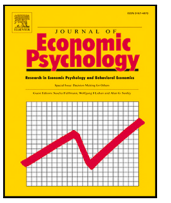
Journal of Economic Psychology 99 (2023) 102666

Contents lists available at ScienceDirect

Journal of Economic Psychology

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



App-based experiments☆

Paolo Pin [a,b], Tiziano Rotesi [c,*]

[a] Department of Economics and Statistics, Università di Siena, Italy
[b] BIDSA, Università Bocconi, Italy
[c] Department of Economics, Université de Lausanne, Switzerland

ARTICLE INFO

Dataset link: https://doi.org/10.17632/grw74bp3p3.1

JEL classification:
C90
C91

Keywords:
Experimental economics
Laboratory experiments
Smartphone applications

ABSTRACT

We elicit and compare behaviors in the laboratory and on a smartphone application that we developed for this study. Our participant pool consists of university students who are subjected to identical incentives and selection criteria. Behavior is similar across samples in measures of attitudes towards risk, effort, cognitive ability, strategic reasoning, trust, and lying aversion. Additionally, participants show comparable beliefs about the actions of the other players. We also identify certain quantitative differences between the two groups. Specifically, subjects using the app donate more in the dictator game, are faster, and show less consistency. These findings show the potential of using smartphone applications to organize experiments, and emphasize the importance of a clear and simple interface in this environment.

1. Introduction

Given the pervasive nature of smartphones and the success of companies whose business model primarily revolves around mobile apps, smartphone applications have emerged as a setting where individuals engage in critical decisions, including those related to investment and consumption. Moreover, as users tend to interact with their smartphones multiple times a day and across various locations, these devices serve as a prominent tool for collecting user data. For these reasons, the number of researchers relying on apps to organize their experiments or data collection efforts is increasing, as documented by Zhang, Calabrese, Ding, Liu, and Zhang (2018). Nonetheless, the impact of this design choice on participants' behavior remains ambiguous, particularly in terms of how experimental results deviate from those obtained in a laboratory setting. Do participants exhibit different behaviors in smartphone app-based experiments compared to those conducted in a laboratory setting?

To address this question, we set up an experiment to compare the behavior of university students in two distinct environments: a smartphone app and a laboratory setting. Participants performed a series of tasks designed to elicit a range of behavioral traits, including cognitive ability, risk aversion, strategic behavior, and trust. We randomly assigned participants to complete these tasks either in the laboratory – where students responded using desktop computers in quiet rooms with researchers present – or using *App Lab*, an app we developed for this study. Notably, we refrained from imposing any particular location constraints on the students utilizing the app, as our objective was to assess behavior in an environment that genuinely reflects their everyday smartphone usage.

☆ We thank Carlos Al 'os–Ferrer, three anonymous referees, as well as Andrea Albertazzi, Maria Bigoni, Stefania Bortolotti, Marco Casari, Francesco Feri, Friederike Mengel, Luca Merlino, Salvatore Nunnari, Marco Piovesan, David Smerdon. We also thank seminar participants at the University of Verona for their comments and suggestions. We thank our research assistants Nicoló Cardana, Claudia Marangon, and Ludovica Mosillo for their superb work. We gratefully acknowledge funding from the Italian Ministry of Education Progetti di Rilevante Interesse Nazionale (PRIN) grants 2015592CTH and 2017ELHNNJ. The data and code for this project can be found at this link https://doi.org/10.17632/grw74bp3p3.1.

* Corresponding author.
E-mail addresses: paolo.pin@unisi.it (P. Pin), tiziano.rotesi@unil.ch (T. Rotesi).

https://doi.org/10.1016/j.joep.2023.102666
Received 23 August 2022; Received in revised form 7 August 2023; Accepted 8 August 2023
Available online 19 August 2023

To isolate the effect of the platform and avoid selection into one of the two treatments, we designed the recruitment of subjects in two steps. In the first step, all participants downloaded the app and tested whether it worked on their devices by answering a short preliminary survey. In the second step, they reserved a time slot to come to the laboratory. On the day prior to the experiment, we selected a group of participants and asked them to participate using the app instead of coming to the laboratory. We kept the same time slots for the sessions and the same economic incentives. Also, both groups were paid through gift cards a few days after the experiment using the same procedure. We validate our randomization by comparing the answers collected in the preliminary survey: the two groups were balanced across all the observable variables. Furthermore, we do not find evidence of differential attrition between the two groups. Participants had the same probability of participating in the main experiment, regardless of whether their participation necessitated visiting the laboratory or using the app.

Our findings reveal similar behavioral patterns in both the laboratory and app settings across a majority of tasks. We observe no statistically significant differences in the performance of cognitive tasks, nor in measures of attitudes towards risk, strategic reasoning, trust, and lying aversion. However, when we evaluate other-regarding preferences, we observe that subjects using the app seem to be more generous in the dictator game. Differences between the two samples also manifest in areas related to time management and consistency during the experiment. Participants in the laboratory tend to be slower and experience timeouts – in the sense that they hit an exogenously imposed time limit without submitting any response – more frequently than their colleagues using the app. Likewise, they tend to be more consistent in tasks where they face multiple choices at once, as illustrated by a lower tendency of exhibiting reverse switching in a choice list.

Our findings show how the two environments of the laboratory and smartphone applications can yield comparable results. We believe that the evidence provided in this paper is promising for studies that explore the synergy between these two settings. Furthermore, these results suggest areas where researchers should focus their attention if they want to compare behaviors between the laboratory and smartphone apps.

The observed disparities in timeout occurrences between laboratory and app settings in this study suggest potential sources of bias. As such, tasks where participant response speed is critical warrant meticulous consideration. Finally, our findings highlight the importance of having a simple interface to reduce the likelihood of errors, particularly in the context of a mobile app.

