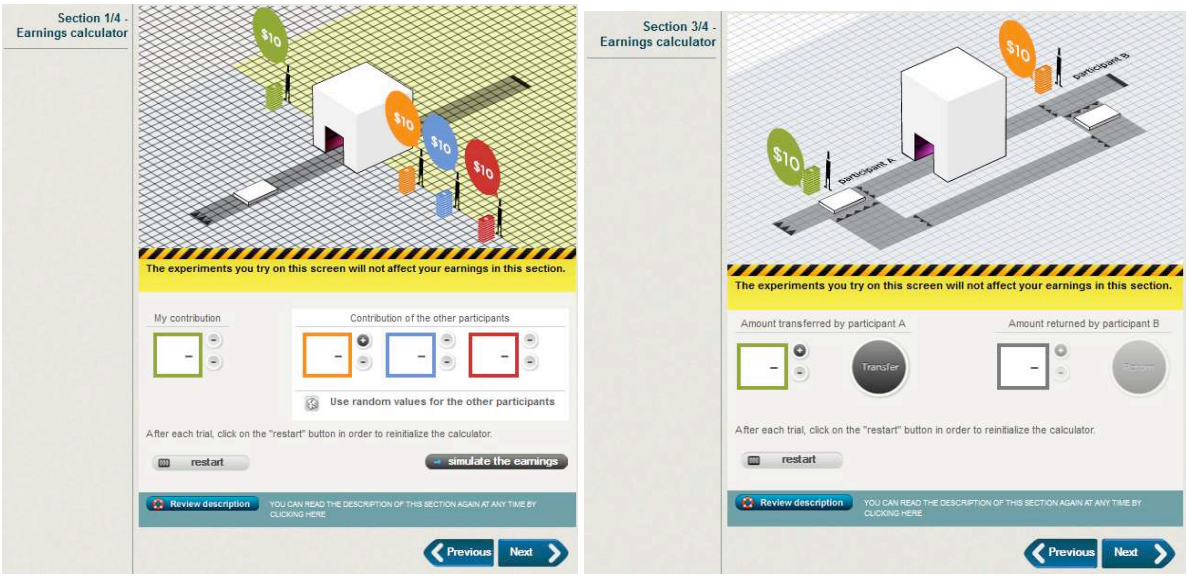
Figure 2. Flash animation for the Public Good game



Since our main objective is to compare behavior between the two implementations, we get rid of this noise by fixing the actual sequence: the loop of concrete examples displayed in the animations is first randomly determined and then fixed for each game. The same loop is displayed to all subjects without any other numeric information than the subjects' initial endowments.

Second, the instruction screens are followed by a screen providing some examples of decisions, along with a detailed calculation of the resulting payoffs for each player. These examples are supplemented on the subsequent screen by earnings calculators. On this interactive page, subjects are allowed to test all the hypothetical scenarios they are interested in before making their decisions in the Public Good and Trust games (English translations of the original earnings calculators in French are provided in Figure 3, (a) for the Public Good game and (b) for the trust game). In contrast to the flash animations, the numeric results of each scenario run by a subject in the earnings calculator screens are explicitly displayed.

Figure 3. Earnings calculators



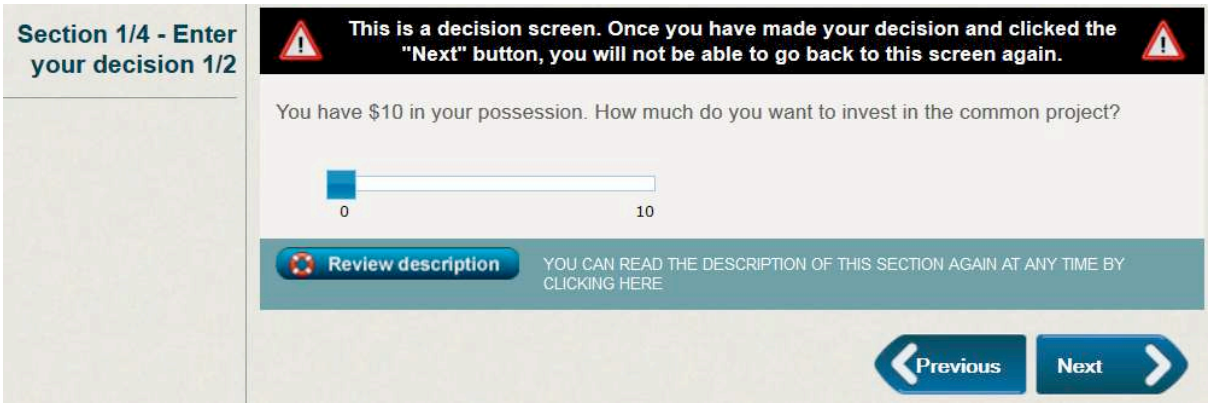
(a) Public Good game (b) Trust game

Last, the system provides quick access to the instructions material at any moment during decision-making. On all screens, including decision-making ones, a "review description" button gives subjects direct access to the instructions displayed at the beginning of the game. The system also allows participants to navigate at will from one screen to another – until a decision screen has been passed – through the "Previous" and "Next" buttons located at the bottom of each screen (Figure 4 provides an English translation of the original decision screen in French for the Public Good game).

A potentially important confound when comparing laboratory and online experiments is the average variation in decision times. Anderhub et al. (2001) report that subjects make decisions more quickly in an online environment. However, an established body of research in psychology indicates that shorter decision times are likely to be associated with *instinctive and emotional reasoning processes* rather than *cognitive and rational* ones (Kahneman 2003), which could cause subjects to make more pro-social decisions on average. In order to generate a control variable for this dimension, the platform recorded detailed data on the time in seconds that subjects spent on each screen of the interface (this timer was not visible to the subjects). But more time on a screen does not necessarily mean longer decision time if, for instance, online subjects leave their computer while answering the survey.

To get further information about whether some subjects were likely to have been distracted from the online experiment at some point, we included an indicator of mouse inactivity in the platform.

Figure 4. Decision screen for the Public Good game



The indicator records both the screen and the duration of inactivity each time the mouse of the subject is inactive for more than 5 minutes.[4]

2. 3 Practical implementation of the experiment

All participants in the experiment were contacted through the subjects database of the experimental economics laboratory of University Paris 1 Panthéon-Sorbonne.[5] The allocation to sessions is intended to minimize differences in the subject pools and avoid self-selection into treatments. We apply a matching procedure that proceeds in two steps. First, subjects are invited to register for a date on which a session takes place. They are told that practical details about the experiment will follow once their registration has been confirmed (as usual, registrations are confirmed on a first-come first-served basis). Indeed, two sessions are scheduled during each time slot: one session online and one session in the laboratory. In the second step, we sequentially allocate subjects either to the laboratory or to the online experimentation according to their registration order.

As the capacity of the laboratory allows for no more than 20 subjects, we allowed 56 persons to register for each time slot, allocating half of them to the laboratory and the other half to the Internet session. In the laboratory, we had to refuse any overbooked subjects who showed up on time. Since no such constraint applied to the online experiment, we allowed all subjects to participate while keeping track of those who logged-in after the target number of 20 participants had been reached. In laboratory sessions, subjects are randomly assigned to a computer upon arrival. The instructions for the experiment are read aloud, and subjects are then left to use all devices at their disposal to check their own understanding (access to the text, earnings calculators, etc.). Each game is described in turn,

[4] The system considered the mouse inactive when it was moving over screens not belonging to the experimental economics platform.
[5] The database is managed using Orsee (Greiner 2004).

following the above-described interface, so that all subjects progress inside the experiment at the same time.

Online subjects are invited to visit the url embedded in their confirmation e-mail at the time their session is scheduled, and to log into the system using their e-mail address, which served as a unique login token. The url was activated during the half-day spanning the time scheduled for the experiment. The computer program allocates online subjects to either participant A or participant B according to their login order (in order to ensure that we get a somewhat equal split of the subject pool between participant As and participant Bs, despite possible dropouts).

At the end of the experiment, subjects are matched using a perfect stranger procedure. Subjects are informed of their earnings in each game only at the end of the experiment. In the laboratory, subjects from a given session are matched together. By contrast, online subjects had their decisions matched with the decision records of subjects who had already completed the experiment.[6] This feature of the platform allowed Internet subjects to perform the experiment independently and at their own pace, thus smoothing the interactions and arguably reducing dropouts.[7] The drawback of this matching procedure is that it breaks the joint determination of payoffs between subjects: when a subject makes a decision, his own payoff is determined by the decision made by some previous participant, while his current decision determines the payoff of another, future participant. Such a sequential matching between current and past decisions can hardly be avoided in online experiments, in which subjects must be allowed to participate at any time they see fit. An alternative way of implementing the online matching, introduced by Cooper and Saral (2013), would have been to compute both subjects' outcomes at a later time, once the second subject has gone through the experiment – thus restoring the joint determination of payoffs inside each pair. We opted for the first solution for two reasons. First, having subjects wait until a future date before they can get their earnings involves inter-temporal preferences and may induce further differences in the saliency of payoffs between the two environments. Second, we were also concerned that the credibility of the experiment would be challenged for online subjects, if they were not informed about their experimental earnings immediately after their participation. Both solutions have advantages and drawbacks, and a more systematic comparison of the consequences of each design is worth investigating in future works.[8]

Laboratory subjects' earnings are paid in cash before subjects leave the laboratory. Internet subjects get paid through an automated PayPal transfer. This guarantees a fungibility similar to that of cash transfers, as money transferred via PayPal can be readily used for online purchases or easily

[6] Since we apply a sequential matching rule for online subjects, the queue has to be initialized somewhere. We used data from 3 pilot sessions in the laboratory run during summer 2010 in preparation for the current study.

[7] Overall, 208 subjects logged in to the platform to participate in the online experiment, of whom 6 dropped out before completion.

[8] Our robustness treatments, presented in Section 4.3, provide some preliminary insights on this issue.

Table 2.1. Summary of the design: common procedures between treatments

Decision problems	Decisions elicited from participant		Games ordering			Sequence of screens
	A	B	1	2	3	
1. PUBLIC GOOD GAME *Elicitation of beliefs*	(i) unconditional contribution (ii) conditional contribution (strategy method) (i) normative view on how much people should contribute (ii) idea about contributions of others at time of decision (iii) estimation of contributions of others at time of decision		1	1	1	- Description (text + animation) - Illustrative examples - Earnings calculator - Decision screen unconditional - Decision screen conditional - Beliefs elicitation
2. DICTATOR GAME	Transfer	*None*	2	4	3	- Description (text + animation) - Decision screen
3. ULTIMATUM GAME	Transfer	Minimum acceptable offer	3	3	2	- Description (text + animation) - Decision screen
4. TRUST GAME *Elicitation of beliefs*	Transfer (i) idea about return at time of decision (ii) estimation of return at time of decision	Amount returned (strategy method) *None*	4	2	4	- Description (text + animation) - Illustrative examples - Earnings calculator - Decision screen - Beliefs elicitation
5. HOLT & LAURY LOTTERIES	Choice over 10 lottery pairs		5	5	5	- Description (text + illustrative table) - Decision screen
Social values survey	Cooperation, altruism, fairness, trust (WVS), general trust, trust in strangers, risk aversion (see table 8)					
Debriefing Questionnaire	(i) demographic control variables (see table 4) (ii) beliefs over the experiment (see table 5)					

transferred to one's personal bank account at no cost. To strengthen the credibility of the payment procedure, we ask subjects to enter the e-mail address that is (or will be) associated with their PayPal account right after the introductory screen of the decision interface.

