THE EXPERIMENTERS' DILEMMA: INFERENTIAL PREFERENCES OVER POPULATIONS

LUCA RIGOTTI ⓡ ALISTAIR WILSON ⓡ NEERAJA GUPTA

Abstract. We examine the experimenter's preferences over different populations using statistical power under a fixed budget as the stand-in for the researcher's utility. We consider five populations commonly used in experiments by economists: undergraduate students at a physical location, undergraduate students in a virtual setting, Amazon MTurk "workers", a filtered MTurk subset from CloudResearch, and Prolific. Focusing on noise due to inattention, observation costs dominate the comparisons, with the larger online population samples superior to the smaller lab samples. However, once we factor in responsiveness to treatment, the lab samples have greater power than either MTurk or Prolific.

1. Introduction

Economic experiments have become a crucial tool for uncovering facets of economic decision making that would be veiled in naturally occurring data. Over the previous half-century, the dominant paradigm for qualitative tests of economic theories was the laboratory experiment: a set of typically undergraduate participants is recruited to a fixed time slot at a physical location, where a tailored set of monetary incentives is then offered to examine and identify the hypothesis. But in the past decade, a number of online populations have emerged to conduct economic experiments. These online populations offer an array of positives: greater convenience, lower barriers to entry, a large number of participants to draw from with greater representation of the wider population. Moreover, for researchers with finite budgets they offer another benefit: a typically much-lower cost per observation than the equivalent lab study. However, one concern that is often raised is that the experimenter has reduced control over the online participants. As participants recruited from online populations will take part in the study on their own devices, and often at their own pace, their possible divided attention and an incentive to complete the study as quickly as possible can lead to noisier data, potentially washing out treatment effects and leading to false-negatives. In contrast, for laboratory studies distractions and study-timing can be controlled, though normally with the greater time commitments/focus typically requiring higher participant payments.

Prompted by some of these trade-offs, we try to assess the inferential bang for your buck across experimental populations, running a horse-race across five different populations

Date: March 2023.

We would like to thank David Huffman, Sera Linardi, and Lise Vesterlund for their helpful comments and feedback as well as participants of the Economic Science Association 2022 North American Meeting. IRB approval for the experiment was obtained at University of Pittsburgh.

using a common task. The exercise starts out with a simple motivating idea: an experimenter is attempting to measure a qualitative treatment effect, to be identified through a difference in means. However, the experimenter has a fixed budget to spend on the project. While recruiting from a laboratory sample might have low noise, costs per observation will be high, and so she will get a smaller sample. In contrast, while an online population might have a noisier response, the lower costs per observation mean she can have a much larger sample for the same overall budget. By assessing our five populations over the ability to detect a treatment effect, and scaling per-participant payments to be representative of the population, we examine the inferential tradeoffs directly. Our results provide a series of insights for experimenters considering these different populations, identifying the merits of each through a coherent inferential lens.

Our first population is a standard laboratory sample, with undergraduate students recruited to a physical lab experiment. The second population holds constant the standard lab sample, but participants are recruited to participate in the experiment online (a mode that became common during the pandemic). Our final three populations use online participants: Amazon's Mechanical Turk (MTurk), likely the most ubiquitous online labor population; the CloudResearch approved list (CloudR), a subset of MTurk workers selected through various measures of attention and data quality; and Prolific, an emerging platform favored by many recent experiments with a more-curated set of participants.[1]

We assess these five populations using four two-player strategic games—where a strategic setting helps to ensure a minimal degree of attention to understand the induced incentives. However, the selected games vary in the degree of strategic sophistication that is actually required to make an informed decision. In particular, for two of our selected games the strictly dominant action for each player also leads to the socially efficient solution. As such, we interpret the behavior in these first two games as a sanity check, a screen to detect whether the induced incentives are understood by the participants. In contrast, for our other two games, we do generate a clear strategic tension, with a prisoner's dilemma (PD) payoff structure. However, here we vary the PD-game payoffs according to the Rapoport index, a behavioral theory for understanding cooperation that leads to a comparative-static hypothesis with clear support in the prior literature. This behavioral comparative static becomes our final yardstick, where the power we have to assess it allow us to compare the different populations from an inferential point of view.

To fairly compare the inferential power by population, our paper enforces a novel constraint: a fixed pot of money available to spend. As such, by respecting the ecologically valid differences in the minimum and expected payments per participant on the different populations—achieved by scaling down the probability of payment for the game decisions, and through the participation payments—the budget constraint leads to differences in the number of observations recruited from each population. As such, even if participants from the online populations are more inattentive, the population may have greater inferential power because the experimenter can collect a larger sample for the

[1]The exact procedure for selecting the approved list of participants is not disclosed by CloudResearch but their claim on the publicly available website is a strict vetting criteria that keeps the population demographically representative of the overall MTurk population.

same overall budget. We formalize the experimenter's *preference* over the different populations using a T-statistic. Supposing that the experimenter is driven by the desire to maximize the power of her inferential tests (minimizing the probability of a type II error) we can coherently compare populations. In particular, we consider populations as varying over the expected per-observation cost and the attenuation in the expected size of the treatment effect. The expected power of a T-test of a difference in means allows us to compare populations from an inferential point of view. Does the population with the large sample but where inattention has wiped out much of the treatment effect fare better than the population with the smaller sample but with a larger effect?

Using data across our four games—as well as an additional framing variation on which action is presented first—we examine how the populations vary across three outcomes: (i) the cost per independent observation; (ii) the inattention to the incentives, identified through the fraction of participants choosing the dominated action in the games without a strategic tension; and (iii) the size of the treatment effect when we compare the cooperation rates in the two PD games, which we decompose into the effect from inattention, and a reduced elasticity of response. From these measurements, we then examine the experimenter's preference over populations, as if analyzing a standard consumer choice problem: drawing both iso-power contours across different population characteristics under a fixed budget (analogous to the indirect utility) and the dual iso-budget contours under a fixed power level (analogous to the expenditure function).

Our results indicate the tradeoffs across the different platforms are not trivial. Fifty-five percent of the MTurk participants make inattentive decisions that are unreactive to the induced incentives (either through random choice, or by choosing the first-available option). However, because the average costs per observation are very low, at \$3.01, MTurk leads to a very large sample. At the other extreme, the average observation costs in the physical laboratory are high at \$22.08, but inattention to the incentives is much lower, at 14 percent. Holding the experimental budgets constant, we examine how the inferential power from a high-attention sample of 75 laboratory participants compares to the low-attention sample of 550 participants on MTurk. (And more trivially, how each compares to the 540 and 380 high-attention participants on CloudResearch and Prolific, respectively). Assuming an unchanged expected effect size for the attentive sample—and so purely considering attenuation in the response due to inattentive participants—we find a clear inferential ranking in favor of the online populations. CloudResearch and Prolific are clearly dominant populations from this point of view, followed by the very inattentive MTurk population, which, if the attentive effect size was held constant, would still yield greater power than either the physical or virtual lab samples.

However, the proportion of inattentive/noisy participants is not the entire story. Moving to the actual behavior in the PD comparative-static assessment, our online populations also exhibit diminished effect sizes. In particular, for two of the online populations, MTurk and Prolific, a reduced response elasticity essentially eliminates the treatment effect (indeed, MTurk moves in the *opposite* direction). Because of this, both lab samples end up being inferentially superior. Despite low costs per observation and relatively high attentiveness to the incentives, the Prolific sample ends up being too inelastic for the behavioral test to be well-powered. What the prior literature led us to think would be a

moderate treatment effect in the laboratory, ends up being too subtle to detect on MTurk and Prolific.[2] However, the CloudResearch sample maintains its dominance over the two laboratory samples. Although there is a reduction in the size of the treatment effect, the reduced elasticity in the response is more than compensated for with the increased sample size.

