Journal of Behavioral and Experimental Economics 106 (2023) 102057

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

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

On the generalizability of using mobile devices to conduct economic experiments

Yiting Guo [a], Jason Shachat [b], Matthew J. Walker [c], Lijia Wei [a,*]

[a] Economics & Management School, Wuhan University, Wuhan HUB 430072, China
[b] Durham University Business School, Mill Hill Ln, Durham DH1 3LB, UK
[c] Newcastle University Business School, Newcastle Upon Tyne NE1 4SE, UK

ARTICLE INFO

JEL Classification:
C90
C93
C70

Keywords:
Mobile device
Digitization
Methodology
Experiment
Generalizability

ABSTRACT

Mobile devices enable experimental economists to collect decision-making data from more heterogeneous samples, thereby increasing the generalizability of their results. This generalizability may be compromised if the device is a relevant behavioural confound. This paper reports on a battery of classic economic games and tasks in which we randomize the decision-making device (computer versus mobile) and the laboratory setup (physical versus remote). Our results offer broad support for conducting decision experiments using mobile devices. For six out of eight tasks, we find robust null results for means of treatment effects as well for differences in variances and general distributions of choices across treatments. This should give researchers confidence to conduct studies via mobile devices and in out-of-lab settings. However, we find two caveats. First, subjects using a mobile device exhibit more risk aversion and offer less during bargaining. These effects persist in the physical lab after controlling for subjects' observed characteristics. Second, response times and the time taken to read instructions using a mobile device are shorter in the remote setup. These qualifications suggest the importance of ensuring device consistency across treatments.

1. Introduction

Researchers are turning to deploying experiments outside of traditional dedicated laboratories and through smartphone devices for several reasons, including the desire to show external validity, the growing number of laboratory-in-the-field studies and the simple potential for lower data collection costs. Confidence in this two-way shift in methodology would be bolstered by an evidence-based argument that behaviours well replicated in controlled laboratory settings with participation through computers also replicate under these new methods. Seeking to provide such evidence, in this paper we report on an experiment comprising a battery of eight classic economic games and tasks in which we randomize the decision-making device (computer versus mobile) and the laboratory setup (physical versus remote).

A traditional methodological strength of laboratory economic experiments is the ability to apply strict control to factors such as communication, information and external stimuli. This control instils confidence in the internal validity of such results, a confidence reinforced by successful replication projects (e.g., Camerer et al., 2016). However, laboratory experiments also suffer from certain longstanding criticisms and limitations. Laboratory studies typically rely heavily on the participation of students, leading to commonly expressed concerns of generalizability. Physical laboratory studies also limit the type of decision tasks and research questions one can pursue. Time limits for tasks and interactions, cohort sizes limitations, and difficulties of establishing true anonymity are all constraints facing someone designing a physical laboratory study.

Recent advances in digital technologies facilitate the implementation of field, online and hybrid experiments using portable and low-cost mobile devices. They also offer the opportunity to reach more heterogeneous samples. According to survey data from Pew Research Center in 2021, 15% of US adults report being smartphone dependent, i.e., not using broadband at home but owning a smartphone. However, smartphone dependency differs significantly by demographic group; it is inversely related to age, income and education and also varies by race. For example, the rates of smartphone dependency among White, Black and Hispanic respondents are 12%, 17% and 25%, respectively (Andrew, 2021). Across most emerging economies, smartphones are the most widespread type of mobile device and only a minority have personal access to a computer or laptop (Silver et al., 2019). These emergent

* Corresponding author.
E-mail address: ljwei@whu.edu.cn (L. Wei).

https://doi.org/10.1016/j.socec.2023.102057
Received 6 February 2023; Received in revised form 9 June 2023; Accepted 13 June 2023
Available online 14 June 2023

patterns of smartphone adoption present researchers who embrace advances in digital technologies and software based on web applications optimized for mobile compatibility[1] with an opportunity to collect more heterogeneous samples.

Mobile phones and tablets are increasingly used to implement field experiments and large-scale surveys (Himelein, 2021). The field provides external validity and fills the gap between the physical laboratory and naturally occurring data (see Harrison & List, 2004, Levitt & List, 2009). They also afford rapid deployment to monitor individual attitudes and behaviours, and to collect digitized and geocoded data instantaneously in times of crisis, such as natural disasters (e.g., Beine et al., 2020) or public health emergencies (e.g., Lohmann et al., 2021). Generalizability would be compromised, however, if the device used to implement the experiment is a relevant behavioural confound that is not adequately controlled for in the design process.

We investigate this claim by systematically varying the decision-making device used by subjects in a set of widely adopted experimental tasks across physical and remote setups. Participants in our experiment are randomly assigned *ex ante* to complete the experiment using either a computer or mobile phone device, in either a physical laboratory or remote setting. Compliance with the assigned device is verified by measuring the device resolution and screen dimensions. This permits two strengths of our experimental design. First, we can identify violations of treatment assignment by subjects, and by eliminating associated observations reducing measurement error in our study. Second, we circumvent endogeneity problems inherent with subjects choosing their device.

We selected eight classic behavioural economics games and tasks to test in our experiment based on their frequent use not just as primary outcome variables, but also as extraneous measurements of underlying preferences. In selecting games to measure social preferences and cooperation, we restricted ourselves to 2-person games that have been the subject of extensive prior study: the Dictator Game, Trust Game, Ultimatum Game, Prisoner's Dilemma game and Stag Hunt game. We also included the canonical game for measuring strategic reasoning, the Beauty Contest.[2] Our final two tasks were a sequence of lottery choices to measure risk preferences, and a number-reporting task to infer dishonest behavior. Preferences for risk-taking and truth-telling are key economic parameters measured in the experimental literature. For its tractability, we chose to elicit risk preferences using the method proposed by Abdellaoui et al. (2011), in which subjects choose between a fixed lottery and an increasing certain amount. To assess lying aversion, we use a variation of the methods developed by Jiang (2013), Kajackaite and Gneezy (2017), and Parra (2022). Our method uses a facet of university administration to generate an equiprobable discrete interval, to which subjects are asked to add a number in their minds. In the spirit of previously used methods (e.g., Fischbacher & Follmi-Heusi, 2013), lying is not detectable at the individual level and it is possible to infer the true distribution of the summation formula under full honesty. Our method avoids requiring the use of dice or other equipment, thereby ensuring the consistency of the physical and remote laboratory setups.

From the perspective of generalizability, our results offer broad support for conducting decision experiments using mobile devices: for six out of eight tasks, we find robust null results in terms of differences in the mean and shape of the outcome distributions across decision-making devices, both in- and out-of-laboratory. The effect sizes observed are small enough in magnitude that we would require very large sample sizes to detect a significant difference at conventional statistical levels. There is only weak evidence of greater variability in the mobile phone data, and this is limited to a subset of tasks. Overall, these findings should provide researchers broad confidence to conduct studies out-of-laboratory and via mobile phones.