2.4 Summary of the design

To sum up, the experiment elicits the same decisions with similar procedures in both treatments. In particular, we recruit from the same subject pool, use the same monetary stakes, the same decision interface, and control the allocation of subjects between the lab and Internet treatments. This is summarized in Table 2.1, which also provides an exhaustive list of all the preferences we elicit.

At the same time, there are some important practical differences between the two kinds of implementations, most of which are due to subjects not being in the same physical space as the experimenter in the online implementation. Obviously, the standard procedure for laboratory experiments does not have to be adapted to such constraints. Our empirical strategy is to stick to common practice with the laboratory implementation, so as to keep the benchmark situation as close

Table 2.2. Summary of the design: differences in implementation between treatments

	Matching	Payment	Participation slot	Overbooked subjects
Inlab	Simultaneous	Cash	At time	Refused
Online	Sequential	Automated PayPal transfer	Any time during the half-day spanning the slot	Identified in the data and allowed to participate

as possible to existing evidence. We tried to choose the most innocuous adaptations when we had no choice but to introduce a difference between the two designs. Table 2.2 summarizes the resulting differences between our two treatments.

We conducted two different sets of experimental sessions, each conducted over a one-week period: 6 sessions (3 in the lab, 3 online) were conducted in November 2010 and 12 sessions (6 in the lab, 6 online) were conducted in November 2011.[9] Overall, 180 subjects performed the experiment in the laboratory and 202 subjects performed it online. We conducted 8 sessions with games order 1 (80 participants in the lab, 85 online), 6 sessions with games order 2 (60 and 67) and 4 sessions with games order 3 (40 and 50). Subjects in both conditions earned on average 21.24€ from the experiment.

3 Social preferences in the online laboratory

This section reports on our main outcome of interest, *i.e.* the reliability of the online elicitation of social preferences, taking laboratory behavior as a benchmark. In the next section, we assess the internal validity of both the online experiment and the comparison with laboratory behavior, based on the analysis of underlying secondary outcomes and additional robustness treatments.

Figure 5 provides a qualitative comparison of the behavioral patterns observed in the lab and online. For all games, the preferences we elicit online are parallel to those generally observed in the laboratory – which our lab condition replicates. While the theoretical prediction in the Public Good game is full free-riding, we do observe a positive amount of contribution that ranges between 35% and 40% of the initial endowment. In particular, the Nash equilibrium of the one-shot game is strongly rejected everywhere, with a high share of subjects making other-regarding decisions.

In the Dictator game (Figure 5.g), we observe three striking variations when preferences are elicited online. In the laboratory, the mode of the distribution is at 0, with 40% of subjects deciding not to give anything to their partner. For behavior online, the share of zero donors falls to half of this

[9] The 2010 version of the experimental economics platform did not elicit subjects' level of confidence in the experimental instructions, nor did it collect detailed data on the time spent by subjects on each screen of the interface. After observing that overall response times did indeed significantly differ between treatments, we decided to include those features before conducting further sessions.

proportion and the mode of the distribution is equal to 5 (*i.e.* equal split). Last, at the upper tail of the distribution, some subjects are willing to send more than 70% of their endowment online while no such behavior is observed in the laboratory. All three inflexions go in the direction of more other-regarding decisions online. In the Ultimatum Bargaining game (Figure 5.e), the shape of preferences for proposers are much more parallel, although we still observe a slightly higher share of zero donors in the laboratory (5%) as compared to online subjects (0%). Similarly, for receivers (Figure 5.f), the observed patterns are very similar with a mode at the equal split threshold, although there exists a slight difference at the bottom of the distribution with the share of low thresholds being 5% higher in the laboratory.

In both the Trust game (Figure 5.c) and the Public Good game (Figure 5.a), the same qualitative variation as in the Dictator game can again be observed: the high share of non-participants in the laboratory (1/4 of senders in the trust game, 1/5 in the Public Good game) is strongly reduced online, falling to around 1/10 in both instances. The remaining shape of the distribution is comparable, which again tends to suggest that players tends to be more pro-social online. Figures 5.b and 5.d describe the decisions elicited through the strategic method. Figure 5.b focuses on the Public Good game and plots the mean of the contributions to the common project made by subjects in the laboratory and Internet conditions, conditional on the average contribution made by the other 3 group members. In both fields, the qualitative pattern is very similar, with conditional contributions that are monotonically increasing in the average contributions of others but with a slope that is strictly lower than one. As this average group contribution increases, the distribution of conditional contributions among Internet subjects tends to dominate the distribution of conditional contributions among laboratory subjects, potentially indicating that online subjects were more prone to conditional cooperation. The overall effect, however, is relatively weak.

Figure 5.d, by contrast, exhibits a much stronger pattern. It plots the mean of the amount returned by participants Bs under laboratory and Internet conditions depending on the amount transferred by participant A. The shape of the social preferences elicited both online and in the laboratory points to the same conclusion: the amount returned by the trustee is strictly increasing in the amount received. The slopes, however, are quite different. The distribution of returns among Internet subjects strictly dominates the distribution of returns among laboratory subjects.

One consistent result in the literature about Trust games is that trustors are generally willing to place some of their resources in the hands of trustees. For their part, trustees typically tend to exhibit positive reciprocity, but the effect is usually not strong enough for this to be profitable to the trustor (Fehr & Camerer 2004). We can see this general pattern in our data, whereby participants Bs exhibit positive reciprocity, but tend to systematically return a lower amount to participant As than they

transferred in the first place. This result no longer holds among Internet subjects, however, in which participants Bs consistently return slightly more on average than the participant As initially transferred.

Last, regarding the risk aversion task, we follow Holt and Laury (2002) and interpret the number of times subjects chose the secure option as a raw measure of their level of risk aversion (Figure 5.h).[10] Again, the overall patterns of risk aversion in each pool of subjects share the same qualitative features: very few subjects are observed at the lower end of the distribution. Most of the sample switches after 5 risky decisions, with the majority of subjects switching between decisions 5 and 9. The figure also shows, however, that the distribution of risk preferences online strictly dominates the distribution in the laboratory, indicating that levels of risk aversion tend to be lower online. This observation confirms the results reported in Shavit et al. (2001).

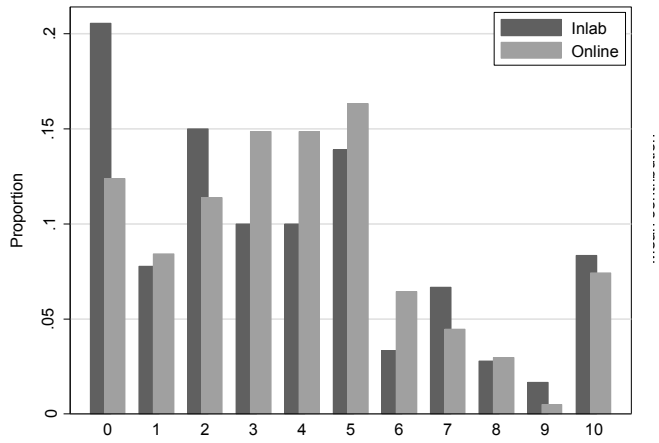
We now turn to a statistical assessment of the comparison. Table 3 reports on univariate non-parametric tests of differences between the two fields in terms of the mean and the dispersion of observed behavior. As regards mean comparisons, most of the differences discussed above induce statistically significant differences between the two elicitation fields (in 11 out of 14 measures). Leaving risk aversion aside, the most economically and statistically significant differences emerge in the Dictator game and the Trust game, especially as regards the behavior of trustees. On average, 58% of participant As in the Dictator game chose to transfer some fraction of their endowment to participant Bs in the lab, as opposed to 81% online. Overall, online subjects in the Dictator game transferred 17% more of their endowment to participant Bs. In the Trust game, they transferred about 9% more of their endowment, with this increase in trust being reciprocated in kind by participant Bs, who exhibited a reaction function to their transfers about 0.44 point steeper than laboratory subjects. Last, online subjects also appear significantly less risk-averse than laboratory subjects. The difference is significant at the 1% level, irrespective of whether we exclude confused subjects from the sample or not.

In their early experiments, Anderhub et al. (2001) and Shavit et al. (2001) both suggest that the variance in preferences tends to be higher when elicited online. Our statistical assessment does not confirm this conclusion. While the behavior in the Dictator game and risk aversion task do seem to be significantly more dispersed online, we actually find it to be significantly less dispersed for one of our measures of conditional cooperation in the Public Good game, and statistically indistinguishable from the variance generated in the lab for all the other measures.

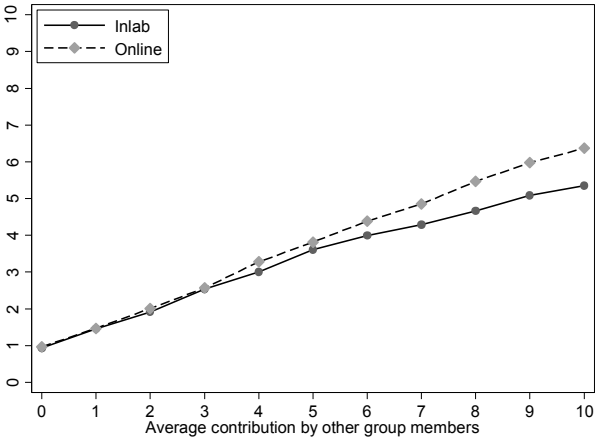
[10] Note that in constructing this figure, we excluded from the analysis the 5 laboratory and 22 Internet subjects who arguably misunderstood the task and choose option A in decision 10. Apart from the last data point, including those subjects has no impact on the figure.