In terms of external relevance, some of our net-effect results could certainly be specific to social dilemmas. Moreoever, many experiments conducted on online platforms require lower attention/comprehension than the strategic setting we examine. However, our results on inattention identified through the response in games without any real strategic tensions are stark enough that the conclusion that MTurk is dominated by the two more-curated online populations likely has more-general implications. For hypotheses examining relatively stark economic institutions, the ability to collect many, many observations for the same fixed budget does favor the more-curated online populations (CloudResearch and Prolific) over the standard lab populations. However, our study also points to the potential benefits of laboratory studies in more-nuanced or complicated economic environments. Despite the expense per observation, lab samples generate larger elasticities in the response, and may be preferable in many contexts given their ability to extract consistent results. Indeed, our laboratory samples are consistent not only in the size of the response between the virtual and physical populations, but also in reproducing the quantitative predictions we would have expected from prior laboratory studies.

The rest of the paper is organized as follows: Section 2 summarizes the related literature and highlights our contributions, section 3 discusses our experiment design, implementation, hypotheses and presents our model for studying inferential preferences across populations. Section 4 presents results for the hypothesis, while Section 5 compares the five populations from the lens of inferential power, and Section 6 summarizes and concludes.

2. Related Literature

A number of studies have compared MTurk with the laboratory population, focusing primarily on whether the empirical regularities observed in the laboratory can be replicated. Our paper's novelty is in making the focus more explicitly on the effective power within each population, taking into account researchers' financial constraints. In particular, we consider how the possibilities for many more independent observations from cheaper online populations interact with the potential for noisier data and/or a more inelastic response.

One of the earliest works examining the use of MTurk in online experiments is Paolacci et al. (2010), replicating three classical behavioral economics results (the Asian disease, Linda and Physician problems), and finding no significant differences between the populations. Similarly, Horton et al. (2011) find no significant differences in cooperation between an MTurk sample and the experimental lab literature on one-shot PD games.

[2] In a robustness exercise, we show that Prolific does detect a significant response to the social dilemma tensions, but only if the induced size of treatment is substantially increased. This extension further cements our interpretation that this is a reduced response elasticity.

In Goodman et al. (2013) an MTurk sample replicates standard decision-making biases.[3] More recently, Thomas and Clifford (2017) suggests that the strict exclusion criterion for "problematic" participants can reduce statistical noise without introducing sampling bias. In Arechar et al. (2018) the researchers uncover the same basic behavioral patterns of cooperation and punishment in a repeated public good experiment in both the lab and MTurk, although dropout can be a challenge to conduct interactive experiments on MTurk.

Snowberg and Yariv (2021) elicit and compare a set of behavioral characteristics using a survey administered to an entire undergraduate cohort, a self-selected laboratory sample, and a representative sample of US participants recruited online from MTurk. Although they look at many different behaviors, their study includes two one-shot PD games (though with the same effective incentives). Similarly to our findings, they do find significant differences in cooperation *levels* between populations for the PD game, the online sample being more cooperative; however, they do find comparable comparative static responses between populations in many other behaviors. Their other overarching results are that behavioral characteristics are similarly correlated across populations and that noise (measured by differences in the response for duplicate elicitations) is higher for online populations. We confirm this last result when it comes to MTurk but not for the CloudR or Prolific samples, though our own measures of noise are based on responses to a more basic check of rationality and a frame change. Our focus though is on inferential power, where we take the existence of the effect as given, and instead focus on the effective power of the population under a fixed researcher budget.

Our findings match growing concerns over a decrease in the quality of MTurk data over the past two years. The literature highlights the limitations of MTurk, including, but not limited to, anticipation of deception by researchers, repeated participation in similar tasks that leads to knowledge acquisition and a resultant change in behavior, unmeasurable attrition, and programmed bots (Hauser et al., 2019; Chmielewski and Kucker, 2020). Aruguete et al. (2019) identify MTurk workers as being more likely to fail attention checks designed to measure haste and carelessness in responses than college students (although our noise measures are in a sample that has successfully passed an understanding quiz).

These growing concerns about the data quality retrieved from MTurk workers has led to growth in the infrastructure for running online experiments (see Fréchette et al. (2022) for an analysis of the trends in experimental literature and discussion about growing use of experimentation platforms. A popular alternative to MTurk has come up in the form of more-curated sub-populations such as the CloudResearch approved list and research-specific platforms like Prolific. While Eyal et al. (2021) find MTurk workers to show an alarming rate of inattentiveness relative to more-curated online populations, how these populations fair relative to the standard lab is an open question. Moreover, out framework outlines a clear framework for assessing populations through inferential power.

[3]MTurk participants exhibit: (i) present bias; (ii) risk-aversion for gains and risk-seeking for losses; (iii) show delay/expedite asymmetries; and (iv) show the certainty effect.

Some of our results also replicate prior findings, though here through a lens focused on inferential power through the fixed budget. Although we do find lower inferential power on our final comparative static assessment for the data from both MTurk and Prolific, our inattention measures point to both Prolific and CloudResearch as having similar attention to the laboratory participants. Instead of noise, the low power on the PD-game comparison on Prolific seems to come about through a much-smaller response elasticity, where the population in general offers much more promise. Finally, the CloudResearch approved list of participants seems to circumvent both the noise and inelastic response concerns, where its relatively low cost per observation suggests it is inferentially superior.

3. Experiment Design

Our experiment has a $5 \times 4 \times 2$ design over:

Population: We use five experimental populations: (i) undergraduate students recruited from the University of Pittsburgh for a physical experiment (the *Lab* sample); (ii) undergraduate students again recruited from the University of Pittsburgh, but now for a virtual online setting where the experiment was conducted online (the virtual laboratory or *VLab* sample); (iii) online workers recruited from Amazon's online labor market Mechanical Turk (the *MTurk* sample); (iv) online workers from the CloudResearch approved-list on Mechanical Turk (the *CloudR* sample); and finally (v) online workers recruited from research platform Prolific (the *Prolific* sample).[4]

Strategic environment: We ask participants to make a binary action choice in four symmetric two-player games (with payoffs indicated in Table 1 and further details below). Although the experiment uses an A/B action labeling, we use a C(ooperate)/D(efect) labeling in the paper, as all four games have joint-cooperation as the socially efficient outcome.

Frame: Our framing change shifts the order in which the actions are presented to the participants, permuting the ordering of the C and D actions in the given game tables (so shifting the assignment of C/D to the A/B choice labels shown to participants).

3.1. **Incentives and Implementation.** Our design collects data across ten between-subject treatments, the five populations and the frame-change over the ordering of the cooperate/defect decision. Each participant is asked to submit their choice between the two actions (A or B) in the four games (though presented to them in a random order). Games are presented to participants as a table with four rows (one for each possible action profile) ordered as (A,A), (A,B), (B,A) and (B,B) for the self/other action. The re-framing therefore moves the socially efficient (C,C) entries from the top row in the table (labeled (A,A) in the experiment) to the bottom row (labeled (B,B)).

[4]CloudResearch was formerly known as TurkPrime, which provided tools for online study recruitment on MTurk. Both our Mechanical Turk samples are collected using CloudResearch, but MTurk sample draws participants from the unfiltered population, including but not limited to those on the "approved list."