There are three important qualifications to our results. First, we find that subjects who are randomly assigned to complete the experiment using a mobile phone on average exhibit significantly greater risk aversion, and offer significantly less during ultimatum bargaining, than those randomly assigned to complete the experiment using a computer device. These findings are robust to addressing the threat of multiple hypothesis testing and persist in the physical lab after controlling for subjects' observed characteristics.[3] The finding on ultimatum offers extends to first-order stochastic dominance of the outcome distribution. Thus, our results suggest that researchers should pay attention to the decision-making device when eliciting risk-related preference measurements. For treatment comparisons, ensuring a consistent device across treatment arms is likely to be sufficient.

Second, we find using a within-subjects analysis that the pairwise correlation between different outcome measurements depends on the decision-making device. That is, the device may mediate behavioural spillover effects between tasks in a non-random manner.[4] Establishing control over the device is therefore particularly important when eliciting more than one preference measurement in a session, which is often the case due to budget constraints. In practice, when conducting experiments remotely and given a mix of observed devices in the underlying sample, controlling for the device in analyses at the individual-level is preferable.

Third, subjects' response times and the time spent on reading task instructions are impacted by both the environment and the device used. We find that response times are shorter for those subjects assigned to the mobile phone remote treatment. We find that the time reading instructions is longer for those assigned to the mobile phone in the laboratory treatment, but shorter for those assigned to the mobile phone remote treatment. Accordingly, researchers should consider adopting device consistent controls in complex experiments where concerns about response time and instruction attention are likely to be more salient behavioural drivers.

Two features of our experimental design are noteworthy. First, the necessity to conduct remote experiments throughout the Covid-19 pandemic forced the hands of the experimental and behavioural economics community to loosen long held norms that the controls offered by physical laboratory experiments are a minimal acceptable research standard. As this community moves towards more prevalent use of mobile devices and remote experimentation, they face a gap in knowing how the change in control affects behavior (see also Buso et al., 2021 and Li et al., 2021).[5] To the extent that the pandemic caused transitory shifts in economic preferences over time, a strength of our experimental

[1] For example, oTree (Chen et al., 2016) and LIONESS (Giamattei et al., 2020).

[2] A search for the terms "Dictator game", "Trust game", "Ultimatum game", "Prisoner's Dilemma game", "Stag Hunt game" and "Beauty contest game" on the EconLit database (5th April 2023) produced 703, 655, 532, 620, 102 and 82 results, respectively. Comprehensive meta-analyses of earlier Dictator Game (Engel, 2011) and Trust Game (Johnson and Mislin, 2011) studies in the experimental economics literature each covered data from approximately 130 papers. Cochard et al. (2021) provides a more recent meta-study on the ultimatum game and dictator game, Embrey et al. (2018) on the finitely repeated prisoner's dilemma and Dal Bó et al. (2021) on the Stag Hunt game.

[3] We thank an anonymous referee for identifying evidence of this nuanced result.

[4] For a discussion of behavioural spillover effects in traditional lab experiments, see Bednar et al. (2012).

[5] The National Science Foundation recognised the potential of the online environment for behavioural research early on (see Bainbridge, 2007). In terms of taxonomy, the online experiment sits in the region between the traditional controlled lab experiment and the more natural setting of the field experiment; Charness et al. (2013) refer to experiments in this region as "extra-laboratory" experiments.

data is that both physical and remote lab responses were collected contemporaneously and before the pandemic in May 2019.[6]

Second, our protocols control for a larger number of elements than previous physical versus remote participation studies. These elements include the subject pool, recruitment, randomization and matching protocols, experimenter communication channel, monetary stakes, payment technology and the device. That is, we tighten the ceteris paribus assumption. Previous studies individually control for only a subset of these factors (see Table 1 for details on an inventory of these studies). The authors of early papers tested for individual-level differences in consumption and savings decisions (Vital Anderhub et al., 2001) lottery evaluations (Tal Shavit et al., 2001) and trust (Charness et al., 2007), finding similar behavior on average but larger variance, lower risk aversion and attenuated social preferences online. Other researchers have successfully replicated experiments and behavioural anomalies on representative MTurk samples (Horton et al., 2011; Amir et al., 2012; Arechar et al., 2018; Gupta et al., 2021; Erik Snowberg & Yariv, 2021) and even virtual world platforms (Chesney et al., 2009; Fiedler & Haruvy, 2009).[7]

Of the studies which use the same subject pool across remote and lab settings, Hergueux and Jacquemet (2015) build an innovative environment that holds constant the stakes and interface. They observe similar social and risk preferences in the online elicitation. Snowberg and Yariv (2021) find qualitatively unchanged behaviours between the lab-going sub-population and overall university student population. Dickinson and McEvoy (2021) observe an increase in dishonesty when moving from the lab to a remote setup. In a study that tests for differences in lab control versus no lab control for elicited time, risk and charitable preferences, Prissé and Jorrat (2022) find no differences except for reduced charitable giving in the no lab treatment. Finally, in a lab experiment with random device assignment, Mograbi (2022) finds no significant difference in elicited risk preferences via smartphone or computer, but does find greater elicited present bias when using smartphones (Mograbi's study does not include an online treatment).[8]

Remote experiments are typically deployed through the web browser and so it is difficult to control the device used by the subject. Use of a smartphone device may increase measurement error due, for example, to smaller screens sizes, lower response times or a greater propensity to multi-task (Lugtig & Toepoel, 2016). Human-computer interactions may also be influenced by the touch interface and ownership (Brasel & Gips, 2014; Melumad & Pham, 2020), whether via psychological channels (e.g., emotional benefits, sense of privacy) or functional mechanisms (e.g., touch interface, compactness of information).

We proceed in the next section by presenting our experiment design which includes descriptions of the battery of tasks, the recruiting process and our implementation procedures. In section three, we present our results with respect to treatment effects of decisions, the statistical power of these effects and then assess correlations of decisions within alternative treatment cells. In section four, we analyze the differences in response time and the length of time spent reading instructions across treatments and how this impacts decision making. Then we offer our concluding remarks in section five.

[6] The mobile phone data used in this study served as the baseline pre-pandemic sample in two related experimental papers examining how the pandemic shifted pro-social and risk preferences (Shachat et al., 2021a, Shachat et al., 2021b).

[7] The variable data quality of online labour platforms has been discussed extensively in the literature. We direct the interested reader to Peer et al., 2022 for an up-to-date account and references Peer et al., 2022.

[8] Although not significant, Mograbi (2022) finds that subjects with a smartphone display a lower degree of risk-taking than with a computer (p-value = 0.101, two-sided t-test), which is qualitatively consistent with the result of our study.

2. Experimental design

2.1. Decision-making tasks

The experiments reported in this study were conducted in the Spring of 2019, using the subject pool database and facilities of the Center for Behavioral and Experimental Research in Wuhan University, China. Each subject participated in seven incentivized economic games or preference elicitation tasks in sequence. These tasks were as follows: Dictator Game (DG); Beauty Contest (BC); Truth-Telling (TT); Stag Hunt (SH); Prisoner's Dilemma (PD); Risk Preference (RP); and Trust game (TG) or Ultimatum Game (UG). Only one of the TG and UG was included in each session to mitigate behavioural spillover effects for second movers between these two tasks (the sample sizes for these two tasks are correspondingly smaller). As we did not wish to provide any information to subjects about the actions of other persons in the session until after the completion of all tasks, we chose not to randomize the order of task presentation. Changing the order in which the extensive-form task (TG and UG) was presented would have violated this informational constraint. We acknowledge that this is a limitation of our design. Nevertheless, to the extent that the task order is consistent across treatments, it should not bias the treatment effects.