1.1. Related literature

The growing interest in the use of apps for organizing experiments, as reported by Zhang et al. (2018), has stimulated researchers to explore the potential of this approach and to compare participant behavior in both laboratory and smartphone app contexts. Li, Lin, Kong, Wang and Duffy (2021) detail the results of two extensive experiments conducted using a platform named MobLab, which is accessible via laptops, smartphones, or tablets. The initial experiment took place in an auditorium in 2019, with a follow-up in 2020 that repeated the same set of tasks remotely. The authors compare their results with data from a set of laboratory experiments already published by other researchers and find the same patterns of behavior. Another paper introducing a platform capable of running experiments on smartphones is Giamattei and Lambsdorff (2019). Our work builds upon these studies by explicitly comparing a traditional laboratory setting with a mobile app environment in an experimental design that minimizes selection bias into either of the two conditions.

We also contribute to the literature comparing online and laboratory experiments. Numerous methodological experiments have been conducted to compare these two environments. Early research employs well-established tasks from the behavioral economics literature to investigate whether results can be replicated with subjects interacting online (Anderhub, Müller, & Schmidt, 2001; Fiedler & Haruvy, 2009; Shavit, Sonsino, & Benzion, 2001). More recently, there has been a shift towards comparisons implemented with between-participant designs, where subjects are recruited with the same procedure and then are randomly assigned either to the laboratory or to the online platform (Bader, Baumeister, Berger, & Keuschnigg, 2021; Buso et al., 2021; Hanaki et al., 2022; Hergueux & Jacquemet, 2015; Li, Leider, Beil & Duenyas, 2021; Ozono & Nakama, 2022; Prissé & Jorrat, 2022; Schmelz & Ziegelmeyer, 2020). With respect to them, we compare behavior in the app with behavior in the laboratory.

Finally, Gillen, Snowberg, and Yariv (2019) and Snowberg and Yariv (2021) are two recent papers that compare three different pools: (i) subjects recruited online with Mechanical Turk, (ii) a representative sample of the US population, and (iii) university students, where the latter are both in the laboratory and online. For the student sample, these papers use a within-participant design, where subjects repeat the same tasks twice in different environments. Our paper complements these studies on three important dimensions. First, we specifically force our subjects to answer using their smartphones and not any device connected to the internet. Additionally, as we use a between-participants design, our subjects experience these tasks only once and not in a fixed order. We can therefore rule out any role played by experience. Third, by randomizing the environment, we can address potential issues related to the time and duration of the experiment.

In the next section, we describe the setup of our experiment and the set of elicitations we consider. Section 3 reports and discusses the main results. Finally, Section 4 concludes.

2. Setup of the experiment

To minimize selection bias, we implemented the experiment using a two-step procedure. In the first step, we asked every participant to install the app and answer a preliminary survey, ensuring that they were able to install and use the app. Then, we divided the subjects into two groups and administered the same survey covering cognitive abilities, effort, attitudes towards risk, strategic reasoning, other-regarding preferences, trust, and lying aversion. In this section, we provide more details about the recruitment process, the randomization method, and the set of elicitations.

Table 1
Balance table.

	Lab (1)	App (2)	Difference (3)	*p*-value (4)
Preliminary survey				
Gender (1 = Female)	0.451	0.477	0.026	0.701
Experience (1 = Yes)	0.876	0.838	−0.038	0.466
Experience (n of times)	5.755	4.874	−0.881	0.145
Location: public venue	0.027	0.023	−0.003	1.000
Location: home	0.655	0.669	0.014	0.892
Location: other	0.018	0.062	0.044	0.111
Location: campus	0.257	0.208	−0.049	0.445
Location: public transportation	0.044	0.038	−0.006	1.000
Have a desktop/laptop with you	0.690	0.746	0.056	0.390
Would prefer desktop/laptop	0.230	0.223	−0.007	1.000
Phone: every 10 min	0.044	0.038	−0.006	1.000
Phone: btw every 10 and 30 min	0.265	0.269	0.004	1.000
Phone: btw every 30 min and 1 h	0.451	0.492	0.041	0.607
Phone: btw every 1 and 2 h	0.230	0.192	−0.038	0.529
Phone: less than every 2 h	0.009	0.008	−0.001	1.000
Correctly counted 1s	0.621	0.613	−0.008	1.000
Attrition				
Participated	0.965	0.946	−0.018	0.551
Observations	113	130		

Notes: The table presents statistics collected in the preliminary survey and compares them across the two groups. Column (1) refers to participants who were then invited to take the full survey at the BELSS Lab at Bocconi University. Column (2) refers to participants who were asked to respond using *App Lab*. Column (3) reports the difference between columns (1) and (2), while column (4) provides p-values for the test of equality between the Lab and App groups. The variable *Experience* refers to previous experiments in which the respondent participated. The variable *Location* indicates the location of the respondent at the time of answering the preliminary survey. *Have a desktop/laptop with you* specifies whether the respondent had access to a desktop or laptop computer when completing the survey, and *Would prefer desktop/laptop* has a value of 1 if the respondent would have preferred to answer the survey on a device other than a smartphone. *Phone* gives a self-assessment by the respondent about how frequently they interact with their phone. *Correctly counted 1s* identifies whether the respondent correctly counted the number of 1s in an 8 × 8 matrix. For indicator variables, we employ Fisher's exact test.

2.1. Preliminary survey and randomization

The experiment took place in Spring 2019. Towards the end of March, we invited students from the lists of the BELSS Lab at Bocconi University. Then, we asked those who had accepted our invitation to download *App Lab*, a smartphone app specifically designed and programmed for this experiment.[1] Participants utilized the app to complete a brief survey, where they indicated their preferred dates to partake in an incentivized experiment. They were aware that this experiment could be conducted either online or in the laboratory. In total, 243 subjects completed the survey and were subsequently invited to participate in the experiment.

Taking into account their date preferences, we arranged the subjects into 8 sessions and informed them of their allocated time slot in early April. Subsequently, we randomly assigned subjects to either the laboratory (treatment *lab*) or *App Lab* (treatment *app*). Finally, two days before the session, we sent participants an email specifying whether they should come in person to the laboratory or complete the tasks remotely using *App Lab*.