TABLE 1. Experiment Design

Panel A	Payoff π_i on action (a_i, a_j)			
	(C,C)	(C,D)	(D,C)	(D,D)
Game PD1 ($\rho = 0.50$)	\$21	\$2	\$28	\$8
Game PD2 ($\rho = 0.71$)	\$19	\$8	\$22	\$9
Game Σ-DOM1	\$17	\$12	\$16	\$10
Game Σ-DOM2	\$15	\$16	\$10	\$11

Panel B	Participants & Expenditure				
	Lab	VLab	MTurk	Cloud-R	Prolific
Participants:					
C-first frame	50	50	368	374	250
D-first frame	24	24	180	167	135
Total	74	74	548	541	385
Expenditure:					
Total	\$1,634.00	\$1,609.30	\$1,647.32	\$1,746.90	\$1,679.76
Per observation	\$22.08	\$21.75	\$3.01	\$3.23	\$4.36

Note: Participant observations exclude those who failed to answer comprehension questions correctly. However, total expenditure includes fixed-payments made to participants who are dismissed on account of over-booking of sessions for the university samples as well as to those dismissed from the online studies for answering the comprehension question incorrectly.

For our horse race between populations, our initial plans were for a budget of \$1,500 per population. However, we ran the Lab study first, and this ended up being more expensive than planned at just over \$1,600. We therefore matched all other samples to this approximate budget (where our later inferential analysis will match the budgets exactly). Within each population, we aimed to spend the budget across the *C*-first/*D*-first frames at a two-to-one ratio, in case pooling the samples was not an option, and so the *C*-first sample would provide a higher power sample.

Our Lab sample consists of undergraduate students recruited from the University of Pittsburgh undergraduate population. Participants were offered a \$6 fixed payment, and were randomly paid for one decision from the four games they made a choice in, after being matched to an anonymous partner in the session.[5] Payments for each action combination in the four games are shown in Table 1. Our total expenditure for 74 laboratory observations was \$1,624, where this figure includes \$72 spent on show-up fees for unused participants.[6]

[5]The experiment was programmed in oTree (Chen et al. (2016) and conducted at the physical Pittsburgh Experimental Economics Laboratory (PEEL) space, where our choice of payments match the PEEL populations conditions for participant payments.

[6]Our methodology here is to include all variable costs for the study. One possible critique is that we do not account for the financial costs of setting up and running the PEEL lab, where our approach is to treat

The virtual laboratory (VLab) experiment follows the in-person Lab experiment very closely, where participants are again recruited from the University of Pittsburgh undergraduate population. However, this experiment was conducted entirely online over Zoom, following the online laboratory protocols that best mirror in-person protocols (Danz et al., 2021). The total expenditure to collect 74 observations was $1,609.30 with the per observation cost being $21.75.[7]

The per-observation costs for the samples drawn from the undergraduate population are therefore approximately $22 in both cases. While we could have offered the same incentives to the online participants in our MTurk, CloudR and Prolific samples, this would have represented a substantial break from the norm, where typical payments are much smaller. As our aim was to match the ecologically valid incentives being offered on each population, and to account for the shifts in the sample size that come from online studies, we scaled down the incentives for our online populations substantially.[8] Participants in our MTurk sample were given a $0.50 fixed fee for taking the 'HIT', and a further $0.50 if they correctly answered a comprehension question to show they understood the instructions.[9,10] While the dollar payments within each game table exactly match the Lab sample, as given in Table 1, the expected payments are scaled down by changing the *likelihood* of payment. One out of every ten pairs of participants are paid for their game decisions, for one of the four game tables.[11] In total our MTurk sample contains data from 548 individuals with a total cost of $1,649 ($3.01 per participant).

The CloudResearch sample follows near identical procedures to the MTurk experiment in terms of incentives, where the only difference was that the population was drawn from

these as sunk costs. As such, inferential comparisons across populations are from the point of view of a researcher who has free access to a turnkey lab space.

[7]Total expenditure on the VLab sample also includes $1.30, the cost of deploying the oTree experiment using a Heroku server. As with the standard Lab sample, our total cost here also include show-up fees of $6 paid to recruited participants who attended an already full session and were turned away.

[8]Focusing purely on the average earnings of the participants (so excluding platform fees and other costs), and dividing by the average time taken to complete the study, the effective wage rates are actually remarkably similar. Across the Lab, VLab, MTurk, CloudR and Prolific samples, the average wage rates are $31.66, $31.13, $31.81, $38.27 , and $40.25 per hour, respectively.

[9]Mirroring standard procedures on online platforms, participants who failed to answer the comprehension questions did not proceed any further. These participants are thus excluded from our sample in the analysis (and our counts of the sample size N); however the costs for these participants, as well as the platform fees for the total samples as charged by Amazon (20 percent) and Cloud Research (4 percent of fixed fee Litman et al., 2017) are included in the total expenditure.

[10]The MTurk/CloudR/Prolific experiments are coded using Qualtrics and recruited participants have the following restrictions: located within the US, and with a 95 percent or better approval rate.

[11]Our instructions give participants a clear rule used to conduct the randomizations, where all draws are made using public randomizations outside of the researchers' control (here public state lottery draws made the evening after the participants made decisions). Moreover, participants are told that if selected for payment, they would be matched to another for-payment participant, where the final bonus-payment would be determined by the joint choices of the pair. As such, conditional on payment selection, the externalities and game-selection chances are identical to our Lab/VLab study.

the CloudResearch approved list. The total cost for the CloudR sample of 541 participants was more expensive than planned at $1,746.[12] The total cost per participant on CloudResearch was therefore $3.23.

Finally, our Prolific sample follows an identical process for the induced incentives in the game payments as the MTurk and CloudR samples. However, Prolific platform rules required larger minimum payments, and so we increased the fixed payment to $1.60.[13,14] The total expenditure on Prolific was $1,680 for 385 observations, so $4.36 per observation.

Our design asks the following core question: given the differential costs for each observation, and the potential quality differences in the data collected, which population is superior? The unfiltered MTurk sample offers the potential for a large number of observations from a fixed budget. However, it is also potentially the noisiest by reputation. On the other extreme, the laboratory is the most expensive per observation. But a question remains on whether this additional expense is warranted through higher quality data, and also whether this is affected by taking standard lab populations from the physical lab to the virtual one. Finally, the more-vetted/curated populations on Prolific and CloudResearch are relatively cheap, and so if they have substantially reduced inattention over MTurk, this might substantially improve the researcher's inferential power.

3.2. **Hypothesis.** We first outline the designed features of the four games participants make choices in. All four games are dominance solvable in terms of the individual payoff, where we relabel the standard notion of strict payoff dominance as:

Definition 1 (i-Dominated action)**.** *Action a is i-dominated if there exists another action a' that gives player i a higher payoff for any selected action of the other player.*

The i-dominant action is to defect in games PD1 and PD2 and to cooperate in Σ-DOM1 and Σ-DOM2. However, there is a large body of behavioral evidence suggesting that many individuals' preferences are other-regarding, and thus sensitive to tradeoffs between the individual payoff and social efficiency. The behavioral literature suggests that many individuals will choose i-dominated actions so long as it improves social efficiency (as measured by the sum of payoffs). As such, a stronger version of dominance can be composed using both the individual and total payoffs to suggest the behavior in games without any tension between the individually and socially efficient actions:

[12]While using the approved list feature on Cloud Research is free of cost, the fees assessed using the platform increased to 10 percent from the earlier 4 percent in the MTurk experiment, on top of the Amazon 20 percent fees levied on all participant payments. The total cost here again includes all fixed fees to participants who incorrectly answered the comprehension question, but where these participants are again excluded in the analysis.

[13]Participants failing the comprehension check received this larger fixed payment, but were not given the chance to get incentive payments from the four games. The total expenditure includes costs for these excluded participants as well as the 33 percent fee imposed by the platform.