Below, we provide a brief description of the players, action sets, and payoffs in each task. The tasks were programmed using oTree software (Chen et al., 2016).

Task 1. DG. Subjects are randomly matched into pairs. Within a pair, subjects are randomly assigned to the role of either Player 1 or Player 2. Player 1 is allotted 5 RMB and decides how much of this endowment to send to Player 2. Player 2 has no decision to make. This task measures pure altruistic preferences.

Task 2. BC. Subjects are randomly divided into groups of four. Within a group, subjects choose an integer between 0 and 100 (inclusive) The subject whose guess is closest to one-half of the average value selected within the group wins 8 RMB (ties broken evenly); the remaining subjects earn zero payoff for the task. This task measures levels of rationality and strategic thinking.

Task 3. TT. Each subject forms their own digit by choosing an integer between 0 and 9 (inclusive) which they are asked to add to the final digit of their student ID number (unobserved by the experimenter).[9] A random integer between 0 and 9 is then displayed on-screen. If this number matches one's digit, then the subject earns 5 RMB; else zero. This task measures preferences for truth-telling: each integer has a one-tenth probability of being realized and so we would expect to observe matches on approximately one in ten reports at the aggregate level. Cheating cannot be detected at the individual level. Subjects need not fear exposure for their dishonesty, even in the physical lab or if they (mistakenly) believe that the experimenters do have access to their student ID, because the choice of integer is made purely in their minds (see also Jiang, 2013).

Task 4. SH. Subjects are randomly matched into pairs. Each player within a pair simultaneous chooses either Option A or Option B. If both players choose A, then both players earn 3 RMB. If both players choose B, then both players earn 8 RMB. If one player chooses A and the other player chooses B, then the first player earns 3 RMB and the second player earns 0 RMB. This task measures how individuals prefer to resolve coordination on risky but more rewarding collective action versus no-risk but lower reward collective action.

Task 5. PD. Subjects are randomly matched into pairs. Each player within a pair simultaneous chooses either Option C or Option D. If both players choose C, then both players earn 6 RMB. If both players

[9] The final digit of students' ID numbers is distributed evenly across integers from 0 to 9 as the university randomly allocates the last four digits of a student ID from 0000 to the total number of students in the entrance year.

Table 1
– Design aspects held constant between remote and physical lab treatments in previous experimental studies.

	Subject pool[a]	Recruitment[b]	Matching protocol[c]	Experimenter communication[d]	Mitigate dropouts[e]	Incentives[f]	Device identified[g]
Anderhub et al. (2001)	×	×	×	✓	✓	✓	–
Arechar et al. (2018)	×	–	–	–	✓	✓	–
Charness et al. (2007)	×	–	–	–	–	✓	–
Dickinson & McEvoy (2021)	✓	✓	–	–	–	✓	–
Fiedler & Haruvy (2009)	×	–	✓	–	–	✓	–
Gupta et al. (2021)	×	–	✓	–	–	✓	–
Hergueux & Jacquemet (2015)	✓	✓	×	–	✓	✓	–
Horton et al. (2011)	×	–	–	–	–	✓	–
Prissé & Jorrat (2022)	✓	✓	–	–	–	✓	–
Shavit et al. (Tal 2001)	×	✓	×	–	–	×	–
Snowberg & Yariv (Erik 2021)	✓	–	×	–	–	✓	–

Notes. A "✓" ("×") means a similar (different) design in lab and remote environment. A "–" means that this aspect was not explicitly mentioned in the article.
[a] The same subject pool.
[b] Invitations sent in advance, no information about the task.
[c] Simultaneous matching at end of session to ensure joint determination of payoffs.
[d] The same communication channel with the experimenter to answer questions on instructions.
[e] Option to rejoin after network disconnection.
[f] Payment based on outcomes of all tasks in local currency.
[g] Device used to complete the remote experiment identified.

choose D, then both players earn 3 RMB. If one player chooses C and the other player chooses D, then the first player earns 0 RMB and the second player earns 9 RMB. This task measures the conflict between self-interest and mutual cooperation.

Task 6. RP. Each subject is presented with a series of nine pairwise choices between a lottery (option A) and a sure amount of money (option B). The lottery remains fixed across all choices: a 50% chance of receiving 9 RMB, and a 50% chance of receiving 3 RMB. The sure amount increases evenly with each choice from 3 RMB up to 9 RMB. After all choices have been made, the system randomly selects one of the nine pairs of options for payment. This task measures risk tolerance (a greater number of lottery choices indicates a greater willingness to take risks).[10]

Task 7. TG. Subjects are randomly matched into pairs. Within a pair, subjects are assigned to the role of either Player 1 or Player 2. Player 1 is allotted 8 RMB and decides how much of this endowment to send to Player 2. Any money sent is multiplied by a factor of three before reaching Player 2. Any money not sent is kept by Player 1. Player 2 observes the multiplied amount sent and decides how much of it to return to Player 1. Any money not returned is kept by Player 2. This task measures levels of trust and reciprocity.

Task 8. UG. Subjects are randomly matched into pairs. Within a pair, subjects are assigned to the role of either Player 1 or Player 2. Player 1 is allotted 8 RMB and decides how much of this endowment to send to Player 2. Player 2 can accept or reject the allocation. In case of rejection, both players receive zero payoff. This task measures fairness preferences.

2.2. Treatments and protocols

All experiments in this study were approved by the Academy of Humanities and Social Sciences, Wuhan University. Our design randomizes two factors: first, the laboratory setup (physical versus remote); second, the decision-making device (personal computer [laptop] versus mobile device [smartphone]).[11] For the remote experiments, the experiment software automatically checks the operating system and screen dimensions to verify device treatment compliance. To check for any effect on decision-making of device ownership, we implemented an additional variant in the physical lab in which we supplied subjects with a laptop computer owned by the laboratory. Hence, there are five treatments, using a between-subjects design (Table 2).[12]

Participants were students registered on a wide range of academic majors at Wuhan University (mean age = 20.6, 65% female; for details of the sample balance across experiment conditions, see Table A1 in the online appendix). Both in the physical and remote laboratory, the recruitment of participants, randomization protocol and payment transfers were executed using the Ancademy platform for conducting social science experiments (https://www.ancademy.org/). Ancademy is based on the open interface of WeChat.[13] Invitations were sent at random to members of this database (c. 9000 members) to participate in a session at a scheduled time. The day before a session, all subjects received a confirmation message specifying the device needed for participation (computer or mobile phone) and details of how to sign in

Table 2
– Treatment matrix ($N = 581$).

	Computer Personal	Public	Mobile Device Personal
Physical Lab	$n = 108$ (3• 28 + 1• 24)	$n = 112$ (4• 28)	$n = 112$ (4• 28)
Remote	$n = 160$ (4• 28 + 2• 24)		$n = 112$ (4• 28)

Notes: Terms in parentheses are (number of sessions x number of subjects in the session). We exclude 23 subjects from our final dataset for using the wrong device to complete the experiment (4 in Lab/Computer, 13 in Remote/Computer, 6 in Remote/Mobile); the results are qualitatively unchanged by including these subjects in either the treatment selected into, or the treatment originally assigned.