A total of 243 students were invited, and 232 participated in the experiment. Attrition – defined as the rate at which enrolled students actually participated – was evenly balanced between the lab and the app treatments. We had 110 students in the lab and 122 in the app. Each of the 16 sessions had an average of 14.5 subjects. Given this sample size and the allocation ratio between the two treatments (0.9), we can detect an effect size (Cohen's d) as small as 0.38 with 80% power and $\alpha = 0.05$.

Table 1 presents the variables obtained from the preliminary survey. It illustrates that they are balanced between the two treatments. Additionally, the last row under *Attrition* indicates the absence of any differential attrition.

Reviewing the answers provided in the preliminary survey, we find that the majority of respondents had already participated in behavioral experiments. Most of them responded from either their home or the university campus. They also considered it natural to respond to the survey with a smartphone, even if they had a desktop or laptop computer at their disposal. In response to the question about how often they check their phone throughout the day, most respondents stated at least once an hour. Finally, the majority of respondents correctly counted the number of ones in a task resembling those used in the main survey.

2.2. The main survey

The main survey, which was identical for both groups, was programmed using oTree (Chen, Schonger, & Wickens, 2016). Participants in the lab completed the tasks on a desktop computer in a traditional setting for behavioral experiments. Conversely, participants using the app responded via their smartphones, having the freedom to choose their location.

[1] At the time of the experiment, *App Lab* was available on both the Play Store and the App Store. Further information regarding the app's functionality is provided in the Online Appendix A.

The economic incentives were identical for all participants, including a show-up fee of €4 and an additional reward ranging from €1.1 to €20 (with an average of €10.7). The potential payoffs were displayed at the beginning of each task. Outcomes of tasks based on interactions among subjects were computed by matching subjects ex–post. Crucially, participants were not informed of the final reward for each task. Instead, they were to receive payment for two randomly selected tasks, unaware of which tasks had been chosen. All payments were made as Amazon gift cards, which were received a few days after the experiment to standardize the timing of payment across treatments. We reminded participants that Amazon allows customers to register after receiving a gift card and highlighted that, in Italy, it is not possible to transfer the gift card to someone else.

In the introductory explanation, we stated that we would record everything that the participants saw and typed on the screen. However, we did not explicitly state that their interactions were solely with individuals from the same session and treatment arm. Subjects completed 18 main tasks, along with 9 additional ancillary tasks related to some of the primary ones. The ancillary tasks were also incentivized. The specific scoring rule used in each case was detailed in the experimental instructions, available in Online Appendix C.

The sequence of the primary tasks was randomly determined for each session, with the ancillary tasks consistently following their related main tasks. All tasks had set time limits, typically two to three minutes each, with longer durations allocated to tasks assessing effort and focus. Participants from both groups could view a countdown, indicating the remaining time for each task, on their screen. Upon completing a task, participants had to wait until all other participants in the same session either completed the task or exhausted their allotted time before they could move on to the next task. On average, sessions lasted about 50 min.

2.3. Description of elicitations

We followed Hergueux and Jacquemet (2015) and Snowberg and Yariv (2021) among others in considering a broad set of elicitations. With respect to them, we included additional tasks involving strategic interaction and two tasks related to *lying aversion*.

Some of the tasks detailed below were accompanied by ancillary questions designed to measure beliefs. In these questions, we asked respondents to focus on the behavior of unrelated individuals (Folli & Wolff, 2022), and we incentivized participants using either a linear scoring rule or the *frequency method* (Charness, Gneezy, & Rasocha, 2021). We opted for these methods due to their simplicity, as opposed to other, more complex techniques found in the literature (Burdea & Woon, 2022; Palfrey & Wang, 2009).

1. Cognitive Tasks. We used two types of cognitive tasks, similar to Snowberg and Yariv (2021). Each of these tasks was followed by ancillary questions in which subjects estimated the number of correct answers they provided and predicted how many out of ten randomly selected subjects performed better than themselves. These tasks were paid €1 each.

Raven's Matrices. Subjects were required to complete five Raven's matrices, where the task was to identify geometric regularities within a set of 3×3 matrices. Each question had a time limit of 30 s. For every correct answer, subjects received €1.

Cognitive Reflection Test (CRT). These questions were adapted from the original cognitive questions utilized by Frederick (2005), tailored to suit the context of students from Bocconi University. Subjects had a 30-second window to answer each question, with the potential to earn €1 for every correct response.

2. Counting 1s. The aim of this task was to measure effort. Subjects were presented with up to 38 square matrices of size 8×8, filled with either 0's or 1's. They were given a total of 5 min to complete the task. A similar task was used in Abeler, Falk, Goette, and Huffman (2011). This task was followed by two ancillary tasks, analogous to those explained in the previous point.

3. Risk Elicitation. We used three different tasks to elicit attitudes towards risk.

Risky Projects. This task was based on cases a and b in Gneezy and Potters (1997). In both the *Risky Project 50%* and *Risky Project 35%* tasks, subjects could choose how much to invest in a risky project. The former had a probability of success of 50% and paid 2.5 times the amount invested (up to €4), while the latter succeeded with probability 35% and paid 3 times the sum invested (up to €3.5).

Gamble Choice. This task was based on Eckel and Grossman (2002), where subjects had to pick one out of 6 gambles, listed in order of increasing risk.

Safe Gambles. This task was based on Holt and Laury (2002). Subjects had to compare a series of lottery pairs. In each pair, one lottery was relatively safe (yielding either €4 or €5, with probabilities p and $1 - p$) while the other was relatively risky (providing either €0.25 or €9.6, with the same probabilities). Subjects were asked to pick one lottery from each pair. The complete list of choices is in the Online Appendix, where we report the full survey.

When analyzing the results for this task, we follow the methodology described in Holt and Laury (2002) and focus on the total number of times the participant chooses Gamble A, which represents the safest option.