[14]We conducted a pilot of 20 participants on Prolific to understand the median time taken, as the minimum fixed-fee payment for the Platform was a function of this time. However, the incentives for this pilot study were different from the calibrated ones in the main study, and so for the sake of comparability, neither this pilot data, nor the costs for acquiring it are considered in our analysis.

Definition 2 (Σ-Dominated action)**.** *Action a is Σ-dominated if there exist another action a′ such that a is both i-dominated by some action a′, but where the sum of the player payoffs is also smaller under a than under a′, for any selected action of the other player.*

Games Σ-DOM1 and Σ-DOM2 are constructed so that the D action is Σ-dominated by C. As such, the D action, if understood, is hard to justify with any other-regarding preference interested in social efficiency.[15] Taking as given that participants are driven by a preference that is strictly increasing in both the own and social monetary reward, we will assume that any Σ-dominated choices are a consequence of the participant not fully understanding the environment. Games Σ-DOM1 and Σ-DOM2 thereby provide our first attentiveness measure, where our null hypothesis is that:

Hypothesis 1 (Dominated-play null)**.** *The five populations have similarly small proportions of Σ-dominated play.*

Our second attentiveness measure for the participants is based on the frame change. One plausible heuristic for inattentive participants is that they move as quickly as possible through the offered choices without understanding the incentives, selecting the first-available option. If the C action is always presented first, we might fail to detect inattentive choices using the Σ-dominant notion. By altering the frame to present the D action first, and changing nothing else about the offered incentives, we can identify this first-option heuristic by looking at the differences in Σ-dominated play across the frame change.

Hypothesis 2 (Reframing null)**.** *The five populations have similar shifts in Σ-dominated play across the re-framing.*

Our first two hypotheses concern the participant's attentiveness, examining whether they respond to money as a reward medium, as the two games we assess behavior in do not have a strategic tension between individuals. In contrast, for our final hypothesis—and the one that we will build our inferential horse race over—we induce a real strategic tension in the two games we will compare, between what is socially efficient (joint cooperation) and what is individual payoff dominant (defection). Games PD1 and PD2 are both prisoner's dilemmas; however, we change the intensity of the strategic tensions across the two games. While pure payoff maximizing Nash would not predict a difference, the longer history of behavioral results suggests a clear behavioral comparative static.

The behavioral theory underlying our comparative-static hypothesis make uses of a parametric index known as the Rapoport ratio (cf. Rapoport, 1967), which has been shown to be predictive of cooperation in PD games. The Rapoport ratio is given by the following function of the PD-game payoffs:

$$\rho = \frac{\pi_i(C,C) - \pi_i(D,D)}{\pi_i(D,C) - \pi_i(C,D)}.$$

[15] Game Σ-DOM1 is designed to satisfy an even stronger ordering: the Pareto order. However, we do not find that this has any additional predictive content, so we focus purely on Σ-Dominance, as this will provide us better identification of random play, with two binary choices rather than one.

The hypothesis from the literature is that the expected rate of cooperation will be increasing in the Rapoport ratio. In our experimental setting, the PD1 and PD2 games have Rapoport ratios of 0.50 and 0.71, respectively. As such, we would expect cooperation to be greater in PD2 than PD1. Beyond simply assessing noise due to inattention in the first two games, the wider inferential aim of our study is to identify a significant *directional* effect for the behavior between games PD1 than PD2. Across our five populations we expect to find the following directional comparative static (assessed against a null of no effect):

Hypothesis 3 (PD comparative static)**.** *Following the Rapoport ratio prediction, each of the five populations will have more cooperation in game PD2 than PD1.*

3.3. **Inferential preferences.** The behavioral PD comparative static hypothesis above is used to generate a horse race between the populations over their inferential power. Given that participants' decisions are binary, any comparative static test over the differences in cooperation between PD1 and PD2 will simply be a function of the observed cooperation rates, and the common sample size N in the two games.[16] In a standard comparison without any further control, the pooled T-statistic for inferences on a null hypothesis of no effect, uses the sample size N and the two cooperation rates of P_1 and P_2 (in games PD1 and PD2, respectively) as follows:

$$T(P_1, P_2, N) = \frac{\sqrt{2 \cdot N} \cdot (P_2 - P_1)}{\sqrt{\left(\frac{P_1+P_2}{2}\right)\left(1 - \frac{P_1+P_2}{2}\right)}}$$

For a qualitative alternative hypothesis that there is more cooperation with a higher Rapoport ratio, we therefore want the T-statistic to be greater than 1.64 to attain 95 percent confidence for the one-sided hypothesis (90 percent for a two-sided test).

Modeling the number of cooperation decisions within the samples as $N \cdot P_1$ and $N \cdot P_2$ via binomial draws from N with true proportions of p_1 and p_2, respectively, it is possible to calculate the chance the experimenter will make a type-II error on this T-test (assuming that $p_2 > p_1$is true). To make things concrete here, we use as a baseline the results from a recent laboratory study of one-shot PD games that varies the Rapoport ratio Charness et al. (2016). Using their results, we can generate an out-of-sample prediction for the inferential power in the Lab. We expect a cooperation rate difference between games PD1 and PD2 of 17 percentage points.[17]

If all of our populations have the same expected cooperation rates and zero attenuation due to inattention, then the power of our tests will simply be a function of the sample-size N. For any fixed experimental budget, all else equal, whichever population had the

[16]Greater statistical power can be generated if we also use the within-subject nature of the data, however, for simplicity we focus on a more-standard between-subject comparison.

[17]We estimate this from the four treatments in Charness et al. via a logit model with the Rapoport ratio as the sole predictor. The estimated model predicts a cooperation rate given by:

$$\text{Coop}(\rho) = \frac{1}{1 + 5.66 \cdot e^{-3.32\rho}},$$

where this implies cooperattion rates of $p_1 = 0.481$ and $p_2 = 0.653$

cheapest observations would yield the greatest power. However, this calculation assumes that the quality of the data from each populations are the same. But, motivated here by a wariness about online samples, and their potential for reduced control, it may be that the effect sizes are washed out in the online samples. For example, participants in the online studies may choose to multi-task while taking part, and therefore fail to pay enough attention to the incentives to make considered choices. As such, while potentially cheaper, if a large enough fraction of the participants are inattentive, there may not be any significant response to treatment. To model this, we consider each population as having two fundamental properties: a dollar cost per observation c; and a noise/attenuation parameter γ that reduces the effect size.

We model the attenuation parameter γ as affecting the population-level expected behavior in both games, attenuating it towards a coin flip choice as γ tends to one. The expected cooperation rate in game G_j with attenuation rate γ is therefore modeled as $\gamma \cdot \frac{1}{2} + (1-\gamma) \cdot p_j$.[18]

Considering each population as an observation-cost/attenuation-rate–bundle (c, γ), we can model the experimenter's *preference* as we would in a consumer-choice problem. Here we put statistical power in place of the consumers' utility function, so that Population A is preferred to Population B under a fixed budget m, if the probability of making a Type-II error via the T-statistic in (3.3) is smaller for population A.[19] Using this idea in Figure 1 we indicate the experimenter's indifference curves over (c, γ) bundles for the Rapoport Hypothesis 3. In particular in panel (A) we indicate iso-power contours under a fixed experimental budget of \$1,650 (the approximate budget in each population), analogous to thinking about the indirect utility function in consumer choice. In contrast, in panel (B) we indicate iso-budget lines under a fixed power level (90 percent for the two-sided test), analogous to the expenditure function in the dual consumer choice problem.