Figure 5. Behavior in the decision problems between treatments

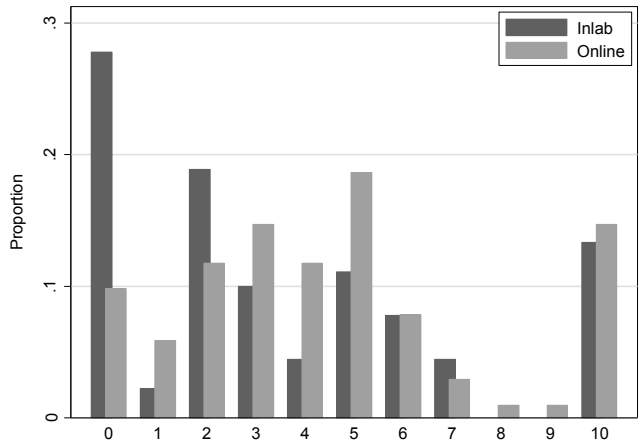
(a) Distribution of unconditional contributions in the Public Good game



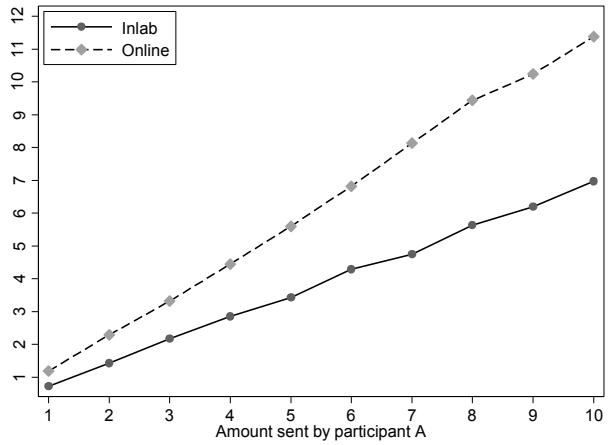
(b) Conditional contributions in the Public Good game



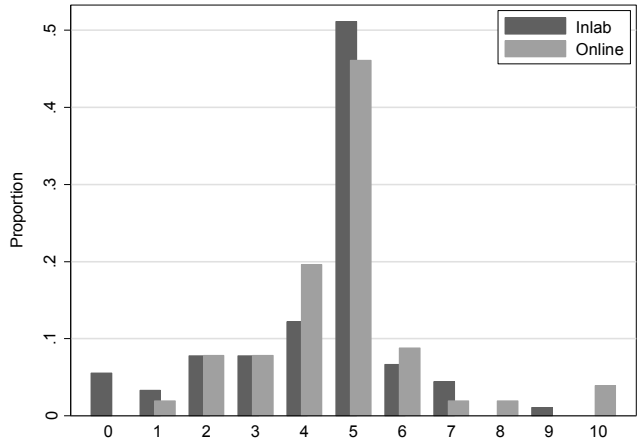
(c) Distribution of transfers in the Trust game



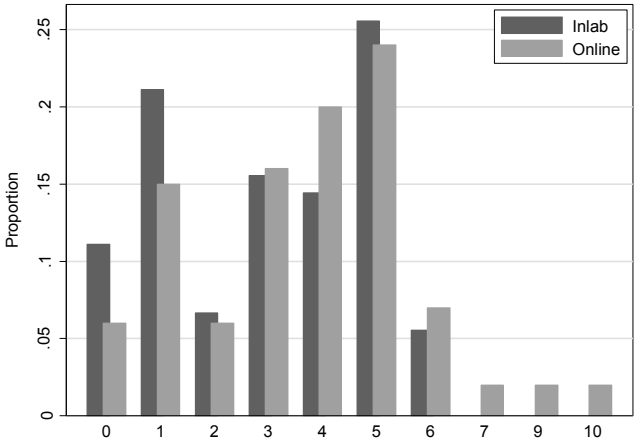
(d) Amounts returned in the Trust game



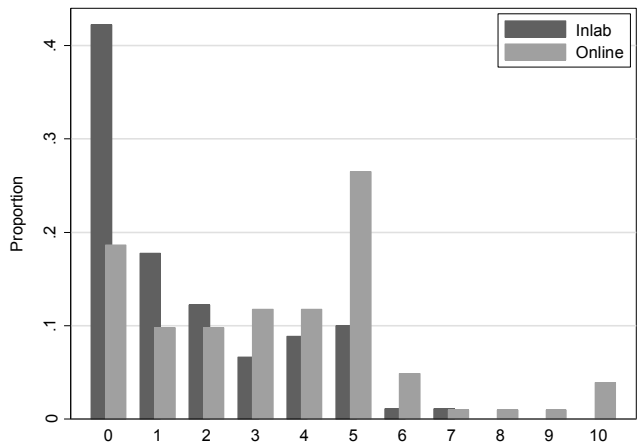
(e) Distribution of transfers in the Ultimatum game



(f) Distribution of minimum acceptable offers in the Ultimatum game



(g) Distribution of transfers in the Dictator game



(h) Risk aversion levels in the Holt&Laury task

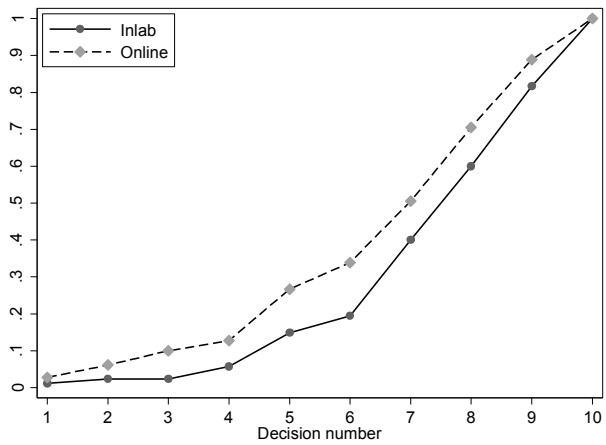


Table 3. Descriptive statistics

	Nb Of Obs.		Mean behaviors			Standard deviation		
Variable	Inlab	Online	Inlab	Online	*p*-value	Inlab	Online	*p*-value
Public Good Game								
Contribution	180	202	3.64	3.89	0.2028	3.06	2.73	0.1202
Mean conditional contributions	180	202	3.35	3.74	0.0394**	1.99	2.10	0.4567
Slope against low contributions others	180	202	0.53	0.57	0.6866	0.56	0.52	0.2870
Slope against high contributions others	180	202	0.35	0.51	0.0437**	0.73	0.61	0.0178**
Dictator Game								
Positive transfer	90	102	0.58	0.81	0.0004***	0.50	0.39	0.0203**
Transfer	90	102	1.62	3.36	0.0000***	1.88	2.53	0.0048***
Ultimatum Bargaining Game								
Transfer	90	102	4.28	4.72	0.4133	4.28	4.72	0.7469
Transfer threshold	90	100	3.00	3.69	0.0556*	1.90	2.14	0.2582
Trust Game								
Amount sent	90	102	3.54	4.45	0.0193**	3.32	3.01	0.3360
Mean amounts returned	90	100	3.85	6.29	0.0001***	3.72	4.33	0.1473
Slope against low amounts sent	90	100	0.67	1.10	0.0007***	0.72	0.82	0.2397
Slope against high amounts sent	90	100	0.71	1.20	0.0016***	0.91	0.98	0.4624
Holt&Laury lottery choices								
Nb of safe choices	180	202	6.76	6.15	0.0021***	1.78	2.03	0.0771*
Nb of safe choices w/o confused	164	152	6.80	6.18	0.0075***	1.70	2.01	0.0345**

Notes: *, ** and *** denote statistical significance at the 10, 5 and 1% levels.*p*-values are from Wilcoxon-Mann-Whitney tests (for differences in distributions) and two-sided variance comparison tests (for differences in variances), respectively. *Public Good game*: *Contribution* = unconditional contribution to the common project; *Mean conditional contributions* = mean of conditional contributions to the common project; S*lope against low contributions others* = slope of the reaction function for average contributions of other group members from 0 to 5; *Slope against high contributions others* = slope of the reaction function for average contributions of other group members from 6 to 10. *Dictator game*: *Positive transfer* = transfer in the Dictator game is positive; *Transfer* = transfer in the Dictator game. *Ultimatum game*: *Transfer* = transfer in the Ultimatum game; *Transfer threshold* = minimum acceptable offer in the Ultimatum game. *Trust game*: *Amount sent* = amount transferred in the Trust game; *Mean amounts returned* = mean of the amounts returned to participant A; *Slope against low amounts sent* = slope of the reaction function for amounts transferred by participant A from 1 to 5; *Slope against high amounts sent* = slope of reaction function for amounts transferred by participant A from 6 to 10. *Holt&Laury lottery choices*: *Nb of safe choices* = number of times (out of 10) the subject chose the secure option (*i.e.* option A); *Nb of safe choices w/o confused* = number of times (out of 10) the subject chose the secure option (*i.e.* option A) excluding the sub-sample of inconsistent subjects, *i.e.* all subjects who either chose the secure option (*i.e.* option A) in the last decision or switched back from option B to option A at least once.

Last, our risk aversion elicitation task allows us to directly investigate the issue of the quality of the data collected online. Overall, there were 13 inconsistent subjects in the laboratory as opposed to 44 online (two-tailed t-test, $p<0.01$).

There was also a fair proportion of subjects who clearly misunderstood the task and chose option A in the last decision. 5 subjects did so in the laboratory, as opposed to 22 over the Internet (two-tailed t-test, $p<0.01$). Consistent with previous findings, those results indicate that it is somewhat more difficult to obtain good quality data with web-based experiments, which should be compensated for by the ease with which the Internet allows to recruit larger samples.

To sum up, the comparison concludes that there is strong parallelism between the patterns of preferences elicited online and those elicited in a physical laboratory. We do observe some point differences between the two settings, though. Beyond the difference in risk attitudes (online subjects being less risk-averse), the most important differences in terms of social preferences are the intensity of altruistic behavior in the Dictator game and of the reciprocity of trustees in the Trust game. What is more, whether the differences are statistically significant or not, they always go in the direction of stronger other-regarding preferences when the elicitation takes place online. We now turn to additional evidence intended to assess the robustness of this surprising result as regards existing theories of social preferences applied to online environments.

4 Do subjects actually behave more pro-socially online?

To assess the robustness of our comparison, we first focus on factors that may impede the internal validity of our observations: composition effects in the subjects' pool, differences in the perceived credibility of the instructions, order effects and increased confusion online. Second, we investigate the differences between treatments as regards the companion measures delivered by our experiment, to see whether the differences that we identified could be explained by induced differences in secondary outcomes that might drive revealed preferences. Last, we report on companion treatments in the laboratory intended to assess the effect of the main differences in design between the online and the in-lab treatments.

4. 1 Internal validity of the comparison

Our design aims to control for any treatment-specific variation in the pool of subjects by matching participants according to their registration order. Still, our sample is not large enough to guarantee a perfectly balanced sample in terms of all demographic characteristics. If any of these demographics are correlated with social preferences, then the observed differences could be driven by pool composition effects rather than the online elicitation procedure.

Table 4 provides a comparison between the two pools along all demographics available from the experiment. Out of the 12 demographic characteristics that we tested, the randomization procedure failed on one: there seem to be 7% more subjects in the laboratory sample who were not born in

Table 5. Beliefs over the experiment

	(1)	(2)	(3)	(4)	(5)
	Believes others are human subjects	Believes final payment will be proceeded	Has read the instructions carefully	The environment was calm	Has already participated in similar study
Online	0.0655	-0.0408	-0.0198	-0.1510**	-0.0107
(*p*-value)	(0.579)	(0.662)	(0.788)	(0.021)	(0.832)
N	265	271	382	382	382
R^2	0.001	0.001	0.000	0.014	0.000

Notes: OLS estimates with baseline=Inlab. *p*-values are reported in parenthesis. *, ** and *** denote statistical significance at the 10, 5 and 1% levels. Constants not reported.