4. Strategic Reasoning. We employed three distinct tasks to assess strategic reasoning.

2X2 Games. In this task, subjects sequentially participated in a *prisoners' dilemma*, a *stag hunt*, and a *hawk–dove* game. The payoffs corresponding to these games can be found in Table 2. Subjects were informed that their payoffs would be determined by pairing them at random with another participant. Each of these games was succeeded by an ancillary task where subjects were asked to estimate the proportion of other players who selected action A.

Beauty Contest. The beauty contest task was a variation of the classical one–shot game presented in Nagel (1995). Subjects were informed that they were paired with another person in the session and that they could win €10 by guessing the closest number to $\frac{2}{3}$

Table 2
Payoffs in euros for the 2 × 2 games.

Prisoners' dilemma		
	A	*B*
A	6,6	0,10
B	10,0	2,2

Stag hunt		
	A	*B*
A	10,10	0,5
B	5,0	2,2

Hawk–dove		
	A	*B*
A	5,5	5,10
B	10,5	2,2

of the mean (€5 in case of a tie). We adopted this explicit pairing to focus subjects' attention on the fact that they were interacting directly with other subjects. Moreover, the strategic implications are simpler when pairing, as playing 0 is the unique dominant action, without the need to use iterated elimination of strictly dominated strategies. Subjects had 2 min to decide.

Money Request Game. In this task, participants were instructed to choose a whole number between 0 and 10. Players were paired, and the monetary payoff was determined as follows. If players i and j selected numbers a_i and a_j respectively, their earnings were $€a_i \cdot 0.50$ and $€a_j \cdot 0.50$. However, if a_i equaled $a_j - 1$, player i received an additional €5 (and the same symmetrically for player j).

Compared to the beauty contest, the money request game shares a similar structure but requires more strategic reasoning as the iterated elimination of strictly dominated strategies does not suffice to predict an equilibrium. This game has a unique Nash equilibrium in mixed strategies, with support given by the set $\{5, 6, 7, 8, 9, 10\}$, and Arad and Rubinstein (2012)[2] used it as a test for measuring how many iterated steps of strategic reasoning players do, in the spirit of level–k reasoning (Stahl & Wilson, 1994, 1995). Subjects had 2 min to make their choice.

5. Other-Regarding Preferences. We used two different tasks to elicit other-regarding preferences.

Dictator Game. Participants were given 2 min to determine how much of a €10 endowment they wished to retain for themselves and how much to allocate to a randomly selected peer. The marginal conversion rate was decreasing in generosity, as shown in Fig. 1. This concept, inspired by Offerman, Sonnemans, and Schram 1996, aimed to reduce the occurrence of extreme outcomes of full selfishness.

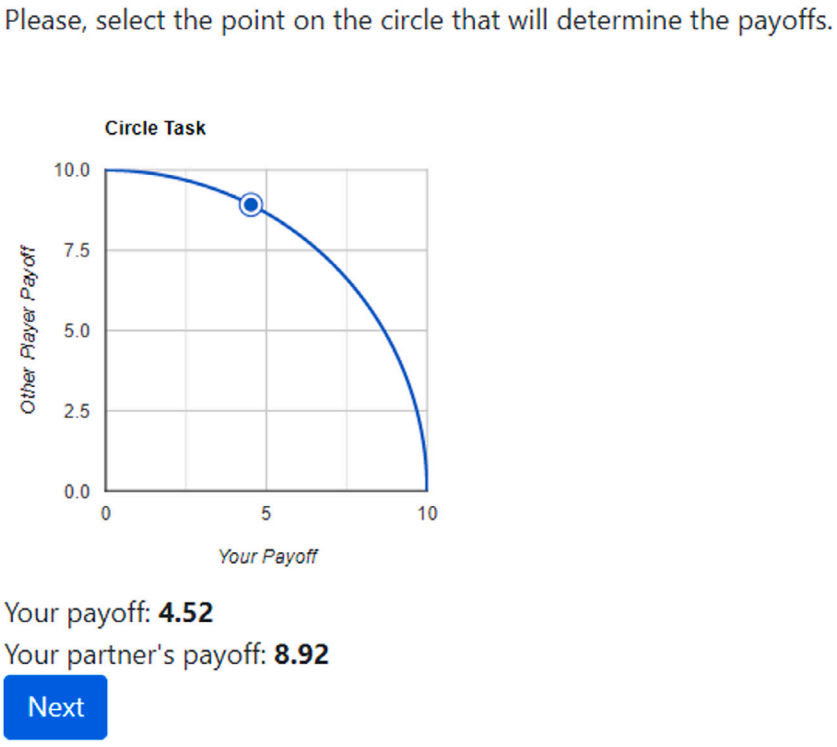


Fig. 1. Screenshot of the dictator game.

[2] In the original paper the range is between 0 and 20, but the qualitative analysis of strategies and equilibria does not change. We chose a smaller range to simplify the comprehension.

Ultimatum Game. In this task, subjects played a standard ultimatum game (Güth, Schmittberger, & Schwarze, 1982), with a total amount of €10. For this game, we used the strategy method and we asked subjects what they would do as Player A and as Player B depending on the 11 possible choices of Player A. Subjects were allocated 1 min to determine their decision as Player A, and an additional minute to decide on their action as Player B, selecting from an ordered table featuring 11 *yes/no* options. This game was followed by two ancillary questions: (i) to guess the modal action of all other subjects when playing as A; and (ii) to estimate the percentage of other participants who, when playing as B, would reject €1.

6. Trust. In this task, subjects played a trust game (Bonowski & Minnameier, 2022; Holm & Nystedt, 2008; Johnson & Mislin, 2011; Sofianos, 2022). Both players, A and B, started with 5 tokens each. Player A had to decide on the number of tokens (a discrete value $a \in 0, \dots, 5$) to send to Player B. Subsequently, Player B would receive $3 \cdot a$ tokens and then determine the number to return to A (a discrete value $b \in 0, \dots, 3 \cdot a + 5$). Each token was valued at €0.70. For this game, we used strategy elicitation and asked our subjects to specify their actions as both Player A and Player B, considering the 6 potential actions of Player A. This game was followed by an ancillary task, in which subjects were asked to guess the average transfer made by the other players when acting as Player A.