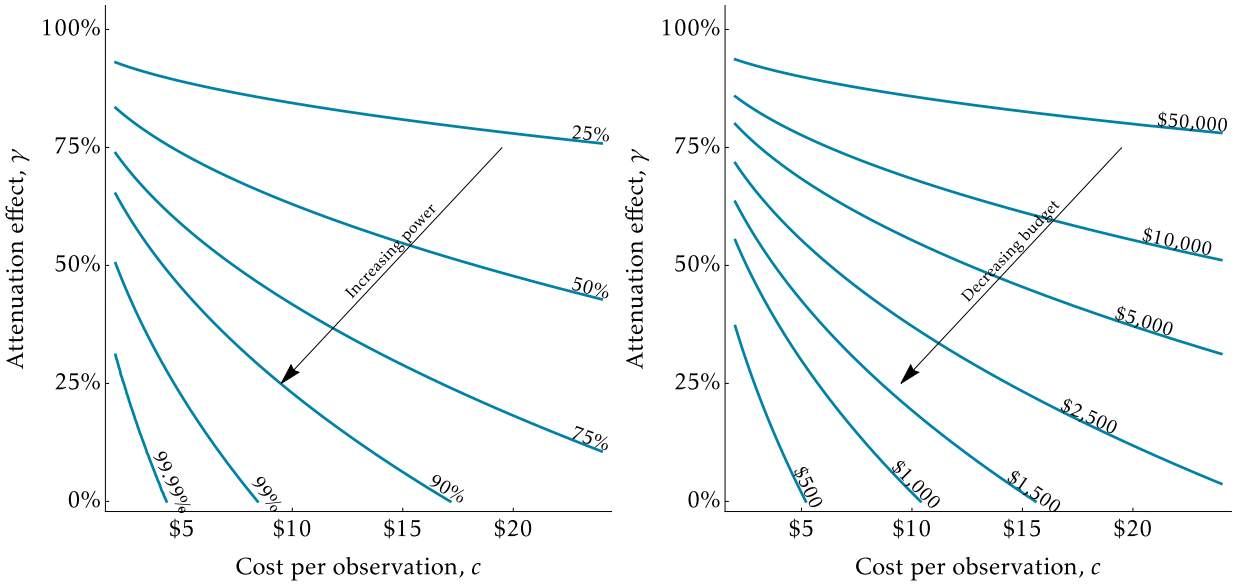
The final assessments of our horse race will therefore be over the populations' inferential power. Using both inattention in the response (assessed through Σ-dominant play) as well as changes in the effect size (measured through the observed cooperation-rate difference between games PD1 and PD2), we will attempt to fairly assess each population from the experimenter's point of view.

4. Results

We now summarize the results from the experiments, before presenting evidence for them: (i) the physical laboratory sample, the CloudResearch sample and the Prolific sample are all relatively similar over the fraction of participants making Σ-dominated choices

[18] While there will be second-order effects via the sample variance in the denominator of (3.3), the main effect of the parameter γ is to scale down the size of the expected difference in behavior in the numerator to $(1-\gamma)(p_2 - p_1)$. As such, focusing purely on the numerator, γ can also represent different forms of attenuation. While our initial focus will be on inattention pushing the response in both games towards a coin flip, the parameter can also be interpreted as a reduction in the response elasticity, scaling down the expected treatment effect from $(p_2 - p_1)$ to $(1-\gamma) \cdot (p_2 - p_1)$.

[19] While satisfying local non-satiation, both, γ and c are 'bads' from the experimenters point of view. Alternatively, one can consider preferences over 'goods' by considering a sample-size/signal bundle $(N, 1-\gamma)$, with $N = m/c$, though the final inferences will be the same.



(A) Iso-Power Contours ($1,650 budget) (B) Iso-Budget Contours (90 percent power)

FIGURE 1. experimenter inferential preferences: Noise versus Cost

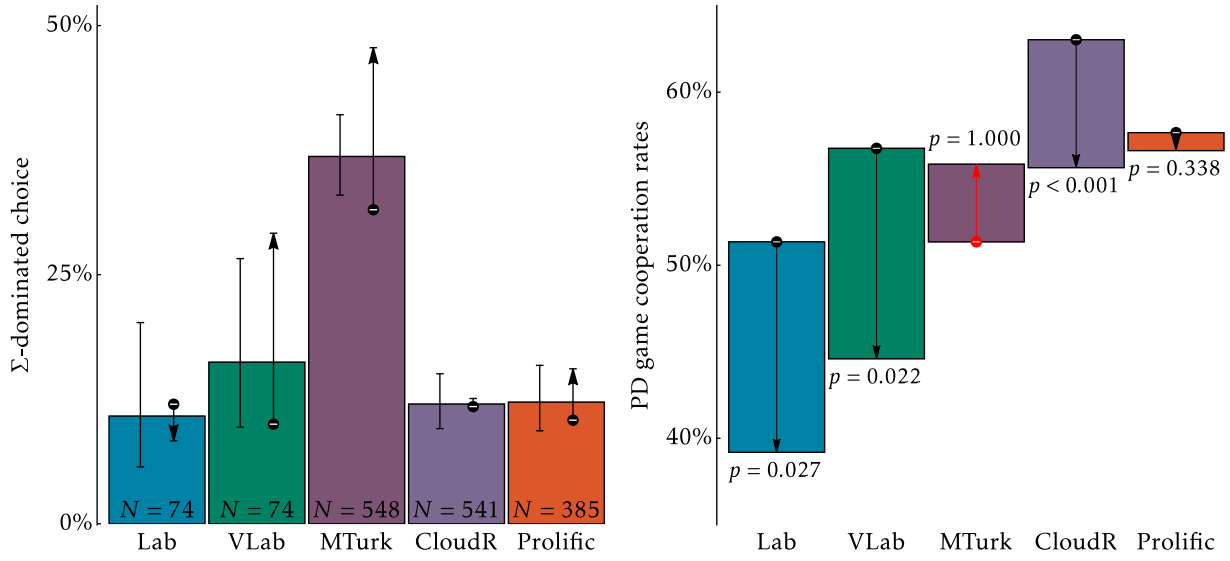
Note: Panel (A) shows iso-power contours (where labels indicate the two-sided probability of rejecting the null) for an experiment with a $1,650 budget, while panel (B) shows iso-budget contours for two-sided test at 90 percent power using a two-sample T-test, with the PD cooperation rates derived from Charness et al. (2016).

(~11–12 percent), while this proportion is slightly larger for the virtual-lab sample (~16 percent) and substantially larger in the MTurk sample (~37 percent); (ii) Changing the order of actions has no significant effects in the Lab, CloudR, and Prolific samples, while the VLab and MTurk samples do exhibit swings (19 and 16 percentage points, respectively) in favor of the first-listed choice; (iii) Both laboratory samples exhibit large and significant shifts in the cooperation rates across the two PD games, as predicted by the Rapoport ratio. For the online populations, only the CloudR sample exhibits a significant result in the predicted direction, where both the Prolific and MTurk samples are essentially inelastic in response to shifts in the PD-game tensions.

The core average behaviors in our experiments are illustrated in Figure 2. Panel A indicates the rate at which participants in each population make a mistake with respect to the offered incentives, choosing a Σ-dominated action in either game Σ-DOM1 and/or Σ-DOM2. Panel B indicates the treatment effect sizes, where the shaded regions indicate the difference in cooperation rates between games PD2 and PD1 (arrows indicate the direction of the difference, expected to be downward).[20]

Inspecting Figure2(A), the proportion of participants making a Σ-dominated choice is not statistically distinguishable between the Lab, VLab, CloudR or Prolific samples with approximately 12% of participants making a defect choices in the last two games. In

[20]For more detailed quantitative results, see Table A.1 in the Appendix.