[10] In our construction of the risk tolerance variable below, risk neutrality corresponds to a score of 5.5.

[11] In practice, all subjects in the computer treatments used a laptop rather than desktop computer to complete the experiment. Based on screen dimension data, we were able to verify this for both the physical and remote treatments.

[12] This study was not pre-registered; the research objective defined ex ante was to test the stability of economic decision-making across devices and experiment setting. We discuss this point further in the conclusion.

[13] Although WeChat is a ubiquitous social media platform in China, it has a low adoption rate in other regions. A successful demonstration of using the social media platform WhatsApp to build a subject pool is provided by Jorrat (2021). In that study, a single iteration "snowball" recruiting is conducted with a cluster randomization technique to produce experimental cohorts.

to the Ancademy platform at the scheduled time (whether remote or in the physical lab).

Each session followed the same procedure. After all subjects had signed in, a six-digit quick join code was distributed; subjects were informed that this code would enable them to rejoin the session quickly in case of disconnection. During the session, a private communication channel with the experimenter was available via WeChat for questions or clarifications. For the two-player games, matching was conducted simultaneously at the end of a session to ensure joint determination of payoffs. To mitigate wealth effects, no feedback about earnings was provided until after the completion of all tasks. Upon conclusion of the session, earnings were transferred directly to subjects' WeChat wallets within 24 hours. Subjects were paid based on choices in all decision tasks and no feedback was provided until the completion of all tasks.

We conducted a total of 22 sessions across the five treatments: 12 sessions in the physical lab (4 sessions with mobile phone, 4 sessions with public computer, 4 sessions with personal computer); and 10 sessions remotely (4 sessions with mobile phone, 6 sessions with personal computer). Average earnings were 40.6 RMB (approximately 6 US Dollars), including a show-up fee of 10 RMB. A session lasted approximately 30 min. We recruited 28 subjects for each session.[14] We exclude data from 23 subjects who failed to comply with the assigned device to complete the experiment; the results are qualitatively unchanged by including these subjects in either the treatment selected into, or the treatment originally assigned (see the online appendix). Thus, the final sample size is $N = 581$.[15] Across the computer treatments, we find no significant differences in behavior between use of a personal and public computer and we pool this data.

In summary, across treatments we hold constant the subject pool, recruitment and protocols, availability of the experimenter communication channel, monetary stakes, and payment technology. Our design mitigates involuntary dropouts (there is no attrition) and permits verification of the randomly assigned device in both physical lab and remote settings.

3. Results

3.1. Generalizability of the mobile device

Table 3 presents the pooled summary statistics for sessions in which subjects were randomly assigned to use either a computer or a mobile phone device for the decision-making tasks. As no feedback is provided until after the completion of all tasks, we use the subject as the independent level of observation. To address the threat of multiple hypothesis testing and the possibility of false positives, we calculate False Discovery Rate (FDR) adjusted q-values across the ten outcome measurements (Benjamini et al., 2006), based on two-sided t-tests (mean) and Wilcoxon rank-sum tests (shape).

There are no significant differences in the mean or shape of the outcome distributions for pure altruistic preferences, trust or cooperation between the computer and mobile phone treatments (all q-values $>$ 0.320). Amounts sent by dictators in the DG are around 30% of the endowment independently of the assigned device. There are similar relative differences between amounts sent in the DG and amounts offered in the UG in both sets of treatments, although the mobile phone sample exhibits higher variance (see the online appendix). Trustees send around 40% of the endowment in the TG on both devices, and this is a breakeven strategy on average based on the trustor's response. We also observe similar rates of cooperation in the PD game - around one-third of subjects choose to cooperate - and in the SH game - around 88% of subjects attempt to coordinate on the efficient but risky outcome.

A preference for truth-telling is slightly higher in the mobile phone data, where 67% report matches, than in the computer data, where 73.3% report matches, but this difference is not significant for either the mean or shape of the distribution (q-values $>$ 0.264). Thus, for both samples, the aggregate data reveals a higher-than-expected frequency of reports that yield a higher payoff in the TT task.[16] There are no significant differences in the mean or shape of the distribution of guesses between samples using different devices in the task designed to measure strategic reasoning (q-values $>$ 0.781). Mean guesses in the BC game (rounded to the nearest integer) are 29 in the computer sample and 28 in the mobile phone sample.

Result 1. *For six out of eight decision-making tasks, we find no significant differences in the mean or shape of the outcome distributions between the computer and mobile phone devices*.

However, we observe significantly lower ultimatum offers (q-value $= 0.025$ [shape] and 0.076 [mean]) and lower acceptance rates (q-value $= 0.044$ [shape] and 0.086 [mean]) when subjects bargain with a mobile phone. Average UG offers are 43.1% of the endowment when using a computer versus 37.6% of the endowment when using a mobile phone; the acceptance rates are 90.2% and 74.5%, respectively. Subjects in the mobile phone treatments also display significantly greater risk aversion than those in the computer treatments. Although subjects in both samples exhibit risk aversion in the RP task overall (score < 5.5), the degree of risk aversion is larger for those subjects using a mobile phone (q-value $= 0.012$ [shape] and 0.007 [mean]).

The cumulative response distributions reveal that the computer sample exhibits a first-order stochastic dominance relationship with the mobile phone sample for the RP Tolerance and UG Offer measurements, although after adjusting for multiple hypothesis testing we only reject the null hypothesis that the two distributions are drawn from the same population for UG offers (q-value $= 0.030$, two-sided Kolmogorov-Smirnov test).[17]

Result 2. *Subjects on average display greater risk aversion and offer less during ultimatum bargaining when randomly assigned to complete the experiment using a mobile device*.

There are no significant differences in the mean or shape of the outcome distributions between the physical lab and remote setups for any of the eight decision-making tasks (see Table A2 in the online appendix, all q-values > 0.385). This result supports previous studies (e.g., Hergueux & Jacquemet, 2015) as to the high internal validity of the remote setup after holding a range of other design aspects constant.

Result 3. *We find no significant differences in the mean or shape of the outcome distributions between the physical and remote laboratory setup in any task*.

The key to the veracity of any null result is the power of the establishing hypothesis test. In Table 4, we calculate – separately for each of the behavioural measurements – the required sample size for our study to detect a significant effect at the 5% statistical level and with 80% power, alongside the actual effect size and the minimal detectable effect size given the number of subjects in our four between-subjects samples. On these bases we argue the null results observed above are robust and not likely to be due to a lack of statistical power. In the six decision-making tasks for which we found no difference in means across treatment arms (Result 1), the effect sizes observed in our experiment would require very large sample sizes for a well-powered study to detect a significant effect. The required sample sizes range from 687 to 8025,766

[14] Due to no-shows on the day, three sessions only had 24 subjects (see Table 2).

[15] For the RP task only, we exclude 10 "inconsistent" subjects who switch from the lottery to the safe option more than once in the list. We return to discuss these inconsistent responses below in the context of data variability among treatments.