7. Lying Aversion. The last set of tasks consisted of three tasks in which subjects could misreport some of the information they received, possibly to their advantage. These tasks are related to recent literature on *lying* and *cheating* (Chua, Chang, & Riambau, 2022; Fischbacher & Föllmi-Heusi, 2013; Gneezy, Rockenbach, & Serra-Garcia, 2013; Rosenbaum, Billinger, & Stieglitz, 2014), but we cannot exclude the possibility that most cases of misreporting in our setting depend solely on involuntary errors if the tasks have been perceived as cognitive rather than honesty-related. All three tasks employed a digital random number generator, allowing participants to virtually roll a die. It is important to mention that our initial instructions explicitly conveyed that we would capture all on-screen activities during the experiment, so that participants were aware that lying was observable (Hermann & Brenig, 2022).

The dice sum: Subjects had one minute to throw a die 5 times and report the sum of the 5 throws. We paid them €0.30 for every point they reported, capping the potential reward at €9 if they reported a total of 30. The reward was based on the report and was independent of the actual outcome of the dice.

The dice, count 6. The count 6's task was similar to the previous one, but in this case, the subjects had 1 minute to throw the die 10 times and report how many 6's they had obtained, earning €1 for each reported 6.

Report Game This task was adapted from Gneezy et al. (2013). It shares similarities with the previous two tasks, but in this case, there was also a strategic interaction, it involved trust, and misreporting had a negative externality on others (Innes, 2022).

In the report game, two players interacted sequentially. Player A rolled a die and then reported an outcome $a \in 1, \dots, 6$, earning a reward equivalent to a euros. Then, Player B had to decide whether to trust Player A's report. If Player B believed the report and it turned out to be accurate, she would earn €5. If she believed it and it was false, she would earn nothing. Alternatively, if she did not trust Player A and indicated so, she would receive €1.50.

For this game, we used the strategy method. Thus, subjects specified their strategy for both roles. As Player A, they had 1 minute to write what they would declare, based on the 6 outcomes of the die. Meanwhile, as Player B, they had 1 minute to choose whether they would trust or not the report, depending on the 6 possible reports.

2.4. Sample size

As mentioned in Section 2.1, the full sample consists of 232 participants: 110 in the laboratory and 122 using the app. Table 3 shows sample sizes for each task.

There are several reasons why we may observe sample sizes that are smaller than the full sample. One common cause is timeouts, which occurred when a participant took too long to provide a response. Additionally, in the task of Counting 1s, we had a lower number of observations due to some participants disconnecting during the task. This was a result of technical issues with the Heroku server used in the first session. However, the problem was resolved in subsequent sessions.

3. Results

This section presents the results from the comparison between the two groups, namely *app* and *lab*. Upon analyzing the elicitations outlined earlier, we note minor variations across all tasks, with the exception of the dictator game. The same applies when studying beliefs, which are measured using the ancillary tasks and are reported in the Online Appendix, Section B.1. However, there are differences in behavior, as we find evidence that subjects in the app group are faster and tend to be more inconsistent.

3.1. Similarities between App and Lab

Table 4 displays the primary outcomes spanning the entire array of elicitations. For each previously discussed task, the table presents the mean results for both the lab and app groups, as well as the difference between the two. Additionally, the table lists the p-value for each comparison, both unadjusted for multiple hypothesis testing and adjusted according to the methodology outlined in Romano and Wolf (2016). Finally, it reports the p-value for the test of equality of group variances.[3]

[3] We thank an anonymous referee for suggesting this.

Table 3
Sample size by treatment.

	Treatment	
	Lab (1)	App (2)
Risky Project 50%	110	122
Risky Project 35%	110	122
Gamble Choice	110	122
Safe Gambles	110	122
Counting 1s, Correct Matrices	103	101
Raven, Correct Answers	110	122
CRT, Correct Answers	110	122
Beauty Contest Guess	108	121
Money Request Game	110	122
Number honest messages	92	117
Dictator Game, Amount kept	110	110
Ultimatum Game, Amount kept	109	121
Ultimatum Game, Min accepted	110	121
Trust Game, Amount sent	110	122
Prisoner's dilemma	110	122
Stag hunt	110	122
Hawk-Dove	110	122
The Dice sum	107	118
Count 6	110	122

Notes: Column (1) refers to participants in the BELSS Lab at Bocconi University. Column (2) refers to participants who answered using App Lab.

The main pattern that emerges from the table is that the average measures are similar across the two groups. This is true across all elicitations, with the sole exception being the dictator giving task. A similar observation can be made from Table A2 located in the Online Appendix, where it is evident that beliefs about oneself or others are comparable between the two groups.

To test the hypothesis that behavior varies under the two conditions, we also conducted an additional test, which is not reported in the table. Using the laboratory as a reference point and considering its mean and standard deviation (SD), we introduced an alternative hypothesis regarding the difference, denoted by Δ, between the means of the two treatments. Specifically, we considered H1: $|\Delta| = 0.3$ SD, a value that sits between what is generally considered a small and medium effect (Cohen, 1977). Beta errors calculated with this method exceed 0.2 only for the measures "Dictator Game, Amount kept" and "Trust Game, Amount sent". Thus, we conclude that the data do not support the alternative hypothesis of a significant difference between the app and lab results, a finding that is reassuring given the scope of our paper.

As previously mentioned, the only exception, albeit only when considering unadjusted p-values, is observed in the dictator game. In this case, subjects using the app appear to be more altruistic, keeping a lower amount for themselves on average (€8.31 in the app, €8.70 in the lab). We will revisit this point in the following section when discussing these findings.