France.[11] There are no significant differences between samples in subjects' age, mothers' origin, degree level, degree level of parents, salary, student status, participation in civic organizations or religiosity.

A second concern in the comparison of the two elicitation fields is a potential difference in subjects' perception about the credibility of the instructions and the payment method. Table 5 provides a summary of the self-reported assessment of the experiment stated by our subjects. Laboratory and Internet subjects report similar levels of confidence in the fact that they interact with real human partners during the experiment and will be paid at the end of the experiment as described in the instructions. We interpret these results as supportive of the internal validity of our online experimentation procedure. Further, there are also no significant differences between treatments in the care that subjects reported taking in reading the experimental instructions or in the proportion of subjects who report having participated in a similar study in the past. The only statistically significant difference that arises from this table is how calm subjects report their environment to have been when they performed the study, although the magnitude of the reported difference is small (-0.15 for Internet subjects on a 4-point scale).

Thanks to the controlled allocation of subjects across treatments, very few observable differences between the two pools arise. Moreover, the common decision platform and the overall design of the experiments have generated very few differences between subjects as regards their assessment of the credibility of the instructions. The two exceptions are the proportion of subjects who were not born in France and how calm subjects report their environment to have been when they performed the experiment. To assess the robustness of observed behavior to these dimensions, we perform separate regressions on each outcome of interest that control for all covariates (of which coefficients are omitted) and in particular these two significant differences. One last dimension that may influence our results is the possible presence of order effects. We include controls for this dimension as well. The results are reported in panel A of Table 6. We observe that the "not born in France" and "calm

[11] The table actually reports two statistically significant coefficients: one associated with the fact of not being born in France, the other associated with the fact of having a father not born in France. It turns out that these two variables are heavily related in the sample (*corr*=0.51; $p<0.001$).

Table 4. Demographic characteristics between treatments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Age	Female	Not born in France			Highest degree completed			Salary	Student	Participates in civic organization	Religious Person
			Subject	Father	Mother	Subject	Father	Mother				
Online	0.0436	0.0564	-0.0706*	-0.103**	-0.0237	-0.192	-0.371	-0.200	-0.0034	-0.0151	0.0717	0.0272
(*p*-value)	(0.969)	(0.269)	(0.0865)	(0.0423)	(0.642)	(0.213)	(0.169)	(0.431)	(0.977)	(0.760)	(0.104)	(0.548)
N	382	382	382	382	382	381	262	266	372	382	382	382
R^2	0.000	0.003	0.008	0.011	0.001	0.004	0.007	0.002	0.000	0.000	0.007	0.001

Notes: OLS estimates with baseline=Inlab. *p*-values are reported in parenthesis. *, ** and *** denote statistical significance at the 10, 5 and 1% levels. Constants not reported.

environment" variables have no significant impact on behavior, except for a positive and marginally significant effect of the former in the Public Good game. Similarly, the order in which games occur seems of secondary importance – as can be expected from the absence of feedback until the end of the experiment. The only exceptions concern the transfers in the Ultimatum game (order 3) and the Trust game (order 2). Importantly, we find that none of these control variables affect the estimated point differences in social preferences elicited online as compared with the laboratory.

While Table 5 shows that subjects trust the experimental instructions online and in the lab equally, we also observed in Section 3 that many more subjects appeared confused in the online risk aversion elicitation task. This raises the question of a relatively worse understanding of the instructions in this elicitation context, even though subjects reported similar levels of care in reading them. To assess the effect of this dimension, we replicate the statistical analysis of Table 6 on those subjects who showed no sign of confusion in the risk aversion task – thus using confusion in this decision problem as a proxy for confusion in the whole experiment.[12] We do not find any difference in either the significance level or even the magnitude of the relevant parameters.[13]

[12] Here we define confusion as either choosing the secure option (*i.e.* option A) in the last decision or switching back from option B to option A at least once. The results are available from the authors upon request.

[13] We ran two additional robustness checks confirming the reliability of these results (results are available from the authors upon request.). First, we excluded from the Internet sample all subjects who logged in to the online platform *after the target of 20 participants per experimental session had already been reached* (so that we obtained a perfectly balanced sample between laboratory and Internet subjects). We thus explored the possibility that our findings were driven by those Internet subjects who logged in to the experiment last in each session. Second, we ran the analysis on social preferences while explicitly controlling for individual levels of risk aversion in the Holt & Laury task. Contrary to Internet subjects, laboratory subjects had to incur some physical and monetary costs in order to get to the lab and play. Those costs incurred *a priori* could have made laboratory subjects relatively more willing to secure their earnings from the experiment, which could be the reason behind the higher levels of risk aversion in decision-making that we observed among laboratory subjects. This higher level of risk aversion, in turn, could have induced laboratory subjects to behave in a more conservative way (*i.e.* less pro-socially) in certain games. In neither case, however, do we find any impact on the magnitude and significance of our estimates.

Table 6. Regression analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Public Good				*Dictator*	*Ultimatum*		*Trust*				*Holt&Laury lotteries*	
	Contribution	Mean conditional contributions	Slope against low	Slope against high	Transfer	Transfer	Transfer threshold	Amount sent	Mean amounts returned	Slope against low	Slope against high	Nb safe choices	Nb safe choices *w/o confused*
Panel A: Includes controls for (i) demographic characteristics (ii) beliefs over the experiment and (iii) games ordering controls													
Online	0.187	0.0938	0.0636	0.112	1.945***	0.609*	0.653	1.102*	1.996***	0.337**	0.292	-0.683***	-0.822***
	(0.60898)	(0.72433)	(0.36483)	(0.18484)	(0.00000)	(0.05607)	(0.11764)	(0.05041)	(0.00902)	(0.02273)	(0.10589)	(0.00688)	(0.00248)
Not born in France	0.816*	0.599*	-0.0137	-0.126	0.644	0.564	-0.0129	0.847	0.543	0.138	-0.0621	-0.0731	-0.0701
	(0.08282)	(0.07961)	(0.87859)	(0.24456)	(0.16216)	(0.13873)	(0.98214)	(0.20748)	(0.60463)	(0.49630)	(0.80348)	(0.81990)	(0.84369)
Calm environment	0.184	-0.0119	-0.00846	-0.0215	0.0868	-0.0611	0.158	0.204	-0.329	-0.0747	-0.165	0.266	0.0455
	(0.55870)	(0.95841)	(0.88832)	(0.76675)	(0.77946)	(0.81170)	(0.67515)	(0.65192)	(0.63028)	(0.57342)	(0.31130)	(0.21746)	(0.84918)
Games order 2	0.199	0.133	-0.0389	-0.0157	0.595	0.384	0.496	1.382**	-0.405	-0.0997	-0.0674	0.155	-0.202
	(0.64939)	(0.67594)	(0.64324)	(0.87654)	(0.19400)	(0.31031)	(0.31252)	(0.03979)	(0.64863)	(0.56335)	(0.75003)	(0.60526)	(0.53396)
Games order 3	0.551	0.222	0.00994	0.0269	0.170	1.031***	-0.134	0.308	-1.419	-0.204	-0.262	0.00726	0.0255
	(0.21740)	(0.49261)	(0.90736)	(0.79319)	(0.70648)	(0.00667)	(0.79230)	(0.64088)	(0.12614)	(0.25645)	(0.23456)	(0.98103)	(0.93672)
Constant	0.0904	3.111**	0.508	0.189	-0.610	3.772**	1.372	-0.554	7.451*	1.040	2.580**	4.515***	6.312***
	(0.96118)	(0.02178)	(0.15391)	(0.65779)	(0.74134)	(0.01474)	(0.55247)	(0.83707)	(0.07734)	(0.20231)	(0.01085)	(0.00046)	(0.00002)
R^2	0.103	0.085	0.059	0.087	0.343	0.162	0.114	0.200	0.205	0.186	0.219	0.090	0.119
Panel B: Same as Panel A and (iv) game specific decision times													
Online	-0.516	-0.200	-0.106	0.144	1.847***	0.688**	0.609	1.140*	2.596***	0.472***	0.410**	-0.636**	-0.733***
	(0.28799)	(0.55616)	(0.23494)	(0.18396)	(0.00001)	(0.03475)	(0.14384)	(0.06146)	(0.00170)	(0.00300)	(0.03868)	(0.01050)	(0.00634)
Game specific timing	-0.581*	0.00596	-0.137**	0.113	0.339	0.508	-0.0711	0.575	1.097*	0.248**	0.216	-0.705***	-0.589**
	(0.07475)	(0.98117)	(0.03887)	(0.16210)	(0.46458)	(0.15725)	(0.81799)	(0.32477)	(0.07731)	(0.03857)	(0.15055)	(0.00897)	(0.03323)
Game specific timing x online	0.0913	-0.541	-0.0137	-0.197*	-0.965*	-0.504	-0.778*	-0.732	-0.196	-0.0705	-0.106	0.616**	0.458
	(0.85310)	(0.11845)	(0.87945)	(0.07612)	(0.07045)	(0.20893)	(0.05103)	(0.31528)	(0.79790)	(0.63239)	(0.56624)	(0.04433)	(0.14946)
Constant	0.339	2.719**	0.582	0.0629	-0.119	3.690**	0.296	-0.973	6.676	0.849	2.388**	4.963***	6.810***
	(0.85516)	(0.04541)	(0.10119)	(0.88428)	(0.94873)	(0.01783)	(0.89629)	(0.72344)	(0.11092)	(0.29144)	(0.01963)	(0.00012)	(0.00000)
R^2	0.121	0.108	0.099	0.097	0.378	0.177	0.182	0.208	0.256	0.247	0.241	0.115	0.141
N	257	257	257	257	131	131	126	131	126	126	126	257	207

Notes: OLS estimates with baseline=Inlab. *p*-values are reported in parenthesis. *, ** and *** denote statistical significance at the 10, 5 and 1% levels. Demographic controls are all variables from table 4. Beliefs over the experiment controls are all variables from table 5. Game specific timing variables are standardized. *Public Good Game*: *Contribution* = unconditional contribution to the common project; *Mean conditional contributions* = mean of conditional contributions to the common project; *Slope against low* = slope of the reaction function for average contributions of other group members from 0 to 5; *Slope against high* = slope of the reaction function for average contributions of other group members from 6 to 10. *Dictator game*: *Transfer* = transfer in the Dictator game. *Ultimatum game*: *Transfer* = transfer in the Ultimatum game; *Transfer threshold* = minimum acceptable offer in the Ultimatum game. *Trust game*: *Amount sent* = amount transferred in the Trust game; *Mean amounts returned* = mean of the amounts returned to participant A; *Slope against low* = slope of the reaction function for amounts transferred by participant A from 1 to 5; *Slope against high* = slope of the reaction function for amounts transferred by participant A from 6 to 10. *Holt&Laury lotteries*: *Nb safe choices* = number of times (out of 10) the subject chose the secure option (*i.e.* option A); *Nb safe choices w/o confused* = number of times (out of 10) the subject chose the secure option (*i.e.* option A) excluding the sub-sample of confused subjects, *i.e.* all subjects who either chose the secure option (*i.e.* option A) in the last decision or switched back from option B to option A at least once.