[16] The self-reported consistency rate in our sample is quite high compared to other studies (see, e.g., Abeler et al., 2019). As Jiang (2013, p.329) points out, cheating for a higher payment may be perceived by subjects as less deliberate (and so more acceptable) when it requires only an "internal twist of the mind".

[17] For more details, see Figure A1 and Table A17 in the online appendix.

Table 3
– Decision-making using a computer versus mobile phone device.

	Computer			*q-value*	*Mobile phone*		
	n	Mean	SD		n	Mean	SD
DG Sent [0,5]	185	1.575	0.992	0.321 [0.431]	106	1.416	1.079
BC Guess [0,100]	363	28.725	19.823	0.868 [0.781]	218	27.592	17.192
TT Match {0,1}	363	0.733	0.443	0.264 [0.278]	218	0.670	0.471
SH Efficient {0,1}	363	0.879	0.327	0.945 [0.944]	218	0.881	0.325
PD Cooperate {0,1}	363	0.325	0.469	0.924 [0.923]	218	0.317	0.466
RP Tolerance {1, 2,…,10}	355	4.789	1.268	0.012 [0.007]	212	4.443	1.111
TG Sent [0,8]	93	3.188	2.524	0.850 [0.901]	55	3.345	2.612
TG Return [0,3• *Sent*]	93	2.930	4.584	0.470 [0.854]	55	3.309	3.983
UG Offer [0,8]	92	3.448	0.951	0.025 [0.076]	51	3.010	1.051
UG Accept {0,1}	92	0.902	0.299	0.044 [0.086]	51	0.745	0.440

Notes: Multiple hypothesis testing adjusted False Discovery Rate (FDR) *q*-values (10 comparisons) based on two-sided Wilcoxon rank-sum tests [t-tests].

Table 4
– Required sample size and minimal detectable effects.

	Computer vs. Mobile phone			Remote vs. Physical Lab		
	Required sample size	Cohen's d effect size	Min. detectable effect	Required sample size	Cohen's d effect size	Min. detectable effect
DG Sent [0,5]	687	0.155	0.358	8025,766	0.001	0.330
BC Guess [0,100]	4568	0.060	4.372	1404,109	0.003	4.456
TT Match {0,1}	853	0.139	0.111	27,382	0.025	0.107
SH Efficient {0,1}	461,543	0.006	0.078	132,855	0.011	0.076
PD Cooperate {0,1}	49,190	0.018	0.112	29,806	0.023	0.109
RP Tolerance {1,2,…,10}	204	0.285	0.285	1427	0.107	0.291
TG Sent [0,8]	4346	0.062	1.229	6301	0.051	1.233
TG Return [0,3• *Sent*]	2187	0.087	2.008	58,675	0.017	2.127
UG Offer [0,8]	85	0.444	0.497	33,587	0.022	0.489
UG Accept {0,1}	85	0.444	0.193	33,587	0.022	0.178

Notes: Required sample size in each group is calculated by G*Power with 80% power and 0.05 significance level (based on two-sided t-tests). Cohen's d effect size: calculated by R package effsize. Min. detectable effect with 80% power and 0.05 significance level.

subjects. For all 10 remote versus physical lab statistical comparisons, and for 5 out of 7 null computer versus mobile phone comparisons, the required sample size is 4-digits or more.[18]

3.2. Covariates and interaction effects

The aggregate statistics reported above neither control for observed subject characteristics, nor consider interaction effects. To address these concerns, we regress each of the decision-making outcomes on the full interaction between the device (computer versus mobile) and experiment setup (lab versus remote), controlling for age, gender, monthly expenditure, and academic major. Estimation is using OLS for the continuous outcome measures and logistic regression for the binary outcome measures and we compute heteroskedasticity-consistent standard errors (see Table 5 for the continuous outcomes, which include risk tolerance and ultimatum offers; estimates for the binary outcomes are contained in the online appendix). We again calculate FDR adjusted *q*-values for the mobile phone treatment dummy and these are reported alongside conventional *p*-values in the regression output tables. Further, we conduct Wald tests to check whether there is an effect of the device in the remote setup.

Consistent with the aggregate findings, we find significantly greater risk aversion and lower UG offers in treatments with a mobile device in the physical lab (*q*-value = 0.003 and *q*-value = 0.0498, respectively). Based on the results of the Wald tests, we find no significant effect of the device in the remote setup on any outcome measurement. UG acceptance rates do not differ between device treatments after accounting for the difference in offers (*q*-value = 0.563). There is a strong positive relationship between UG offers and acceptance rates in all treatments. Average guesses in the BC game are significantly higher among female subjects in our sample than among male subjects; no other gender effects in task performance emerge from the pooled data.[19]

[18] In the online appendix, we use the method of Bellemare et al. (2016) to compute the ex-post power of our experimental design given the observed effect sizes. The results are qualitatively unchanged.

[19] Although not the focus of this study, we do uncover some interesting gender effects when splitting the sample between male and female subjects. Whereas we find no significant differences in female decision-making between the remote and physical lab setups, males are more likely to report a match in the TT task online (q-value = 0.083). We also observe that when using a mobile device, males submit lower guesses in the BC game, report fewer matches in the TT task and offer less in the UG game (all q-values = 0.052). Moreover, while females exhibit greater risk aversion in the RP task on a mobile device (q-value = 0.027), this difference is not significant for males (q-value = 0.305), with the caveat of a much-reduced sample size for males. Full details are contained in the online appendix. These observations complement Braut (2023) who finds that female subjects (only) are more risk-averse online.

Table 5
– Linear regression analysis of decision-making in the experiment.

	Dependent Variable					
	DG Sent (1)	BC Guess (2)	RP Tolerance (3)	UG Offer (4)	TG Sent (5)	TG Return (6)
Remote	−0.003	−0.62	−0.24*	−0.07	−0.20	−0.31
	(0.15)	(2.08)	(0.13)	(0.22)	(0.59)	(0.58)
	[0.985]	[0.958]	[0.711]	[0.958]	[0.958]	[0.958]
Mobile	−0.14	−1.12	−0.48***	−0.57***	0.36	−0.25
	(0.16)	(2.07)	(0.13)	(0.22)	(0.73)	(0.72)
	[0.782]	[0.887]	[0.003]	[0.0498]	[0.887]	[0.909]
Remote × Mobile	−0.10	−1.1	0.23	0.30	−0.35	0.56
	(0.25)	(3.23)	(0.21)	(0.34)	(0.94)	(0.76)
Female	0.18	5.70***	−0.16	0.30	−0.19	−0.57
	(0.13)	(1.63)	(0.11)	(0.19)	(0.53)	(0.47)
TG Sent						1.42***
						(0.09)
Constant	3.69***	37.82***	5.76***	4.13***	5.58***	−1.15
	(0.71)	(11.90)	(0.67)	(0.86)	(2.76)	(2.79)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	291	581	567	143	148	148
R-squared	0.08	0.06	0.06	0.15	0.06	0.73
H_0: Remote = Mobile + Remote × Mobile (Is there an effect of the device in the remote setup?)						
F-stat.	0.685	0.166	0.002	0.241	0.045	0.577
p-value (two-sided)	0.409	0.684	0.960	0.625	0.833	0.449

Notes: *$p<0.1$; **$p<0.05$; ***$p<0.01$. Robust standard errors are shown in the parentheses, calculated using the Huber/White sandwich estimator of variance. Multiple hypothesis testing adjusted False Discovery Rate (FDR) q-values in square brackets (10 comparisons) for remote and mobile treatment dummies. Control variables include subject age, gender (female dummy reported above), monthly expenditure, and academic major.