(A) Σ-dominated individual choice (B) PD comparative statics

FIGURE 2. Summary results by population

Note: Panel A: Error-bars indicate binomial-exact 95-percent confidence intervals for the proportion of Σ-dominated play in the pooled population sample, where arrows indicate the change across the framing variable from the *C*-action presented first to the *D*-action presented first. *Panel B*: Shaded region in bars show the cooperation rate gap between the PD2 and PD1 games, with arrows indicating the change (red moving against prediction); provided *p*-values are participant-clustered tests for a difference in proportion against the one-sided alternative (more cooperation in game PD2).

contrast, for the MTurk sample the rate of Σ-dominated choices grows to more than one-in-three, significantly different from all the other population samples.[21]

Moreover, once we take into account the MTurk behavior in the reframed treatment, where the Σ-dominated *D*-action is listed first, the number of participants that are making orthogonal choices to the induced incentives increases still further. While the bar heights in Figure 2A indicate the pooled fraction of participants making a Σ-dominated choice, the arrows in the figure show the change in this proportion as we move from the environment where the *C*-action is listed first, to the alternate frame where the *D*-action is listed first. The largest shifts across the frame change here are in the VLab and MTurk samples, where listing the *D*-action first leads to a 19.2 and 16.3 percentage points increase, respectively, in the fraction of the Σ-dominated choices ($p = 0.037$ and $p < 0.001$, respectively, from two-sided tests). Despite successfully passing the screening questions—where participants must demonstrate their understanding of the game incentives, where those failing the check do not move forward—approximately one half of the MTurk sample are making choices that indicate little awareness of the induced incentives. While approximately a third of this inattention can be attributed to participants choosing the *D* action in games Σ-DOM1 and Σ-DOM2 simply because it is listed first (rather than random choice), the result of such a heuristic is similarly to wash out any treatment effect. As such, just under half of the sample are making choices that are orthogonal to

[21] $p < 0.001$ for all pairwise tests of proportion between MTurk and the other four populations.

the offered incentives (over half once we correct for some mis-classification). In contrast, despite similar costs per observation on CloudR and Prolific, the rate of such inattention is not significantly greater than that the laboratory sample.

We summarize these first two results:

Result 1 (Σ-Dominance)**.** *The MTurk sample exhibits significantly more inattention, measured through choices that cannot be rationalized by any preference that is increasing in both the individual and social payoff. The Lab, VLab, CloudR and Prolific samples are not significantly different based on this inattention measure.*

Result 2 (Response to frame)**.** *The VLab and MTurk samples exhibit significantly more choices that select the first-listed option, regardless of the incentives. While there is also a small effect for the Prolific sample, the effect is only marginally significant. We do not detect any re-framing effect in the CloudR and Lab samples.*

The focus of the above is on measuring the extent to which choices are being driven by inattention—either through mistakes, or to some feature that is orthogonal to the induced monetary incentives. We now examine the PD-game comparison, where we might expect level differences in the behavior across populations. However, our design accounts for such level differences by focusing on the *difference* in cooperation between the two PD games. The prediction from the Rapoport index, validated by the wider behavioral literature, is that a larger share of the participants will cooperate in PD2 than PD1. While the average cooperation levels for each population illustrated in Figure 2(B) can be different, our null is that the populations will have similar differences (the shaded gaps in the figures), with all five having more cooperation in game PD2 (arrows pointing downwards).

The first inference from Figure 2(B) is that the Lab and VLab samples yield identical treatment effects with a 12.2 percentage point cooperation rate difference in both samples.[22,23] In contrast, the MTurk and Prolific samples show substantially different cooperation gaps from the one we expect from the literature ($p < 0.001$ both comparisons), where neither gap moves significantly in the predicted direction. In fact, for MTurk, the data actually moves in the opposite direction (though insignificantly so if we had considered a two-sided test). Of the three online populations, only the CloudR sample has a significant effect in the predicted direction ($p < 0.001$). While the CloudR effect size of 7.4 percentage points is smaller than our literature expectation ($p < 0.001$), the increased power of the large sample leads to the greatest confidence across the five tests.

We summarize the findings across the two PD-games as follows:

[22]While there is not a significant difference in cooperation levels (neither by PD game, nor jointly), we do see that the virtual lab is more cooperative than the physical lab, which was in the opposite direction from our intuition.

[23]For both the Lab and VLab, while we do see a reduced effect size relative to the literature prediction of 17.2 percetage points, we cannot reject this in either treatment ($p = 0.426$ and $p = 0.404$ for the respective two-sided tests).

TABLE 2. Mixture model estimates

Population	Inattentive		Attentive
	First	Random	
Lab	0.000	0.144	0.856
VLab	0.000	0.216	0.784
MTurk	0.107	0.447	0.445
CloudR	0.000	0.160	0.840
Prolific	0.022	0.153	0.825

Result 3 (Behavior Comparison)**.** *The Lab, VLab and CloudR samples replicate the prior behavioral literature, with a significant cooperation drop between games PD2 and PD1, whereas this pattern is not found in either the Prolific or Mturk data.*

5. Inferential Preferences

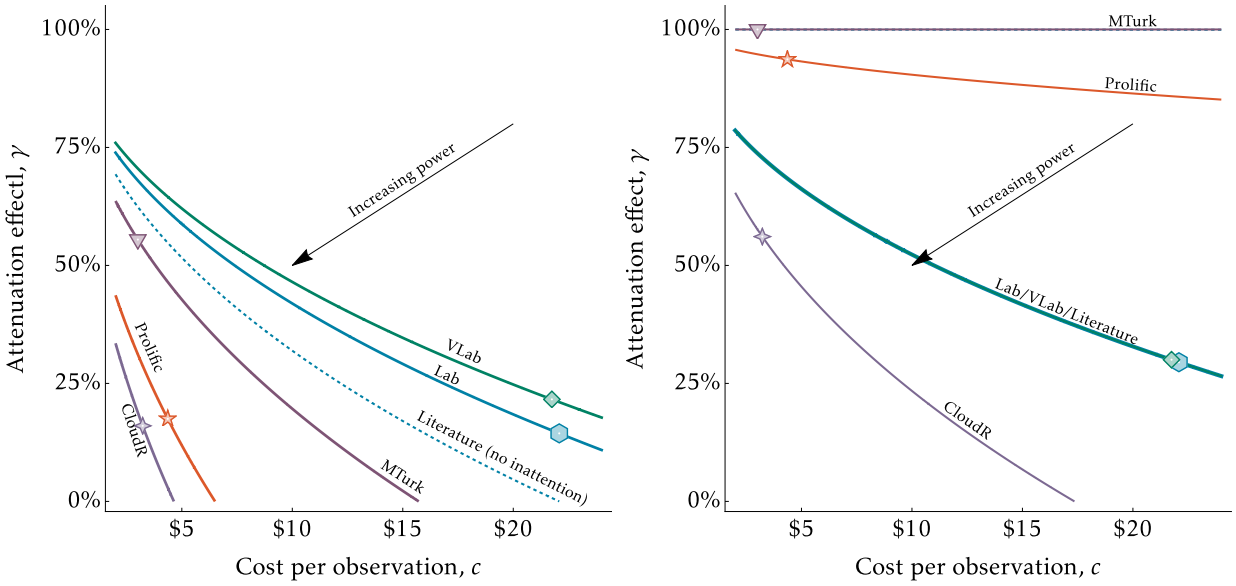
We now turn to our assessment of the difference across populations in terms of their inferential power. Using the results from our experiments, we can compare the five populations with respect to which contour they lie on within Figure 1. We first consider the inferential effects if the reduction in effect size is purely driven by an inattentive response, where we subsequently combine inattention with further reductions in the effect size due to inelastic responses.