Table 7. Difference in median/variance of time spent on the experiment

Number of Observations		*Median time*			*Standard Deviation*		
Inlab	Online	Inlab	Online	Diff.	Inlab	Online	Diff.
120	154	35.01	28.50	6.51***	7.77	17.52	- 9.74***
				p<0.0001			p<0.0001

Notes: p-values are from a Wilcoxon-Mann-Whitney test (for difference in distributions) and two-sided variance comparison tests (for difference in variances), respectively. *, ** and *** denote statistical significance at the 10, 5 and 1% levels.

4. 2 Differences in underlying secondary outcomes

We now turn to a second kind of confounding factor that could challenge the inference drawn from observed preferences: the effect of the field of elicitation on secondary outcomes which may drive revealed preferences. We consider three dimensions in turn: decisions times, self-reported social preferences and the expected behavior of other subjects.

First, Shavit et al. (2001) report that participants in an Internet experiment tend to exhibit shorter decision times than classroom participants, which could, according to the literature, have a sizeable impact on behavior. Table 7 presents evidence regarding decision times in both treatments. We observe that the median time spent with the experiment among Internet subjects is 6.51 minutes lower than among laboratory subjects (Wilcoxon-Mann-Whitney test, p<0.0001), with an average completion time of 34 minutes across treatments. In addition, we also observe that the variance in the time spent on the experiment is significantly higher online (two-tailed F-test, p<0.0001). Notwithstanding this fact, we were surprised that none of our Internet subjects remained inactive for more than 5 minutes at any point when performing the study, which we interpret as good news for its internal validity.[14]

To assess the influence of this treatment effect on the preferences elicited in both fields, we include decision times in the regressions presented in Table 6. For each outcome, the decision time variable is defined as the time spent by the subject on the corresponding decision problem (from the instruction screen to the decision screen). We include it both as an additional control variable and in interaction with the online treatment so as to capture the variation in social preferences online that is induced by variations in decision times. The results are presented in panel B of Table 6. Many timing coefficients are not statistically significant. When they are, however, our estimates suggest that faster decisions are associated with more other-regarding decisions.

[14] Even if online subjects do seem to play faster on average, some of them spent quite a lot of time on the experiment. One extreme case was a subject who spent more than 3 hours on the experiment without once triggering the 5-minute inactivity indicator.

Table 8. Self-reported social preferences between treatments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Cooperation	Altruism	Fairness	Trust (WVS)	General trust	Trust in strangers	Riskaver
Online	0.457	0.148	-0.235	0.0887*	-0.0477	-0.0551	0.300
(*p*-value)	(0.117)	(0.474)	(0.271)	(0.0676)	(0.477)	(0.447)	(0.247)
N	366	376	372	352	370	372	271
R^2	0.007	0.001	0.003	0.010	0.001	0.002	0.005

Notes: OLS estimates of column variables on the online dummy (the baseline is inlab subjects, constants are not reported). *, ** and *** denote statistical significance at the 10, 5 and 1% levels. *Cooperation* = whether subjects consider it justifiable to free-ride on public social allowances; *Altruism* = whether subjects think that people are mostly looking out for themselves as opposed to trying to help each other; *Fairness* = whether subjects think that people would try to take advantage of them if they got a chance as opposed to trying to be fair; *Trust (WVS)* = whether subjects think that most people can be trusted or that one needs to be very careful when dealing with people; *General trust* = subjects' level of general trust in people; *Trust in strangers* = how much trusting subjects are of people they just met; *Riskaver* = whether subjects generally see themselves as fully prepared to take risks or as trying to avoid them.

For instance, a one standard deviation increase in decision time is associated with a 6% decrease in the proportion of the endowment unconditionally contributed and a decrease of 0.14 points in the slope of the reaction function in the Public Good game in the lab (although only for relatively low values of the average contribution of the other group members), as well as a 8.5% decrease in the proportion of the endowment that receivers in the Ultimatum game demand online. Incidentally, it is also associated with an average decrease of 0.71 in the level of risk aversion in the Holt & Laury task (but only in the lab). These results are in line with those reported in Rubinstein (2007), Rand et al. (2012) and Lotito et al. (2013), who report that shorter decision times are associated with more pro-sociality on average.[15] This evidence supports the System 1/System 2 hypothesis that shorter decision times are associated with instinctive and emotional decision processes (Kahneman 2003), which should drive subjects to behave relatively more pro-socially on average. On the other hand, the timing coefficients for the Trust game are at odds with the theory, as they indicate that higher decision times are significantly associated with an *increase* in trustworthiness.

Focusing on our coefficients of interest, we observe that controlling for decision times has no effect on the magnitude and significance of the point differences between treatments. One exception is the difference in levels of trustworthiness exhibited by participant Bs in the Trust game, which even increases.[16]

Next, we explore whether the elicitation field had an impact on subjects' self-reported measures of social preferences, which could in turn have had an effect on their behavior. To do so, the final questionnaire asked subjects to answer a set of traditional survey questions about social preferences.

[15] The evidence reported in Piovesan & Wengstrom (2009) is an exception.

[16] The change in the magnitude of these coefficients is explained by the negative correlation between the Internet treatment and average decision time, which is found to be positively and significantly associated with our measures of trust and trustworthiness.

Table 9: Beliefs about other subjects' decisions by treatment

	(1)	(2)	(3)	(4)	(5)
	How much others should contribute	Idea about how much others will contribute	Estimation of how much others will contribute	Idea about how much the responder will return	Estimation of how much the responder will return
Online	-0.450	-0.0910*	0.202	-0.0719	0.0737
(*p*-value)	(0.147)	(0.0538)	(0.496)	(0.260)	(0.584)
N	381	382	266	192	116
R^2	0.006	0.010	0.002	0.007	0.003

Notes: OLS estimates with baseline=Inlab. *p*-values are reported in parenthesis. *, ** and *** denote statistical significance at the 10, 5 and 1% levels. Constants not reported. (1) is how much subjects think people *should* contribute to the common project in the Public Good game; (2) is whether subjects had an idea of how much the other subjects in their group *would actually* contribute to the common project when they made their decision; (3) is conditional on (2), how much subjects thought the other subjects in their group would contribute on average when they made their decisions. (4) is whether subjects in the role of senders in the Trust game had an idea of how much the responder would return to them when they made their decision; (5) is conditional on (4), proportion of the amount sent that trustors anticipated would be returned to them by the trustee when they made their decision.

The result of the comparison between subject pools is reported in Table 8. We can see that no statistically significant differences arise between laboratory and Internet subjects in self-reported social preferences, except for the WVS and GSS trust question, in which roughly 9% more subjects report that "most people can be trusted" in the Internet sample (p<0.10).[17]

Last, Table 9 provides a comparison of subjects' self reports on the expected behavior of other participants in the Public Good and Trust games between treatments. The point differences in social preferences that we identified especially strongly in the Trust game do not seem to be mediated by a modification of subjects' expectations about the behavior of others depending on the experimental context either. Indeed, the only (marginally significant) difference that arises in terms of expectations is in whether subjects report having an idea of how much the other members of their group contributed when they made their decision in the Public Good game (-9% in the Internet sample, p<0.10).

4. 3 The effect of the Internet-specific differences in design

As stressed in Section 2.4, our strategy in designing the experiment is to make the online and in-lab environments as similar to each other as possible, while ensuring that the in-lab conditions complied with standard practice. This led us to introduce two important differences between the two designs, so as to account for the specific constraints faced when subjects do not come to a physicial laboratory to participate. First, the compensation of online subjects goes through an automated PayPal transfer, which is less immediate, and perhaps less salient, than the cash payment offered to laboratory

[17] Theses measures are very likely to be correlated with unobserved factors determining behavior in our games, and so we do not include them as control variables in the regressions.

subjects. Second, since we wanted to allow online subjects to progress within the experiment at their own pace without having to wait for others to make decisions, we implemented a sequential matching scheme between participants. Importantly, this implies that the decisions made by an online subject do not affect the outcome of his current partner, but the outcome of some future online subject. In this section, we check for the sensitivity of the observed differences in behavior between the two environments to these changes in the design, through additional laboratory experiments involving each feature in turn.

4.3.1 Design of the robustness treatments

We ran two companion treatments in the laboratory. In the *Sequential Matching* treatment, subjects in the laboratory experiment are matched with subjects from previous sessions. In the *PayPal* treatment, participants in the laboratory experiment are paid by an automated PayPal transfer. In order to comply with the general rules of our laboratory, and avoid negative reactions both in the overall subject pool and towards our experiment, this feature of the design had to be announced at the registration stage.[18] More precisely, on the webpage on which subjects confirm their willingness to participate, a preliminary screen informed them that experimental earnings would paid be through PayPal transfers. Subjects were allowed to decline participation at this stage, in which case we recorded the information available in the subject management database if provided by the subjects, *i.e.* their gender, age and student status.

Three sessions of each treatment were run in May 2013. We chose the sequence of games (as described in Section 2.2) so as to balance the overall number of sessions for each order: we ran one session of each treatment with order 2, and two sessions with order 3. Since these sessions took place after our main treatments of interest, and without an online countepart, our control over self-selection into the elicitation field does not apply to these treatments – subjects registered on the usual first-come first-served basis for both treatments. This concern about the composition of the subject pool is reinforced by self-selection at the registration stage of the PayPal treatment, as 20% of subjects actually gave up on their registration when informed of the PayPal payment.[19]

[18] This is unlike our Internet treatment, in which subjects were informed that the final payment would be processed through PayPal right after the introductory screen of the online platform, *i.e.* after they had already registered and logged in to participate (see section 2.3). For the present treatment, self-selection into participation due to the payment system can hardly be avoided for any payment method other than cash. Even if our laboratory usually paid subjects using PayPal (or, say, a bank transfer) we would have had to announce this in the recruitment adds, hence inducing self-selection into the overall population of potential subjects. In that sense, the selection effect that occurs in this treatment replicates the one at stake in a laboratory using PayPal as a way to dematerialize subject's payments.