Result 4. *The effect of the mobile device on risk aversion and ultimatum offers is mediated in the physical lab.*

3.3. Variance-covariance of responses and similarity of correlations

To examine whether decision-making is systematically noisier among subjects randomly assigned to participate using a mobile device, which we might expect if these factors increase measurement error, we consider the coefficients of variation.[20] Based on Feltz and Miller tests (Feltz & Miller, 1996), we find that the UG acceptance rate is more variable in the mobile phone sample than in the computer sample (p-value < 0.001), and weak statistical evidence of greater variation on the mobile for UG offers (p-value $= 0.072$) and reported TT matches (p-value $= 0.060$). There are no significant differences in variability at the 10% level for the remaining seven behavioural measurements.

A further indicator of reduced data quality across the samples is provided by the number of "inconsistent" responses in the RP task, i.e., those subjects who switch from the lottery to the safe option more than once in the list. The frequencies of inconsistent responses are similar on the two devices (1.6% for the computer treatment and 1.8% for the mobile phone treatment, two-sided proportion test: $p = 0.890$). The rates of inconsistent responses in the RP task are also similar between the remote sample (2.0%) and the physical lab sample (1.5%, two-sided proportion test: p-value $= 0.826$).

Finally, as Snowberg and Yariv (2021) point out, experimental economists care about correlations between different behaviours and attributes. As experimentalists are commonly subject to budget constraints and wish to maximize the usability of their datasets, they often elicit more than one preference measurement in a session, for example, when distinguishing between an individual's willingness to trust and willingness to take risks. Consistency of pairwise correlations among these different measurements are thus of interest. We apply Snowberg and Yariv's similarity of correlations method to examine how different behavioural measurements relate between decision-making devices and laboratory setups in our dataset. Specifically, we determine whether a particular statistically significant pairwise correlation recorded when two outcome measurements are elicited using a mobile device is also recorded when those measurements are elicited using a computer. Similarly, whether a particular statistically significant pairwise correlation recorded when two outcome measurements are elicited in the physical lab is also recorded when those measurements are elicited in the remote setup.

The results, presented in Fig. 1, include the sign and significance of pairwise correlations among all outcome measurements. Panel (a) covers mobile phone and computer devices, and panel (b) covers remote and physical laboratory setups. A positive (negative) and significant correlation between two outcome measurements is denoted with a "+" ("-") and an insignificant correlation with a "0". Taking panel (a) as an example, cells in which the pairwise correlations agree in sign and significance between devices are denoted as "Complete agreement" and we use a single symbol in that cell. Cells in which the pairwise correlation is significant on one device and insignificant on the other device are denoted as "Partial disagreement" and the first symbol in the cell corresponds to the correlation observed on the mobile device. Similarly, cells in which the pairwise correlation are significant on both devices but in opposite directions are denoted as "Complete disagreement".

We observe greater inconsistency in pairwise correlations between the mobile and computer samples than between the lab and remote samples. Out of 37 pairwise correlations, 1 case indicates complete disagreement: whereas there is a significant negative association between PD Cooperate and UG Offer in the phone sample, there is a significant positive association between these two measurements in the computer sample. A further 9 cases (24.32%) show partial disagreement, while the pairwise correlations feature a complete agreement in sign and significance (72.97%). No pairwise correlation displays complete disagreement between the lab and remote samples.

4. The mediating effect of response and instruction times

What might explain our above findings that subjects display greater risk aversion and offer less during ultimatum bargaining when randomly assigned to complete the experiment using a mobile device (Result 2)?

One plausible mechanism is related to response time. If subjects are

[20] Li et al. (2021) find that remote experiment data is less noisy when using a webcam-on protocol. We find no significant differences in variability for the physical lab versus remote sample comparisons (see the online appendix).

Mobile Phone vs. Computer	DG Sent	BC Guess	TT Match	SH Efficient	PD Cooperate	RP Tolerance
DG Sent						
BC Guess	+					
TT Match	0-	0				
SH Efficient	0	0	0			
PD Cooperate	0+	+0	-0	+		
RP Tolerance	-0	-0	0	0	-	
TG Sent	+	0	-	0	0	0
TG Return		0	0-	0	0	0+
UG Offer	+	+0	0	0	-+	0
UG Accept		0	0	0	0	0

Panel (a): Decision-making device. Mobile Phone vs. Computer.

Remote vs. Physical Lab	DG Sent	BC Guess	TT Match	SH Efficient	PD Cooperate	RP Tolerance
DG Sent						
BC Guess	+0					
TT Match	-	0				
SH Efficient	0	0	0			
PD Cooperate	0	+	-0	+		
RP Tolerance	0	0	0	0	-	
TG Sent	+	+0	-	0	0	0
TG Return		0+	0-	0	0	0+
UG Offer	+	0+	0	0	0	-0
UG Accept		0	0	0	+0	0

Panel (b): Laboratory setup. Remote vs. Physical Lab.

Complete agreement | Partial disagreement | Complete disagreement

Fig. 1. – Within-subjects correlations across decision-making device and lab setup.
Notes: A "+" denotes a significant (at the 10% level) positive correlation, a "-" donates a statistically significant negative correlation and a "0″ donates an insignificant correlation. We use a single symbol if the two signs in the same cell agree. "Complete agreement": the two signs in a cell are the same. "Partial disagreement": one sign in the cell is significant positive/negative, the other sign is insignificant. "Complete disagreement": one sign in the cell is significant positive, the other sign is significant negative. Only one of the TG and UG was included in each session to mitigate behavioural spillover effects for second movers between these two tasks.

prone to decide more quickly on a mobile phone device than on a computer, perhaps due to differences in the mental allocation of time between tasks on different devices (cf. Rajagopal & Rha, 2009), the naturalness of the decision interface, or decision heuristics, then this might manifest itself via changes in willingness to take risks. Both the RP and UG tasks involve risk, objective or strategic. There is experimental evidence to suggest that increased time constraints may increase risk aversion in the gain domain (Cahlíková and Cingl, 2017, Kocher et al., 2013, Kirchler et al., 2017). An alternative interpretation of response times is provided by Konovalov and Krajbich (2019). They find that when subjects are closer to indifference, they tend to take longer to make their decisions in risk and social choice domains. Thus, response times may also correlate with strength of preference.

To explore this mechanism further, we first conduct separate OLS regressions of response time and instructions reading time on our treatment dummies and individual-level covariates, pooled across all eight tasks (see Table 6). Both variables have long tails to the right and so we take the logarithmic transformation. The results of this analysis suggest that overall response time is significantly lower among those subjects assigned to complete the experiment using a mobile phone device in the remote setup, versus the physical lab with computers (p-value = 0.030). Relative to this benchmark, subjects using a mobile phone device remotely also spend significantly less time reading the instructions (p-value = 0.012), whereas the opposite is true for subjects assigned to complete the experiment in the physical lab using a mobile device (p-value = 0.006).