5.1. **Inattention only.** For pure inattention we focus on the Σ-dominated behavior in games Σ-DOM1 and Σ-DOM2, when there is no real strategic tension. We suppose that there are three types of agents: (i) Inattentive types that chooses the first-listed action in both games, regardless of the offered incentives, with measure γ_F in the population. (ii) Inattentive types that choose their action at random, playing each choice with 50 percent probability, regardless of the incentives, with measure γ_R in the population. (iii) Attentive types that responds to the incentives and so satisfy Σ-dominance, with incidence $\gamma_\Sigma = 1 - \gamma_F - \gamma_R$.

Using the predicted behavior of the three types in games Σ-DOM1 and Σ-DOM2 (for each frame) we estimate a mixture model over the types proportions using maximum likelihood at the population level.[24] Estimated type proportions are given in Table 2 for each population.

The main result that jumps out of Table 2 is that for our MTurk sample, approximately 55 percent of the participants are inattentive, where just 45 percent of the respondents make choices driven by the offered economic incentives. In contrast, the estimated attentive proportion is 83 percent for Prolific, 84 percent for CloudR, 78 percent for the VLab and

[24]Random types choose each combination $(a_1, a_2) \in C, D^2$ for games Σ-DOM1 and Σ-DOM2 with equal probability in both frames. First-action types choose (C, C) in the standard frame and (D, D) in the reversed frame. Σ-dominant types choose (C, C) in both frames. As such, the mixture model will account for a one-in-four chance that random types are mis-classified as Σ-dominant in the raw proportions, and for the first-action type proportion in the standard frame.



(A) Noise only

(B) Noise and effect-size reduction

FIGURE 3. Population power

86 percent for the Lab.[25] While the proportion of inattentive MTurk decisions is certainly large—both through the first-listed action heuristic and through random play—each observation is very cheap. Given a cost of $22 for each Lab participant and $3 for each MTurk participant, we can collect seven MTurk observations for each lab observation. We now show that this disparity in cost should still break in favor of the MTurk sample, even when only a minority of the MTurk sample are attentive. So long as the attentive minority has a similar effect-size to the prior literature, then the MTurk sample would still be inferentially dominant when compared to the lab sample.

In Figure 3(A) we indicate each population's position as a (c, γ) bundle: for the Lab (hexagon), VLab (diamond), MTurk (triangle), CloudR (four-pointed star) and Prolific (five-pointed star). For the cost c, we simply take the sample's average observation cost; for the attenuation we set $\gamma = \gamma_F + \gamma_R$ from the mixture-model estimates. For a given population bundle (marked as a point), we can also draw out the set of power-equivalent bundles under a fixed budget of exactly $1,650 as a curve, with a lower curve indicating a more-powerful population. In this first panel, we assume that the expected effect sizes for the attentive sample are exactly as given from the prior literature, where the inattentive sample chooses randomly.

Using the illustrated results from the Panel A of Figure 3, we can see that if our populations varied solely over the observation costs and the rate of inattention, then both laboratory populations (Lab and VLab) would be inferentially inferior to MTurk. Despite

[25] While the aggregate results did indicate that the VLab sample had a greater rate of Σ-dominated actions in the D-first frame, our estimates from the mixture model instead attribute this to random type, as we do not see substantially more (D, D) choices in this treatment than would be predicted simply from the random types. Instead, the increase in the VLab sample is for (C, D) and (D, C) action choices, so that they are more likely to make *one* dominated choice.

a substantial inattentive proportion (55 percent) the much lower observation costs still leads to a lower chance of making a type II error on the qualitative hypothesis. However, MTurk is itself dominated. While Prolific and CloudResearch are both more expensive than MTurk per observation, the far lower rates of inattention on these populations mean the two more-curated online populations would have much greater inferential power.[26]

Figure 3(A) indicates that so long as there is no change in the expected effect size, at 19.5 percent attenuation due to inattention, Prolific would still be preferable to MTurk even if its cost per observation increased to $12.20. Alternatively, fixing the current MTurk cost, the inattention rate would have to shrink to 32.6 percent to match the Prolific sample's power. While our Prolific and CloudR samples have similar and statistically indistinguishable rates of inattention, CloudResearch emerges as the winner for this first analysis due to a smaller cost per observation at $3.23 (primarily driven by the additional fixed payment requirements on Prolific, which pushed the average observation costs to $4.36).

In these initial results, all three online samples dominate the laboratory samples. But here we are assuming that the effect sizes on the online platforms are similar to the prior literature which was primarily identified using lab studies. However, noise due to inattention is not the only factor to consider. Not only do we want a large fraction of attentive participants that are responsive to money as a reward medium, we also need the populations to have a suitably elastic responses.[27] The final power of each population, at least as it relates to our social-dilemma hypothesis, is the net of both the effect reductions due to inattention, but also any reduction in the effect size. We examine this compound effect in the second panel of Figure 3.

Similar to the first panel, each treatment is again depicted both as a point in cost-attenuation space, along with the iso-power curve for that point (assuming a fixed $1,650 budget). However, here we calculate the total attenuation, relative to the prior literature. The critical attenuation rate γ therefore scales down the literature effect size (here seventeen percentage points) to produce the same power as the actual samples, given the realized behavior in games PD1 and PD2 (the size of the illustrated gaps in Figure 2(B)).

This final analysis of the actual sample power substantially changes the ranking across our populations. First, MTurk has no power to speak to our directional hypothesis, as the realized comparative static has the opposite sign from the hypothesis, and so the attenuation is 100 percent. Prolific is next in our ranking, with only a small amount of power because the realized difference in cooperation between games PD1 and PD2 is just 1 percentage point. The Lab and VLab samples have essentially the same power level

[26] In a robustness check, CloudResearch provided us with worker IDs for the subset of our MTurk sample that were also within the CloudResearch Approved List. Of the 548 participants in our MTurk sample, 162 were in the Approved List (30 percent selection). Rerunning our mixture model on these 162 participants we find that the Approved-List sub-sample has $\hat{\gamma}_F = 0.000$, $\hat{\gamma}_R = 0.189$ and $\hat{\gamma}_\Sigma = 0.811$, which is very similar to the direct CloudR sample.

[27] For example, see Araujo et al. (2016) who demonstrate that while a slider-based real-effort task does show a qualitative response to incentives, the effect sizes are economically very small. DellaVigna and Pope (2017) also find a much smaller change in effort in response to a large change in incentives.

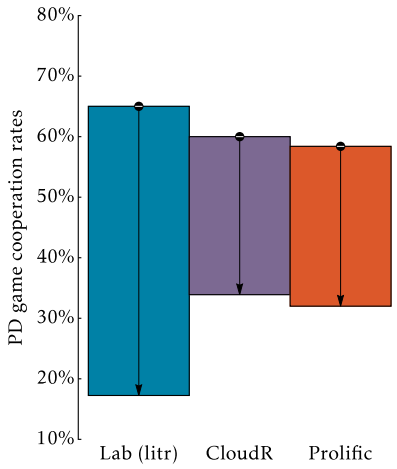


FIGURE 4. Effect Size (extended sample)

Note: Shaded region in bars show the cooperation rate gap between PD games with Rapoport ratio of $\rho = 0.71$ and $\rho = 0.05$, with arrows indicating the change; provided p-values are participant-clustered tests for a difference in proportion against the one-sided alternative (more cooperation with higher Rapoport ratio).

as one another (the Lab and VLab curves are essentially coincident in the figure).[28] In terms of the Rapoport comparative static across the two PD games, the CloudR sample comes out ahead. While there is a slightly reduced effect size (relative to both the assessed lab effects and the literature prediction), this is more than compensated for by the much larger sample size.