[19] As a comparison, it is notable that none of the subjects in the Internet treatment dropped out of the experiment at the level of the PayPal payment screen. According to the data available for this treatment, subjects who gave up on their registration at the

Table 10: Demographic characteristics between the in-lab, sequential matching and PayPal treatments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Age	Female	Not born in France			Highest degree completed			Salary	Student	Participates in civic organization	Religious Person
			Subject	Father	Mother	Subject	Father	Mother				
SeqMatch	2.628	0.0389	-0.0556	-0.178**	-0.0944	-0.246	-0.316	0.0796	0.0679	-0.0889	-0.00556	0.0500
(*p*-value)	(0.110)	(0.600)	(0.382)	(0.0165)	(0.202)	(0.269)	(0.369)	(0.814)	(0.690)	(0.205)	(0.926)	(0.460)
PayPal	-2.824*	-0.178**	0.0444	0.0722	0.106	0.338	-0.203	0.119	-0.299*	0.178**	-0.0222	0.183***
(*p*-value)	(0.0859)	(0.0169)	(0.485)	(0.328)	(0.154)	(0.127)	(0.559)	(0.721)	(0.0843)	(0.0115)	(0.711)	(0.00708)
N	382	382	382	382	382	381	262	266	372	382	382	382
R^2	0.000	0.003	0.008	0.011	0.001	0.004	0.007	0.002	0.000	0.000	0.007	0.001

Notes: OLS estimates with baseline=Inlab. *p*-values are reported in parenthesis. *, ** and *** denote statistical significance at the 10, 5 and 1% levels. Constants not reported.

Table 10 provides an overview of the demographics in the pool of subjects who participated in the SeqMatch and PayPal robustness treatments as compared with the standard laboratory one. Despite the different sample sizes (180 online as opposed to 60 in each additional treatment), we observe very few differences between the in-lab and sequential matching samples. The only significant difference that arises concerns the nationality of the father. The high refusal rate of the PayPal treatment had a greater impact on the composition of the sample, however, as PayPal subjects are on average less likely to be female, more likely to be students and religious and also younger with a lower income (although marginally significantly so).

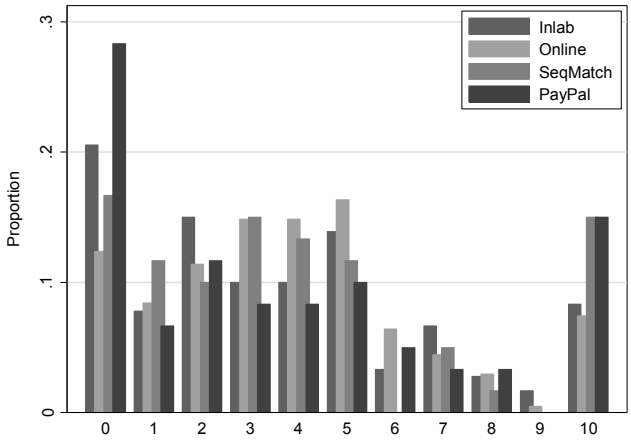
4.3.2 Results

Figure 6 replicates the qualitative description of observed behavior of Figure 5 with the four treatments taken together. In all games, the qualitative patterns in elicited preferences remain the same. One notable feature of the figure is that the relatively low proportion of fully self-interested decisions in the online treatment that we identified in Figure 5 is not replicated by either the sequential matching or the PayPal treatments. Indeed, less than 20% of subjects make no transfer in the Dictator game in the online treatment, while this proportion is more than doubled in the other three laboratory treatments (figure 6.g). For this decision, the online condition is also the only one to have its mode at an equal split of the endowment (decision made by about 25% of online subjects, as opposed to 10% or less in all other samples), while the other three treatments have a mode at zero. Similarly, less than 10% of subjects make no transfer in the Trust game in the online condition, while this proportion is again more than doubled in the other treatments (figure 6.c). This pattern is less clear-cut for the contribution decisions in the Public Good game (figure 6.a) and the transfer and

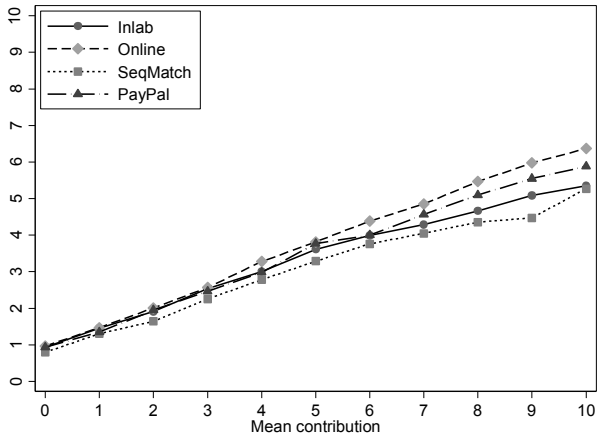
stage of the PayPal payment explanation screen were on average 23.3 years old (as opposed to 24.6 for those who eventually participated in the experiment), 30% female (as opposed to 35%) and 56% students (as opposed to 82%).

Figure 6. Behavior in the decision problems between treatments (including the SeqMatch and PayPal treatments)

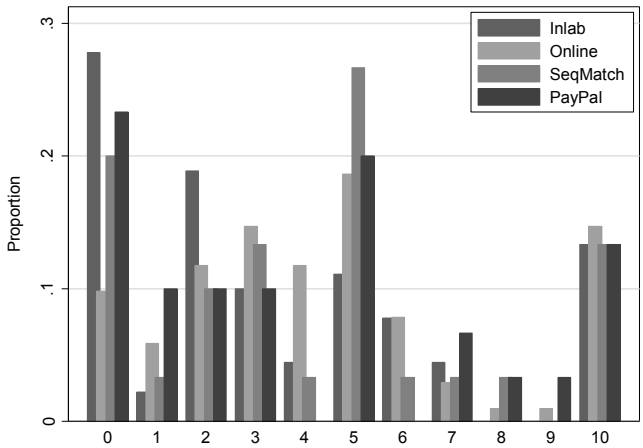
(a) Distribution of unconditional contributions in the Public Good game



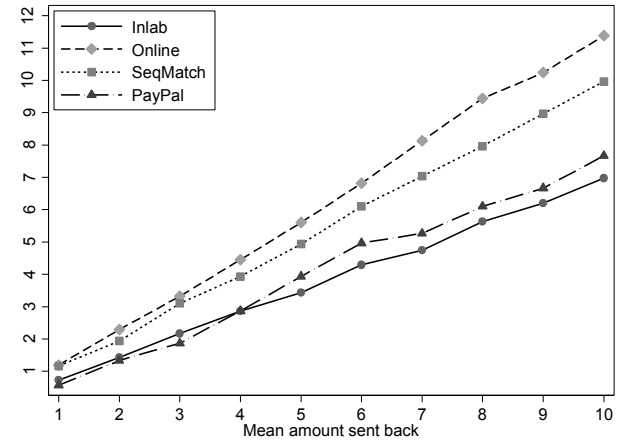
(b) Conditional contributions in the Public Good game



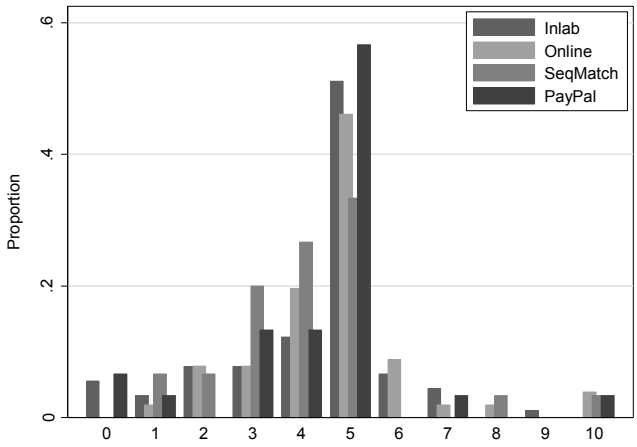
(c) Distribution of transfers in the Trust game



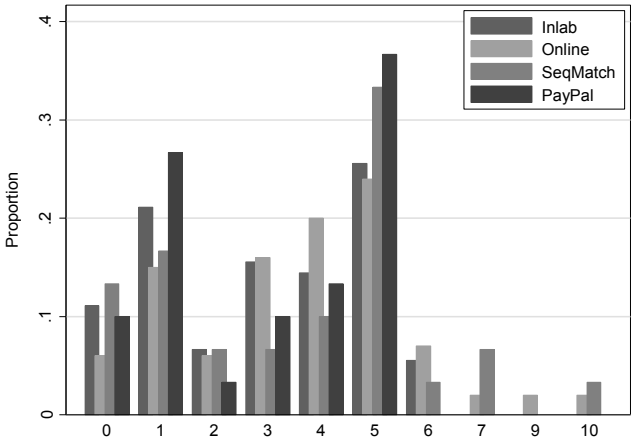
(d) Amounts returned in the Trust game



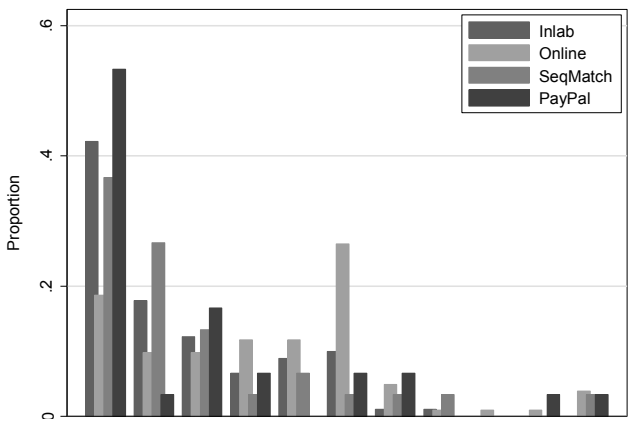
(e) Distribution of transfers in the Ultimatum game



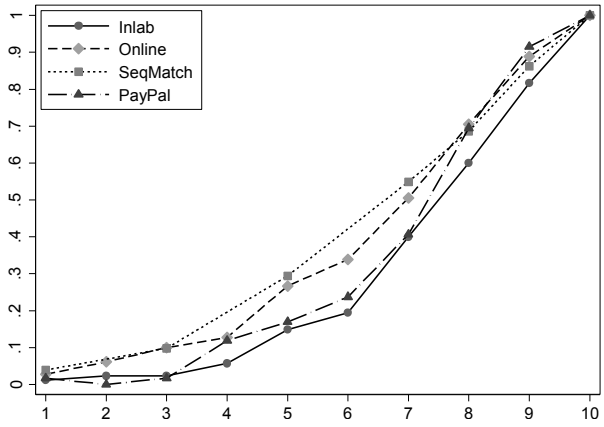
(f) Distribution of minimum acceptable offers in the Ultimatum game



(g) Distribution of transfers in the Dictator game



(h) Risk aversion levels in the Holt&Laury task



threshold decisions in the Ultimatum game (figures 6.e and 6.f, respectively), but remains visible.