Upon examining group variances, we find a comparable degree of variability in the data for both groups. Statistically significant differences only emerge in two tasks requiring strategic reasoning, namely the "Beauty Contest, Guess" and the "Money Request Game", as well as in the beliefs regarding others' actions in the Hawk-Dove game. In all these instances, variability is more evident in the laboratory context.

3.2. Time management and consistency

One dimension in which there is a difference between the two treatments is time management. Table 5 shows that in tasks where time constraints are more binding, subjects using the app tend to experience fewer timeouts.

First, we examine how subjects perform in the Raven's Matrices and in the Cognitive Reflection Test, where each question had a 30-s time limit. In the lab, subjects are more likely to run out of time without providing an answer, resulting in a higher number of timeouts. In contrast, subjects using the app more often provide an incorrect response before time runs out. The number of timeouts is also higher in the Report Game, where subjects face multiple choices at once due to the strategy method. When examining the time spent on each matrix in the Counting 1s task, we find that subjects using the app, on average, spend less time and therefore attempt to answer a higher number of matrices.

Regarding this, we must acknowledge the potential impact of time constraints. We decided to impose a time limit on all the questions to prevent the possibility of subjects using the app from losing their focus on the experiment. However, this approach also leads to the loss of potentially valuable information. Specifically, we may miss insights into whether subjects in the lab could have increased their accuracy rate if they had been granted more time.

Table 5 also reveals differences in inconsistent behavior, which we identify by examining two tasks and making minimal assumptions about rationality.

In the Safe Gambles task, we assume a minimal level of rationality under risk and label respondents as consistent only if they do not switch from the riskier Gamble B to the safer Gamble A as the probability of winning increases. Using this definition, the fraction of respondents showing this inconsistency is higher in the app (22%) than in the lab (7%).

Table 4
Similarity between App and Lab.

	Treatment		Mean		Variance
	Lab (1)	App (2)	Difference (3)	*p*-value (4)	*p*-value (5)
Risky Project 50%	243.973	250.811	6.839	0.615	0.905
(out of 400)	(10.140)	(9.106)	(13.589)	[1.000]	
Risky Project 35%	148.300	148.205	−0.095	0.994	0.667
(out of 350)	(9.374)	(8.174)	(12.382)	[1.000]	
Gamble Choice	3.373	3.164	−0.209	0.305	0.773
(1 to 6)	(0.141)	(0.145)	(0.203)	[1.000]	
Safe Gambles	5.236	5.115	−0.122	0.596	0.710
(0 to 9)	(0.166)	(0.158)	(0.229)	[1.000]	
Counting 1s, Correct Matrices	15.000	15.762	0.762	0.287	0.137
(number)	(0.527)	(0.480)	(0.714)	[0.990]	
Raven, Correct Answers	2.218	2.213	−0.005	0.976	0.610
(out of 5)	(0.123)	(0.117)	(0.170)	[1.000]	
CRT, Correct Answers	1.673	1.885	0.213	0.191	0.138
(out of 5)	(0.120)	(0.109)	(0.162)	[0.969]	
Beauty Contest Guess	45.722	47.099	1.377	0.690	0.038
(out of 100)	(2.781)	(2.115)	(3.452)	[1.000]	
Money Request Game	6.709	7.000	0.291	0.371	0.083
(out of 10)	(0.256)	(0.204)	(0.324)	[1.000]	
Number honest messages	2.739	3.120	0.381	0.201	0.557
(player A)	(0.217)	(0.200)	(0.296)	[0.969]	
Dictator Game, Amount kept	8.701	8.306	−0.395	0.026	0.461
(out of 10)	(0.121)	(0.128)	(0.176)	[0.367]	
Ultimatum Game, Amount kept	6.321	6.107	−0.214	0.357	0.906
(player A)	(0.171)	(0.156)	(0.232)	[1.000]	
Ultimatum Game, Min accepted	2.736	2.628	−0.108	0.642	0.796
(player B)	(0.169)	(0.160)	(0.233)	[1.000]	
Trust Game, Amount sent	1.964	2.311	0.348	0.102	0.872
(player A)	(0.152)	(0.147)	(0.212)	[0.806]	
Prisoner's dilemma	0.336	0.328	−0.008	0.891	–
(cooperate)	(0.045)	(0.043)	(0.062)	[1.000]	
Stag hunt	0.655	0.730	0.075	0.216	–
(cooperate)	(0.046)	(0.040)	(0.061)	[0.990]	
Hawk-Dove	0.609	0.590	−0.019	0.769	–
(yield)	(0.047)	(0.045)	(0.065)	[1.000]	
The Dice sum	0.850	0.856	0.005	0.908	–
(correct report)	(0.035)	(0.032)	(0.047)	[1.000]	
Count 6	0.909	0.926	0.017	0.635	–
(correct report)	(0.028)	(0.024)	(0.036)	[1.000]	

Notes: Column (1) refers to participants in the BELSS Lab at Bocconi University. Column (2) refers to participants who answered using App Lab. Column (3) reports the difference between column (1) and (2). Column (4) reports p-values for the test of equality between the Lab and App groups. When comparing continuous variables we use linear regressions. For indicator variables we use probit regressions. In brackets, p-values adjusted for multiple hypotheses testing calculated according to Romano and Wolf (2005a, 2005b, 2016) as implemented in Clarke (2021). When defining the list of hypotheses to test, we include the variables reported in Table 5 as well. Column (5) reports p-values for the Brown–Forsythe test for the equality of group variances. * p<0.05, ** p<0.01, *** p<0.001. Standard Errors in parentheses.

Similarly, in the Ultimatum Game, a subject playing as B can be defined as consistent only if they either accept all offers, reject all offers, or establish a unique switching point above which they accept offers. Under this definition, the share of inconsistent players is 17% in the app and only 4% in the lab.