Table 6
– Regression analysis of response time and instructions reading time (in seconds).

Dependent variable (ln):	Response time (1)	Instructions reading time (2)
Computer & Remote	0.004	0.05
	(0.04)	(0.04)
Mobile & Physical Lab	−0.03	0.11***
	(0.05)	(0.04)
Mobile & Remote	−0.15**	−0.16**
	(0.07)	(0.06)
Female	0.03	−0.002
	(0.04)	(0.04)
Age	0.02***	0.01
	(0.01)	(0.01)
Constant	2.63***	3.41***
	(0.27)	(0.20)
Control variables	Yes	Yes
Observations (clusters)	3777[a] (581)	4067[b] (581)

Notes: *$p<0.1$; **$p<0.05$; ***$p<0.01$. This table reports coefficient estimates from OLS regressions. Standard errors clustered at the individual level are reported in parentheses. The dependent variable in column (1) is how much time subjects spend on making their decision in a given task. The dependent variable in column (2) is how much time subjects spend on reading the instructions in a given task. Both variables are in natural logarithmic form. Control variables include subject age (reported), gender (reported), monthly expenditure, and academic major. The constant term represents the Computer & Physical Lab treatment cell.

[a] (7 tasks x 581 individuals) – (290 recipients in the Dictator Game) = 3777.

[b] 7 instruction sets x 581 individuals = 4067.

Depending on one's favoured interpretation, these findings may indicate more careful and deliberative thinking or greater strength of preference in the traditional physical lab setting, perhaps because the tasks are more salient or there are fewer distractions. There is heterogeneity in the size of the difference in response times among tasks. After adjusting for multiple hypothesis testing, we find that the difference in response time between devices is significant for the BC game (q-value = 0.019), PD game (q-value < 0.001) and RP task (q-value = 0.019), although the direction of the effect is the same across most tasks (see the online appendix).

To test whether differential response or instructions time can plausibly explain variation in behaviours in our experiment, we re-estimated the regressions from Section 3.3 after controlling for response and instructions reading time as covariates. The significant negative effect of the mobile treatment dummy on risk tolerance remains a robust finding (q-value = 0.007). The corresponding negative effect on UG offers no longer survives multiple hypothesis testing (q-value = 0.094). We find no economically or statistically significant effect of variability in individual-level response or instructions reading time on risk aversion, or of instruction reading time on UG offers. There is, however, a negative association between UG offers and response time, which is significant at the 5% level. This suggests that faster decision times on the mobile device may mediate the lower observed offers.

We note, tangentially, that for our sample there is a strong inverse statistical relationship between response time and the level of (naïve) equilibrium behavior in the BC game – naïve as lower guesses do not necessarily increase the probability of winning. This is consistent with earlier evidence that deciding more slowly produces faster convergence to equilibrium behavior (Kocher & Sutter, 2006). Subjects who decide more quickly are also significantly more likely to cooperate in the PD game and SH game (see Rand et al., 2012, on the intuitiveness of cooperation).[21]

5. Concluding remarks

Rapid developments in digital technology have led to a paradigm shift in opportunities for the conduct of incentivized decision-making experiments in settings where traditional laboratory experiments are not feasible. With a shift towards more remote learning and working likely to persist, and a potential reduction in the time cost of experiment management, implementation of experiments using low-cost mobile phone devices will remain an attractive option in the experimentalist's toolkit. To investigate whether generalizability might be compromised by the nature of the decision-making device, we presented evidence from a battery of economic games and decision-making tasks in which we randomly assigned the device (computer versus mobile phone) and the laboratory setup (physical versus remote), holding constant the subject pool, experimental protocols, communication channel, monetary stakes, and payment technology.

This study was not pre-registered and we acknowledge that this is a limitation of our research design. The study was designed to provide a transparent evaluation of the generalizability of behavior in classic economic games and tasks across device and lab setting. In the data analysis, we attempt to mitigate some reproducibility concerns by correcting our statistical comparisons for multiple hypothesis testing. The data and materials to replicate the experiment tasks and results reported in this paper are available in the project repository at the Open Science Framework.[22]

As noted in the introduction, the mobile phone treatment data collected for this study served as the baseline sample in separate work that we conducted investigating the behavioural consequences of the Covid-19 pandemic (Shachat et al., 2021a; Shachat et al., 2021b). The ability to deploy incentivized decision-making experiments via mobile devices in Wuhan at the onset of the pandemic was crucial in enabling the monitoring of individual attitudes and behaviours in real time. We used the mobile phone treatment data, rather than the computer treatment data, as the baseline sample in that work to ensure consistency of the decision-making device between pre- and post-pandemic samples. That is, we were concerned that generalizability might be compromised if the device used to implement the experiment is a relevant behavioural confound that is not adequately controlled for in the design process.

While the findings of the present study offer support for conducting decision experiments using mobile devices across a class of common behavioural economics instruments designed to measure pro-sociality, cooperation, and strategic reasoning, they also suggest that we were right to be cautious. In our study, subjects who are randomly assigned to complete the experiment using a mobile phone in the physical lab are more risk averse and offer less during bargaining than those who are randomly assigned to use a computer. Within-subjects correlational analyses also indicate systematic behavioural differences that are influenced by the decision-making device.

We identify response and instruction time as a potential behavioural mechanism for divergent behavior across devices. However, we admit this is likely not a complete explanation as there remains residual differences in assessed risk aversion. One behavioural explanation, of which our study design does not permit evaluation, is that individuals experience a greater endowment effect and aversion to the risk of loss when making decisions using their mobile phone (Kahneman et al., 1991). Hein et al. (2011) observe that, compared with traditional laptop and desktop computers, tablet and smartphone devices may induce a

[21] The strength of evidence on the intuitiveness of cooperation is disputed (see, e.g., Tinghög et al., 2013). There is no evidence in our sample to suggest that subjects who take more time to decide are less likely to report a TT match (see Shalvi et al., 2012, for evidence that honesty requires time).

[22] https://osf.io/bpqz8/.

greater association with an individual's extended self. Wang and Nelson (2014) argue that even if this is seen as a relationship role instead of an extension of self, the bond with touch devices is closer than the bond with other kinds of devices. As this channel focuses on the "self" rather than on the "other", it would also not contradict our null finding between devices in our measures of "pure" social preferences (altruism, trust and cooperation).

In the hierarchy of List (2021), our study constitutes a "Wave 2″ study that delves deeper into the boundary conditions required for the generalizability of decision-making experiments in settings where physical lab experiments are nether the feasible nor natural choice. There are of course many background factors that may influence human behavior in a particular setting. Some factors are likely to be more important than others. The traditional methodological strength of laboratory economic experiments is the ability to apply as much control over those factors as possible. Future work will continue to explore threats to the generalizability of online, field and hybrid experiments, in which control is necessarily reduced. As List (2021) surmises, it is through the discovery of such factors that science progresses.

Data availability

Data will be made available on request.

References

Abdellaoui, M., Driouchi, A., & l'Haridon, O. (2011). "Risk aversion elicitation: Reconciling tractability and bias minimization". *Theory and Decision, 71*, 63–80.