Another insight from Figure 6 is that the distribution of returns for online subjects in the Trust game continues to dominate the distribution of returns for all other laboratory subjects (figure 6.d). It is striking, however, that when compared with the patterns of trustworthiness exhibited in the in-lab and PayPal treatments, the pattern exhibited in the sequential matching treatment is much closer to that of the online treatment. This suggests that the point differences in trustworthiness levels that we identified between our lab and Internet conditions might be at least partly due to the sequential matching that we implemented between online subjects. This result is surprising, as one might have expected the indirect reciprocity induced by this matching procedure to weaken rather than strengthen trustworthiness.

We now turn to a more formal statistical assessment of the four treatments. We proceed in two steps. First, in panel A of Table 11 we provide estimates of the treatment effects using the same specification as in panel B of Table 6 above. We observe few differences between the baseline laboratory treatment and the sequential matching and PayPal treatments, as virtually all coefficients on those robustness treatments are insignificant. Strikingly enough, one prominent exception is the level of risk aversion, which is significantly affected by the sequential matching procedure implemented in the lab. This result is surprising, as this decision problem is the only one that does not involve interactions with other subjects.

These regression results stand as a rather weak robustness test, as they may be affected by the differences in sample size between treatments. As an additional more rigorous test of the robustness of the comparison, the two bottom panels of Table 11 provide mean comparison tests against each treatment. We compare the preferences elicited online with those elicited in each robustness treatment as a benchmark in turn. These comparisons thus inform about how well online behavior is replicated by behavior in a laboratory experiment in which subjects are, respectively, matched sequentially or paid by automated PayPal transfers. Remember that only two out of the three orders considered in our treatments of interest are implemented for the robustness treatments. We thus control for order effects in the mean comparison tests reported in the table. In line with the pattern observed in the qualitative discussion, we observe that some of the previously significant differences are no longer significant when the laboratory sessions incorporate the differences in design. Focusing on social preferences, sequential matching in the laboratory seems to replicate the higher levels of trust and trustworthiness found online in the Trust game. The higher level of donation in the Dictator game, by contrast, is robust to both changes and appears to be specific to the online elicitation field. In line with the top panel of Table 11, the risk preferences elicited online are no longer different from the ones observed in the lab, when it features either PayPal payment or sequential matching.

Table 11. The effect of sequential matching and PayPal payment on behavior

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	Public Good				*Dictator*	*Ultimatum*		*Trust*				*Holt&Laury lotteries*	
	Contribution	Mean conditional contributions	Slope against low	Slope against high	Transfer	Transfer	Transfer threshold	Amount sent	Mean amounts returned	Slope against low	Slope against high	Nb safe choices	Nb safe choices *w/o confused*
Panel A: All treatments pooled – Baseline=Inlab treatment													
Online	-0.287	-0.160	-0.0395	0.0864	1.867***	0.594*	0.631	1.158*	2.808***	0.479***	0.470**	-0.612**	-0.682**
	(0.54645)	(0.63026)	(0.61771)	(0.38877)	(0.00000)	(0.06187)	(0.10752)	(0.05352)	(0.00019)	(0.00094)	(0.01078)	(0.01405)	(0.01016)
SeqMatch	-0.560	-0.394	-0.0551	0.0267	0.516	0.0158	0.270	0.580	1.489	0.205	0.222	-0.695**	-0.929***
	(0.29866)	(0.31632)	(0.55626)	(0.82177)	(0.35814)	(0.97291)	(0.59898)	(0.46843)	(0.11413)	(0.25960)	(0.34017)	(0.03586)	(0.00765)
PayPal	-0.401	0.128	-0.000327	-0.00278	0.524	0.0111	0.130	0.637	1.324	0.322*	0.0768	-0.513	-0.437
	(0.47030)	(0.75264)	(0.99730)	(0.98180)	(0.33252)	(0.98042)	(0.80956)	(0.42938)	(0.17955)	(0.09202)	(0.75288)	(0.13495)	(0.20333)
Constant	0.748	3.125**	0.529*	-0.0689	-0.0818	4.828***	0.202	-0.735	4.034	0.534	1.129	5.101***	6.949***
	(0.66367)	(0.01337)	(0.07774)	(0.85590)	(0.96068)	(0.00049)	(0.91083)	(0.76353)	(0.22149)	(0.40156)	(0.16765)	(0.00001)	(0.00000)
N	369	367	367	367	185	185	184	185	184	184	184	368	304
R^2	0.080	0.046	0.072	0.067	0.387	0.186	0.158	0.187	0.224	0.204	0.154	0.096	0.137
Panel B: Comparison of the Online and SeqMatch treatments (controls for games orders included)													
Online	0.353	0.499	0.0320	0.0647	1.971***	0.871**	0.354	0.435	0.501	0.153	0.0280	-0.201	-0.114
	(0.44963)	(0.15388)	(0.70821)	(0.52592)	(0.00081)	(0.02722)	(0.48965)	(0.53888)	(0.62290)	(0.43184)	(0.90162)	(0.55129)	(0.76333)
N	262	261	261	261	132	132	130	132	130	130	130	262	200
R^2	0.010	0.023	0.010	0.011	0.087	0.044	0.029	0.010	0.011	0.007	0.017	0.007	0.007
Panel C: Comparison of the Online and PayPal treatments (controls for games orders included)													
Online	0.566	0.102	-0.0320	0.0465	1.776***	0.618	1.002**	0.641	2.069**	0.286	0.325	-0.475	-0.589*
	(0.23520)	(0.77762)	(0.69658)	(0.65696)	(0.00338)	(0.12082)	(0.03950)	(0.37804)	(0.03431)	(0.11991)	(0.14835)	(0.14064)	(0.08408)
N	262	262	262	262	132	132	130	132	130	130	130	262	208
R^2	0.013	0.006	0.007	0.008	0.073	0.027	0.044	0.010	0.047	0.022	0.038	0.011	0.017

Notes: OLS estimates with baseline=Inlab. *p*-values are reported in parenthesis. *, ** and *** denote statistical significance at the 10, 5 and 1% levels. Panel A compares the Inlab treatment to the other three treatments. Demographic controls are all variables from table 4. Beliefs over the experiment controls are all variables from table 5. Game specific timing variables are standardized. . Panels A and B compare the Online treatment to the SeqMatch and PayPal treatments, respectively (constants not reported; regressions control for games ordering effects only). *Public Good Game*: *Contribution* = unconditional contribution to the common project; *Mean conditional contributions* = mean of conditional contributions to the common project; *Slope against low* = slope of the reaction function for average contributions of other group members from 0 to 5; *Slope against high* = slope of the reaction function for average contributions of other group members from 6 to 10. *Dictator game*: *Transfer* = transfer in the Dictator game. *Ultimatum game*: *Transfer* = transfer in the Ultimatum game; *Transfer threshold* = minimum acceptable offer in the Ultimatum game. *Trust game*: *Amount sent* = amount transferred in the Trust game; *Mean amounts returned* = mean of the amounts returned to participant A; *Slope against low* = slope of the reaction function for amounts transferred by participant A from 1 to 5; *Slope against high* = slope of the reaction function for amounts transferred by participant A from 6 to 10. *Holt&Laury lotteries*: *Nb safe choices* = number of times (out of 10) the subject chose the secure option (*i.e.* option A); *Nb safe choices w/o confused* = number of times (out of 10) the subject chose the secure option (*i.e.* option A) excluding the sub-sample of confused subjects, *i.e.* all subjects who either chose the secure option (*i.e.* option A) in the last decision or switched back from option B to option A at least once.

Overall, this exercise leads to mixed conclusions. On the one hand, the comparison confirms our main conclusion that, contrary to what is generally thought, other-regarding preferences are no less intense online than in the laboratory. For the Dictator game, the higher level of transfers even remains strongly significant in comparison to all three laboratory situations. On the other-hand, both PayPal payment and sequential matching of subjects in the lab seem to influence revealed preferences, and account for part of the point differences we observe. This raises interesting questions, as dematerialized payment is most likely to become the standard way to remunerate subjects in online experiments, and as the indirect reciprocity involved in sequential matching could have been expected to weaken rather than strengthen social preferences. As for the purpose of this study, these results show that design choices compatible with online experimentations are not neutral on behavior, and deserve systematic experimental investigation.

5 Discussion

From the results developed in the previous sections, our main methodological conclusion is in favor of the internal validity of the preferences elicited online, thanks to the additional controls of our design. In particular, no significant difference between treatments appeared in subjects' self-reported beliefs about the accuracy of the experimental instructions. In the same vein, we found that none of our online subjects seemed to have been distracted from the experiment for more than 5 minutes (although major distractions may occur in an even shorter time-range) and that a relatively modest number of online subjects (6 out of 208) eventually dropped out of the experiment before its completion. Importantly, unlike earlier studies (*i.e.* Anderhub et al. (2001) and Shavit et al. (2001)), the dispersion of preferences that we elicit online is often statistically indistinguishable from that of the lab.

The experiment does highlight some specificities of online elicitation of behavior, though. Consistent with the above-mentioned seminal studies, we find that it is relatively more difficult to collect good quality data over the Internet, as 22 subjects on the Internet failed to select option B in the 10^{th} decision (in which subjects had the choice between earning 20€ or 38.5€ with certainty) as compared with 5 in the laboratory. However, it should be possible to compensate for this extra noise in the data by leveraging the Internet to recruit larger samples. Finally, we find that online subjects play significantly faster on average than laboratory subjects, with sometimes a sizeable impact on behavior. Depending on the kind of experimental data, including controls for this dimension of behavior can therefore be important.

These observations speak in favor of the reliability of Internet data. The second important question this paper aimed to answer is the reliability of Internet-based inference – taking behavior in the laboratory as a benchmark. The qualitative patterns in the data unambiguously answer yes to this question, as the Internet-based experiment generates social preferences that are similar to the laboratory ones. Subjects interacting in an online setting exhibit pro-social behavior, are conditionally cooperative on average, often altruistic in the Dictator game, reveal a taste for fairness in the Ultimatum game that other subjects anticipate in the form of higher average transfers, and exhibit both trust and trustworthiness in the Trust game.

Beyond the reliability and the internal validity of social preference elicitation online, we also find that the magnitude of other-regarding behavior is not weakened by social interactions online. The amount sent in the dictator game, and the amount returned in the trust game is even significantly higher for online subjects . A more exacting assessment of the data in this regard would consist in looking statistically at the simultaneous coincidence (or difference) in social preferences elicited in both fields. To define the null of such a test, however, one has to choose which outcomes or measures are worth considering. For instance, one could focus on one outcome variable per decision role in each game, or include all averages described in Table 3, account for decisions times as well, or even add differences in variance and the like. Instead of reporting the statistics on the joint significance of all imaginable combinations of outcomes of interest, or choosing a few particular combinations, we decided to report all results with the *p*-values of univariate comparisons. The Bonferroni correction for multiple comparisons can then be applied to test for joint equality of any combination of the results reported (Bland & Altman 1995). According to the correction, the threshold used to conclude on the equality of k outcomes of interest in order to replicate a Type I error equal to α is α/k. Given the strength of the statistical differences in both the trust game and the dictator game, such an exercise concludes in most instances that there is a significant difference in behavior between the two settings,[20] in the direction of higher other-regarding preferences online.