4. Discussion

This paper examines and compares the behavior of subjects in both laboratory and app (*App Lab*) settings, employing a wide range of elicitations. The findings presented in this study provide compelling evidence that subjects exhibit similar behavior in both environments, indicating that smartphones have the potential to be used as reliable tools for eliciting subjects' characteristics. This applies to tasks that measure cognitive ability, effort, and risk attitudes. This similarity is also noteworthy when considering subjects' beliefs about the behavior of others and in tasks involving strategic interaction. Furthermore, there does not appear to be a difference in tasks where subjects may provide incorrect reports for monetary gain. However, we identify differences in three areas: (*i*) other-regarding preferences, (*ii*) time management during time-constrained tasks, and (*iii*) consistency.

Overall, these results confirm and extend to smartphone apps what the previous literature has found when comparing online and laboratory experiments. Specifically, Hergueux and Jacquemet (2015) found that a higher number of online respondents tend to be inconsistent, that they tend to respond faster, and that they are inclined to transfer more in the dictator game and the trust game.

Table 5
Time management and inconsistent behavior.

	Treatment		Mean	
	Lab (1)	App (2)	Difference (3)	*p*-value (4)
Counting 1s, Time	17.942	16.422	−1.519	0.038
(average n. of sec.)	(0.580)	(0.438)	(0.729)	[0.490]
Raven, Timeouts	1.364	0.762	−0.601*	0.000
(out of 5)	(0.134)	(0.087)	(0.157)	[0.010]
CRT, Timeouts	1.664	1.156	−0.508*	0.002
(out of 5)	(0.135)	(0.097)	(0.164)	[0.030]
Report Game, Timeout	0.164	0.041	−0.123*	0.001
(player A)	(0.035)	(0.018)	(0.039)	[0.030]
Safe Gambles	0.073	0.221	0.149*	0.001
(switch from B to A)	(0.025)	(0.038)	(0.046)	[0.020]
Ultimatum Game	0.036	0.174	0.137*	0.000
(refuse high and accept low)	(0.018)	(0.035)	(0.040)	[0.010]

Notes: Column (1) refers to participants in the BELSS Lab at Bocconi University. Column (2) refers to participants who answered using App Lab. Column (3) reports the difference between columns (1) and (2). Column (4) reports p-values for the test of equality between the Lab and App groups. When comparing continuous variables, we use linear regressions. For indicator variables, we use probit regressions. In brackets, p-values adjusted for multiple hypotheses testing calculated according to Romano and Wolf (2005a, 2005b, 2016) as implemented in Clarke (2021). When defining the list of hypotheses to test, we include the variables reported in Table 4 as well. * p<0.05, ** p<0.01, *** p<0.001. Standard Errors in parentheses.

One possible explanation for the observed differences could be the presence of *experimenter demand effects* (Zizzo, 2010). In the context of this experiment, this might occur because students are asked to participate in a laboratory experiment in a setting they naturally associate with their education. Therefore, subjects using the app might feel less pressure to adhere to norms typically associated with educational environments. This could cause students in the lab to pay more attention to their choices during the experiment, thereby reducing the speed of decision-making and the likelihood of making inconsistent choices. For the same reason, subjects using the app might display slightly more generosity than those in the lab, as they may feel less pressure to conform to the behavior expected of the classically selfish economic agent.

A second possible explanation for this behavioral difference could be the physical effort required to attend a laboratory setting. Specifically, while subjects do not know beforehand if they will participate via the app or in the laboratory, those who are ultimately assigned to the lab might feel they are more deserving of the incentives than those participating through their smartphones. This could account for why, particularly in the dictator game, they display behavior more aligned with selfishness. This observation could align with findings from studies such as Zitek, Jordan, Monin, and Leach (2010), where subjects who felt they had been treated unfairly demonstrated less prosocial behavior in a dictator game.

A third possible explanation focuses on the core difference between the laboratory and the app, which is the device used for the experiment and the environment in which it takes place. Although the experiment was programmed in the same way, participants using the app would view the interface on a smaller screen, potentially operate in a noisier environment, or have different perceptions of time's opportunity cost. This may have affected the time required to read the texts or the ability to submit their choices, particularly for those using smartphones with smaller screens. Indeed, we have some missing data for the app on the dictator game task, because people did not choose any point in the circle from Fig. 1 before time expired (see Section 2.4). This could possibly be due to difficulties in clicking precisely on the most selfish answer. The presence of ambient noise could also explain the higher level of generosity observed in the app. Any deviation from a purely selfish choice due to an error would inevitably lead to a shift towards increased generosity. For further discussion on this point, please refer to List (2007) and Bardsley (2008).

To explore this further, at the end of the experiment we asked subjects using the app about their location. Among the possible choices, 44 selected "University campus", while 56 selected "Home". We compared these two sub-groups using the same tests that we presented in Tables 4 and 5. The "Money Request Game", the "Trust Game", and inconsistent behaviors in the "Ultimatum Game" are the only three measures showing differences that are significantly different from zero. Interestingly, the group of respondents at home provided answers that were, on average, closer to the answers given by laboratory participants. This finding suggests that the aforementioned considerations may indeed play a role. However, it is important to note that once subjects knew they were in the App treatment, their choice of the location in which to take the experiment was endogenous, so we cannot establish causality.

These considerations are likely important in guiding the design of data collections performed using smartphone apps. In this environment, it is crucial to have a clear and simple interface and brief sessions to minimize errors. Additionally, it is beneficial to include questions aimed at assessing the quality of answers to further enhance the accuracy and reliability of the data collected (Meade & Craig, 2012; Wolff, 2019).

Finally, we cannot rule out the possibility that some of the subjects using the app were completing the experiment together, and this may have increased their willingness to be more altruistic. This could occur even though the pairing was anonymous, and we were not explicit about the fact that they were interacting only with people from the same session and treatment arm.

Future research could explore additional factors that might impact the comparison between laboratory and app experiments. The present study was limited to university students, and it remains unclear whether the results would generalize to other demographic groups. For instance, app experiments conducted with an older population, or with a less homogeneous group in terms of education,

wealth, and cognitive abilities, may yield different outcomes. Moreover, app-based experiments offer unique research opportunities, such as the ability to collect measures over an extended time frame or to conduct experiments with participants from specific geographical locations.