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica: Journal of the Econometric Society, 87*(4), 1115–1153.

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the internet: The effect of $1 stakes. *PloS one, 7*(2), e31461.

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

Andrew, P. (2021). Mobile technology and home broadband 2021. Pew Research Center. URL: https://www.pewresearch.org/internet/2021/06/03/mobile-technology-and-home-broadband-2021/.

Arechar, AA., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics, 21*(1), 99–131.

Dal Bó, P., Fréchette, G. R., & Kim, J. (2021). The determinants of efficient behavior in coordination games. *Games and Economic Behavior, 130*, 352–368.

Bainbridge, W. S. (2007). The scientific research potential of virtual worlds. *Science, 317* (5837), 472–476 (New York, N.Y.).

Bednar, J., Chen, Y., Liu, T. X., & Page, S. (2012). Behavioral spillovers and cognitive load in multiple games: An experimental study. *Games and Economic Behavior, 74*(1), 12–31.

Beine, M.A.R., Charness, G., Dupuy, A., and Joxhe, M. (2020). Shaking things up: on the stability of risk and time preferences. CESifo Working Paper No. 8187. Available at SSRN: https://ssrn.com/abstract=3570289.

Bellemare, C., Bissonnette, L., & Kröger, S. (2016). Simulating power of economic experiments: The powerBBK package. *Journal of the Economic Science Association*, (2), 157–168.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika, 93*(3), 491–507.

Brasel, S. A, & Gips, J. (2014). Tablets, touchscreens, and touchpads: How varying touch interfaces trigger psychological ownership and endowment. *Journal of Consumer Psychology, 24*(2), 226–233.

Braut, B. (2023). Lab vs online experiments: Gender differences. Working Paper. Available at SSRN: https://ssrn.com/abstract=4246559.

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.

Cahlíková, J., & Cingl, L. (2017). Risk preferences under acute stress. *Experimental Economics, 20*(1), 209–236.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436 (New York, N.Y.).

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

Charness, G., Gneezy, U., & Kuhn, M. A. (2013). Experimental methods: Extra-laboratory experiments-extending the reach of experimental economics. *Journal of Economic Behavior & Organization, 91*, 93–100.

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.

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

Cochard, F., Le Gallo, J., Georgantzis, N., & Tisserand, J. C. (2021). Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game. *Journal of Behavioral and Experimental Economics, 90*, Article 101613.

Dickinson, D. L., & McEvoy, D. M. (2021). Further from the truth: The impact of moving from in-person to online settings on dishonest behavior. *Journal of Behavioral and Experimental Economics, 90*, Article 101649.

Embrey, M., Fréchette, G. R., & Yuksel, S. (2018). Cooperation in the finitely repeated prisoner's dilemma. *Quarterly Journal of Economics, 133*(1), 509–551.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics, 14*, 583–610.

Feltz, C. J., & Miller, G. E. (1996). An asymptotic test for the equality of coefficients of variation from K populations. *Statistics in Medicine, 15*(6), 647–658.

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), 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.

Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). LIONESS Lab: A free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association, 6*(1), 95–111.

Gupta, N., Rigotti L. & Wilson. A. (2021). The experimenters' dilemma: Inferential preferences over populations. Working Paper. ArXiv Preprint, arXiv:2107.05064. Available at: https://doi.org/10.48550/arXiv.2107.05064.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42* (4), 1009–1055.

Hein, W., O'Donohoe, S., & Ryan, A. (2011). Mobile phones as an extension of the participant observer's self: Reflections on the emergent role of an emergent technology. *Qualitative Market Research, 14*(3), 258–273.

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

Himelein, K. (2021). Improved targeting for mobile phone surveys: A public-private data collaboration. Retrieved from https://blogs.worldbank.org/impactevaluations/improved-targeting-mobile-phone-surveys-public-private-data-collaboration-guest.

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

Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior & Organization, 93*, 328–336.

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

Jorrat, D. (2021). Recruiting experimental subjects using WhatsApp. *Journal of Behavioral and Experimental Economics, 90*, Article 101644.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives, 5*(1), 193–206.

Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior, 102*, 433–444.

Kirchler, M., Andersson, D., Bonn, C., Johannesson, M., Sørensen, E.Ø., Stefan, M., et al. (2017). The effect of fast and slow decisions on risk taking. *Journal of Risk and Uncertainty, 54*(1), 37–59.

Kocher, M. G., & Sutter, M. (2006). Time is money—Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization, 61*(3), 375–392.

Kocher, M. G., Pahlke, J., & Trautmann, S. T. (2013). Tempus fugit: Time pressure in risky decisions. *Management Science, 59*(10), 2380–2391.

Konovalov, A., & Krajbich, I. (2019). Revealed strength of preference: Inference from response times. *Judgment and Decision Making, 14*(4), 381–394.

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review, 53*(1), 1–18.

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.

List, J. A. (2021). Non est disputandum de generalizability? A glimpse into the external validity trial. Working Paper. Available at NBER: http://www.nber.org/papers/w27535.

Lohmann, P., Gsottbauer, E., You, J., and Kontoleon, A. (2021). Social preferences and economic decision-making in the wake of COVID-19: Experimental evidence from China." Working Paper.Available at SSRN: https://ssrn.com/abstract=3705264.

Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review, 34*(1), 78–94.

Melumad, S., & Pham, M. T. (2020). The smartphone as a pacifying technology. *Journal of Consumer Research, 47*(2), 237–255.

Mograbi, E (2022). Decision-makers are more impulsive on smartphones than on computers. *Journal of Behavioral and Experimental Economics, 100*, 101916.

D. Parra "The impact of lying aversion and prosociality on cheating." Available at SSRN 3960906 (2022).

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54* (4), 1643–1662.

Prissé, B., & Jorrat, D. (2022). Lab vs online experiments: No differences. *Journal of Behavioral and Experimental Economics, 100*, Article 101910.

Rajagopal, P., & Rha, J. Y. (2009). The mental accounting of time. *Journal of Economic Psychology, 30*(5), 772–781.

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

Shachat, J., Walker, M. J., & Wei, L. (2021a). How the onset of the Covid-19 pandemic impacted pro-social behaviour and individual preferences: Experimental evidence from China. *Journal of Economic Behavior & Organization, 190*, 80–494. a.

Shachat, J., Walker, M. J., & Wei, L. (2021b). The impact of an epidemic: Experimental evidence on preference stability from Wuhan. *AEA Papers and Proceedings, 111*, 302–306. b.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science, 23*(10), 1264–1270.

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.

Silver, L., Smith, A., Johnson, C., & Jiang, J., et al. (2009). Mobile connectivity in emerging economics. Pew Research Center. URL: https://www.pewresearch.org/internet/2019/03/07/mobile-connectivity-in-emerging-economies/.

Snowberg, E., & Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review, 111*(2), 687–719.

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., et al. (2013). Intuition and cooperation reconsidered. *Nature, 498*, E1–E2.

Wang, Z., & Nelson, M. R. (2014). Tablet as human: how intensity and stability of the user-tablet relationship influences users' impression formation of tablet computers. *Computers in Human Behavior, 37*, 81–93.