Given that the Internet is often viewed as the realm of anonymity (and rightly so), one might have expected the increased social distance between Internet-based subjects to drive measures of social preferences down, compared with the traditional laboratory setting. For instance, Hoffman et al. (1996) show that subjects tend to decrease the amount of their transfers in the Dictator game when

[20] The exact *p*-value on the test of mean equality in transfers in the dictator game from Table 3 is 7.39e-7, which drives rejection even if one accounts for more than 1000 outcomes. If we instead focus separately on positive transfers and conditional transfers, *i.e.* restricting to positive contributions only, the p-value of the difference in contributions in the dictator game is 0.0003 leading to more mix conclusions (in the trust game, the p-value on the share of positive returns is 0.015, it is 0.0212 for the comparison in mean amounts returned if positive). For instance, the equality in social preferences between the in-lab and online treatments is rejected at the 1% level if we consider that each game yields one outcome of interest per decision role (*i.e.* k=6, adjusted threshold=0,0017), or if we consider each variable reported in Table 3 as one outcome of interest (i.e. k=14; adjusted threshold =0.0007). The conclusion is reversed if the variance of outcome behavior (14 outcomes), as well as the beliefs over the experiment (5) and the self reported measures of trust (5) are accounted for (k=38; adjusted threshold =0.00026).

social distance (*i.e.* isolation) increases and Glaeser et al. (2000) report that measures of trust and trustworthiness tend to increase with the level of demographic similarity between both players. As regards social distance theory, two alternative conclusions can be drawn from this observation. It challenges either the generally acknowledged greater social distance that prevails on the Internet (Fiedler et al. 2011), or the prediction of social distance theory per se. Our data cannot distinguish between these two views of our results.

A tentative alternative explanation can be found in the nature of many of the social and economic interactions in which individuals tend to engage online, which they may bring to the experiment through its contextual implementation. As the Internet is an environment in which it is difficult to enforce contracts, trust and trustworthiness are likely to be major devices through which to secure online transactions and build a reputation for oneself (Greif 2006). So perhaps the strong anonymity that prevails in Internet-based interactions does not come at the expense of social preferences.[21] The prominent role of trust and trustworthiness in Internet-based economic transactions has already been demonstrated in the case of a popular online auction site (Resnick et al. 2006). In a similar fashion, the drastic reduction in communication and coordination costs brought about by the Internet has made it easier for individuals to behave altruistically towards one another, as exemplified by the impressive growth of question-driven online message boards and customer review systems.

In a recent paper, Hoffman and Morgan (2011) explored the hypothesis that selection pressures resulting from high competition, low entry and exit barriers and agents' anonymity in online business environments should drive individuals with strong social preferences out of those markets. They got professionals from the Internet domain trading and online adult entertainment industries to perform a series of social preference experiments and compared the results to those obtained from a population of undergraduate students. Contrary to what they initially expected, they found that Internet business people are significantly more altruistic, more trusting, more trustworthy and less likely to lie. They interpreted these findings as support for the idea that social preferences are rewarded in the Internet environment, where they help to smooth interactions and are thus beneficial in the long run. Again, our study was not designed to test this explanation against any of a possible set of alternative hypotheses. Future studies should dig into the precise nature of this "Internet effect" that we have found.

[21] The lack of an "institutional" way of securing social and economic interactions over the Internet is often invoked as a reason why many Internet users who value their anonymity online are nonetheless willing to stick to and invest in a unique online identity or pseudonym.

6 Conclusion

The Internet is becoming increasingly attractive to experimenters, both as a *medium* through which to target larger and more diverse samples with reduced administrative and financial costs, and as a *field* of social science research in its own right. In this paper, we report on a randomized experiment eliciting social preferences and risk aversion both online and in the laboratory based on the same, original, Internet-based platform. To provide a testbed comparison of social experimentations online, our platform seeks to control for most of the dimensions commonly highlighted as possibly challenging their internal validity, including self-sorting, differences in response times, concentration and distraction, or differences in experimental instructions and payment methods, together with their credibility.

This testbed comparison shows that online elicitation of preferences is internally valid, according to the additional controls of our design. In particular we find that the qualitative patterns of preferences elicited in the lab are often indistinguishable from those elicited online, whether in terms of treatment effects, point differences or behavioral variance. We do find, however, that it is relatively more difficult to collect good quality data over the Internet – as shown by the increase in the number of inconsistencies in the risk aversion elicitation task. However, it should be possible to compensate for this extra noise in the data by leveraging the Internet to recruit larger samples. Last, we obtain some interesting counterintuitive results as regards social preferences exhibited online. Irrespective of whether the point differences are statistically significant or not, our results indicate that when compared to subjects allocated to the laboratory condition, other-regarding behavior from subjects in the Internet condition is never weaker – sometimes stronger. Those results are at odds with what social distance theory and common wisdom predict, given that the Internet is often characterized as an environment where anonymity is more stringent. As the online environment arguably relies more on trust to achieve trade and contract enforcement, we suggest that such habits may outperform the effect of increased social distance.

These findings are important to the growing community of researchers interested in using the Internet to run large-scale social experiments online and relating their results to the established laboratory literature. Provided that enough care is taken over specific aspects of the design, Internet-based experimental inference should be considered reliable, and the results obtained from online experiments can be compared to those obtained in the lab. These results are also potentially important for social scientists wishing to use social experiments to research the Internet as a field.

Our study raises several unanswered questions. First, we apply our methodology to the elicitation of social preferences – because there were strong reasons to doubt the parallelism between the two

fields – but many other dimensions of preferences or strategic decision-making could vary between the two environments. Second, while our design appears to be adequate to guarantee the internal validity of the preferences elicited over the Internet, our experiment was not designed to differentiate the specific dimensions that were most crucial to achieving this outcome. This is an important issue to investigate in the future, as our results have shown that some design choices compatible with online experimentations are not neutral to behavior. Last, insofar as we do observe some differences in revealed social preferences between the two elicitation fields, we are unable to conclude which of the two measures is closer to actual economic behavior. Actual differences in revealed preferences depending on the field of decision elicitation, and which field scholars should trust more, warrants a more systematic investigation which we leave open for future research.

References

Akerlof, G.A., 1997. Social Distance and Social Decisions. *Econometrica*, 65(5), pp.1005–1027.

Amir, O., Rand, D.G. & Gal, Y.K., 2012. Economic Games on the Internet: The Effect of $1 Stakes. *PLoS ONE*, 7(2), pp.1–4.

Anderhub, V., Müller, R. & Schmidt, C., 2001. Design and evaluation of an economic experiment via the Internet. *Journal of Economic Behavior & Organization*, 46(2), pp.227–247.

Bainbridge, W.S., 2007. The Scientific Research Potential of Virtual Worlds. *Science*, 317(5837), pp.472–476.

Bland, J.M. & Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. *BMJ: British Medical Journal*, 310(6973), p.170.

Charness, G., Haruvy, E. & Sonsino, D., 2007. Social distance and reciprocity: An Internet experiment. *Journal of Economic Behavior & Organization*, 63(1), pp.88–103.

Chesney, T., Chuah, S.-H. & Hoffmann, R., 2009. Virtual world experimentation: An exploratory study. *Journal of Economic Behavior & Organization*, 72(1), pp.618–635.

Cooper, D.J. & Saral, K.J., 2013. Entrepreneurship and team participation: An experimental study. *European Economic Review*, 59, pp.126–140.

Dohmen, T. et al., 2011. Individual Risk Attitudes: Measurement, Determinants and Behavioral Consequences. *Journal of the European Economic Association*, 9(3), pp.522–550.

Eckel, C.C. & Wilson, R.K., 2006. Internet cautions: Experimental games with internet partners. *Experimental Economics*, 9(1), pp.53–66.

Fehr, E. & Camerer, C.F., 2004. Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists. *Foundations of Human Sociality*, 1(9), pp.55–96.

Fiedler, M. & Haruvy, E., 2009. The lab versus the virtual lab and virtual field—An experimental investigation of trust games with communication. *Journal of Economic Behavior & Organization*, 72(2), pp.716–724.

Fiedler, M., Haruvy, E. & Li, S.X., 2011. Social distance in a virtual world experiment. *Games and Economic Behavior*, 72(2), pp.400–426.

Fischbacher, U., Gächter, S. & Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3), pp.397–404.

Glaeser, E.L. et al., 2000. Measuring Trust. *The Quarterly Journal of Economics*, 115(3), pp.811–846.

Greif, A., 2006. *Institutions And The Path to the Modern Economy: Lessons from Medieval Trade*, Cambridge University Press.

Greiner, B., 2004. *An Online Recruitment System for Economic Experiments*, University Library of Munich, Germany.

Henrich, J., Heine, S.J. & Norenzayan, A., 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), pp.61–83.

Hoffman, E., McCabe, K. & Vernon L. Smith, 1996. Social Distance and Other-Regarding Behavior in Dictator Games. *The American Economic Review*, 86(3), pp.653–660.

Hoffman, M. & Morgan, J., 2011. Who's Naughty? Who's Nice? Social Preferences in Online Industries. *UC Berkeley Working Paper*.

Holt, C.A. & Laury, S.K., 2002. Risk Aversion and Incentive Effects. *The American Economic Review*, 92(5), pp.1644–1655.

Horton, J.J., Rand, D.G. & Zeckhauser, R.J., 2011. The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3), pp.399–425.

Kahneman, D., 2003. Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, 93(5), pp.1449–1475.

Lotito, G., Migheli, M. & Ortona, G., 2013. Is cooperation instinctive? Evidence from the response times in a public goods game. *Journal of Bioeconomics*, 15(2), pp.123–133.

Piovesan, M. & Wengström, E., 2009. Fast or fair? A study of response times. *Economics Letters*, 105(2), pp.193–196.

Rand, D.G., Greene, J.D. & Nowak, M.A., 2012. Spontaneous giving and calculated greed. *Nature*, 489(7416), pp.427–430.

Resnick, P. et al., 2006. The value of reputation on eBay: A controlled experiment. *Experimental Economics*, 9(2), pp.79–101.

Rubinstein, A., 2007. Instinctive and Cognitive Reasoning: A Study of Response Times. *The Economic Journal*, 117(523), pp.1243–1259.

Selten, Reinhard, 1967. Die strategiemethode zur erforschung des eingeschrankt rationalen verhaltens im rahmen eines oligopol experiments. In Sauermann, H, ed. *Beitrage zur Experimentellen Wirtschaftsforschung*. Tubingen: J.C.B. Mohr, pp. 136–168.

Shavit, T., Sonsino, D. & Benzion, U., 2001. A comparative study of lotteries-evaluation in class and on the Web. *Journal of Economic Psychology*, 22(4), pp.483–491.