Data availability

https://doi.org/10.17632/grw74bp3p3.1

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (GPT 3.5 and GPT 4) in order to improve the language and readability of the paper. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Online appendix

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.joep.2023.102666.

References

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, *101*(2), 470–492.

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

Arad, A., & Rubinstein, A. (2012). The 11-20 money request game: A level-k reasoning study. *American Economic Review*, *102*(7), 3561–3573.

Bader, F., Baumeister, B., Berger, R., & Keuschnigg, M. (2021). On the transportability of laboratory results. *Sociological Methods & Research*, *50*(3), 1452–1481.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, *11*(2), 122–133.

Bonowski, T., & Minnameier, G. (2022). Morality and trust in impersonal relationships. *Journal of Economic Psychology*, *90*, Article 102513.

Burdea, V., & Woon, J. (2022). Online belief elicitation methods. *Journal of Economic Psychology*, *90*, Article 102496.

Buso, I. M., Di Cagno, D., Ferrari, L., Larocca, V., Lorè, L., Marazzi, F., et al. (2021). Lab-like findings from online experiments. *Journal of the Economic Science Association*, *7*(2), 184–193.

Charness, G., Gneezy, U., & Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behaviour and Organization*, *189*, 234–256.

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

Chua, S. L., Chang, J., & Riambau, G. (2022). Lying behavior when payoffs are shared with charity: Experimental evidence. *Journal of Economic Psychology*, *90*, 102512.

Clarke, D. (2021). RWOLF2: Stata module to calculate romano-wolf stepdown p-values for multiple hypothesis testing. Statistical Software Components, Boston College Department of Economics.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Inc..

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behaviour*, *23*(4), 281–295.

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 Behaviour and Organization*, *72*(2), 716–724.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise–an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547.

Folli, D., & Wolff, I. (2022). Biases in belief reports. *Journal of Economic Psychology*, *88*, Article 102458.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Giamattei, M., & Lambsdorff, J. G. (2019). Classex—an online tool for lab-in-the-field experiments with smartphones. *Journal of Behavioral and Experimental Finance*, *22*, 223–231.

Gillen, B., Snowberg, E., & Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the Caltech Cohort study. *Journal of Political Economy*, *127*(4), 1826–1863.

Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, *112*(2), 631–645.

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behaviour and Organization*, *93*, 293–300.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behaviour and Organization*, *3*(4), 367–388.

Hanaki, N., Hoshino, T., Kubota, K., Murtin, F., Ogaki, M., Ohtake, F., et al. (2022). Comparing data gathered in an online and a laboratory experiment using the trustlab platform. *Institute of Social and Economic Research Discussion Papers*, *1168*, 1–22.

Hergueux, J., & Jacquemet, N. (2015). Social preferences in the online laboratory: A randomized experiment. *Experimental Economics*, *18*(2), 251–283.

Hermann, D., & Brenig, M. (2022). Dishonest online: A distinction between observable and unobservable lying. *Journal of Economic Psychology*, *90*, 102489.

Holm, H., & Nystedt, P. (2008). Trust in surveys and games – A methodological contribution on the influence of money and location. *Journal of Economic Psychology*, *29*(4), 522–542.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Innes, R. (2022). Does deception raise or lower lie aversion? Experimental evidence. *Journal of Economic Psychology*, *90*, Article 102525.

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865–889.

Li, J., Leider, S., Beil, D., & Duenyas, I. (2021). Running online experiments using web-conferencing software. *Journal of the Economic Science Association*, *7*(2), 167–183.

Li, Z., Lin, P.-H., Kong, S.-Y., Wang, D., & Duffy, J. (2021). Conducting large, repeated, multi-game economic experiments using mobile platforms. *PLoS One*, *16*(4), Article e0250668.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*(3), 482–493.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437.

Nagel, R. (1995). Unraveling in Guessing games: An experimental study. *American Economic Review*, *85*(5), 1313–1326.

Offerman, T., Sonnemans, J., & Schram, A. (1996). Value orientations, expectations and voluntary contributions in public goods. *The Economic Journal*, *106*(437), 817–845.

Ozono, H., & Nakama, D. (2022). Effects of experimental situation on group cooperation and individual performance: comparing laboratory and online experiments. *PLoS One*, *17*(4), Article e0267251.
Palfrey, T. R., & Wang, S. W. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behaviour and Organization*, *71*(2), 98–109.
Prissé, B., & Jorrat, D. (2022). Lab vs online experiments: No differences. *Journal of Behavioral and Experimental Economics*, *100*, Article 101910.
Romano, J. P., & Wolf, M. (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, *100*(469), 94–108.
Romano, J. P., & Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.
Romano, J. P., & Wolf, M. (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, *113*, 38–40.
Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, *45*, 181–196.
Schmelz, K., & Ziegelmeyer, A. (2020). Reactions to (the absence of) control and workplace arrangements: experimental evidence from the internet and the laboratory. *Experimental Economics*, *23*(4), 933–960.
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), 483–491.
Snowberg, E., & Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, *111*(2), 687–719.
Sofianos, A. (2022). Self-reported & revealed trust: Experimental evidence. *Journal of Economic Psychology*, *88*, Article 102451.
Stahl, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behaviour and Organization*, *25*(3), 309–327.
Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*(1), 218–254.
Wolff, I. (2019). The reliability of questionnaires in laboratory experiments: What can we do? *Journal of Economic Psychology*, *74*, Article 102197.
Zhang, J., Calabrese, C., Ding, J., Liu, M., & Zhang, B. (2018). Advantages and challenges in using mobile apps for field experiments: A systematic review and a case study. *Mobile Media & Communication*, *6*(2), 179–196.
Zitek, E. M., Jordan, A. H., Monin, B., & Leach, F. R. (2010). Victim entitlement to behave selfishly. *Journal of Personality and Social Psychology*, *98*(2), 245.
Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*, 75–98.