5.2. **Response to Increased Tensions.** Given the observed inelastic response in two out of our three online samples, one possible conclusion is that online populations do not respond to social-dilemma tensions in the same way as laboratory participants. That is, ignoring the more positive results from our CloudR sample, maybe the Rapoport ratio effect is a lab-specific phenomenon. To examine this, we ran a robustness study on both the CloudResearch approved list and Prolific. Here we recruited a further set of participants with a budget of approximately $500, and added two further PD games to the studies. These additional games ramp up the PD tensions further still, with Rapoport ratios are 0.05 and 0.25 (for the precise games see Appendix B).

Looking to the Charness et al. (2016) paper to calibrate the expected effect size from the laboratory-identified literature, we estimate that a comparison of the most-extreme PD game pair (Rapport ratios of 0.71 and 0.05) in a lab sample would show a cooperation reduction of 48 percentage points (moving from 65 percent cooperation to 17 percent). This expected effect size is shown in the first bar in Figure 4. In this way, we can stress test the idea that the online samples are showing reduced elasticity by ramping up the input tensions.

[28]Because the realized difference between the PD1 and PD2 cooperate games are the same, but the VLab sample has higher inattention, the model here assumes the attentive sample effect size is *larger* in the VLab sample. However, this nets out to the same effective power. With participant clustering, the VLab sample actually performs better inferentially.

The difference in cooperation across two most extreme PD games for the robustness sample data are illustrated in Figure 4. While CloudR continues to show a high degree of responsiveness to the increased tensions (a 26.1 percentage point effect), we also find a more substantial effect for Prolific. In the lowest Rapoport ratio PD games the Prolific cooperation rate falls to 32.0 percent, with a comparable cooperation rate to the original sample at 58.4 in the highest Rapoport-ratio PD game (see Appendix B for full analysis).[29] While the difference in cooperation for Prolific is now highly significant across the most-extreme PD games ($p < 0.001$) and similar to the effect for CloudR (also $p < 0.001$), the 26 percentage-point reduction still represents approximately half the effect size we would expect from the laboratory literature.[30]

Our results on the extended games suggest that online populations *are* capable of uncovering the same qualitative patterns as the laboratory. However, two caveats are appropriate here: First, the substantial noise on MTurk suggests that the more-curated online populations are superior.[31] Second, the elasticity of response to other-regarding tensions in our online populations is substantially reduced relative to lab studies (and our own lab samples). Under more-nuanced parameterizations, an online populations' response may be too small to be well-powered. Fixing an online sample, if the aim is purely to uncover a qualitative finding, the conclusion from our extension is to eschew all subtlety. So long as the parameterization can generate a moderate effect size, the smaller cost per observation for online populations can begin to show final benefits on inference.

On the flip-side of the coin, our study also points to the continuing utility of laboratory samples—particularly in an age where sophisticated AIs are becoming more ubiquitous. Lab participants, whether physically present for the experiment or participating online, show consistent, replicable responses, and where the experimenter has the benefit of being able to see the participant and verify their identity. In studies where the aim is to educe more nuanced findings—calibrating a non-linear model say, where estimating curvature requires smaller step-sizes in the treatment—then the lab can play a more useful role. Despite the increased expense per observation, the combination of a more-elastic response to the incentives and a low rate of inattention make standard lab samples a still effective tool for researchers. While our study has no variation in the level of complexity, the lab offers a conducive environment for testing knottier economic hypotheses, if greater explanation/instruction is required to induce the economic environment. By controlling participants' outside-option activities and removing distraction, lab samples allow experimenters to test more-complex theories. While there is certainly a place for online samples, given their low cost and ease of acquisition, a lack of control and inelastic response does seem to be a problem for some online populations in use by economic researchers.

[29] We cannot reject differences with the original cooperation levels in the common games PD1 and PD2, despite the increase to 6 choices.

[30] For details, see Appendix B and C

[31] Alternatively, that greater internal validity checks are required, enabling analysis on a population subsample but substantially increasing the effective cost of each usable MTurk observation

6. Conclusions

We examine five populations commonly used by economists for conducting incentivized experiments. Rather than a pure validation of the comparative statics across the differing populations, we take a different tack. Using the idea that academic research faces the same budget discipline we assume in other setting, we focus on the ecologically valid differences in cost and quality of each observation across these populations. Assuming the experimenter preferences over cost and quality of each observation are increasing in inferential power, we can compare these five populations with an intuitive and relevant yardstick. To that end, we examine the precise substitutions a researcher might want to make by trading off some noise in the data for much cheaper observations, which enable larger samples for the same fixed research budget.

Our design first measures the inattention within each population, via a weak assumption on the response to the reward medium (here considering both individual and social welfare). Inattentive choices will have an effect on inference by washing out treatment effects, if participants do not consider the induced incentives. But we also measure a more-nuanced response to a behavioral theory motivated by the prior literature over the cooperation in social dilemmas. Fixing the experimental budgets on each population and varying the scale of the incentives so that they match standards for each population, we create ecologically valid differences in the sample sizes. Using this data, we then assess the inattention and behavioral response on each population sample, measuring the extent to which each population replicates the behavioral comparative-static result from the literature.

In terms of the proportion of inattentive participants, the laboratory sample has the lowest levels of inattention, though both CloudResearch and Prolific also have relatively low levels (which are statistically indistinguishable from the lab levels). In contrast, at 55 percent, our MTurk sample is particularly inattentive, despite standard screens in place to ensure understanding. However, even at this level of inattention, the very cheap observations from MTurk should still dominate the laboratory from an inferential point of view if the the attentive subsample has a similar treatment effect. But, in terms of the pure inattention and cost, the CloudResearch and Prolific samples should themselves dominate MTurk.

The substantial inattention in the MTurk sample may be a recent phenomenon (where recent studies suggest the population decline happened just before the pandemic). However, our analysis suggests that despite being the cheapest of the samples, MTurk might offer a false economy. While slightly more expensive per observation, both Prolific and CloudResearch offer substantially greater attention from the participants inferential power by reducing noise.

Beyond pure inattention, as we move to our actual behavioral comparative static, the results are not as clearly in favor of the online populations. Our findings on the social-dilemma test indicate that even though the sample size for our lab and the virtual-lab samples is small (due to relatively expensive observation costs) both replicate the standard literature finding. However, two of the online samples, with much larger samples—MTurk and Prolific—fail to recover the expected qualitative result. Overall, our CloudResearch

sample wins on the inferential-power horse-race, where we find that it replicates the standard qualitative finding (albeit with a smaller effect size), but where the increased sample size from a low cost per observation leads to greater power than the lab.

A reason for the failure in two of our online samples, is that the lab samples (and the literature results based on lab samples) exhibit a much-greater elasticity of response to treatment. In contrast, the MTurk and Prolific samples are essentially inelastic for the social dilemmas we induce. Only one out of three online samples recovers a significant effect. While the small elasticities of response could be specific to social dilemmas—where our more-generalizable attention estimates clearly outline the power of Prolific and CloudResearch samples—they also suggest some continuing usefulness for lab studies. Despite greater expense per observation, lab samples can offer greater power, and replicability.

References

Araujo, Felipe A, Erin Carbone, Lynn Conell-Price, Marli W Dunietz, Ania Jaroszewicz, Rachel Landsman, Diego Lamé, Lise Vesterlund, Stephanie W Wang, and Alistair J Wilson, "The slider task: an example of restricted inference on incentive effects," *Journal of the Economic Science Association*, 2016, *2* (1), 1–12.

Arechar, Antonio A, Simon Gächter, and Lucas Molleman, "Conducting interactive experiments online," *Experimental Economics*, 2018, *21* (1), 99–131.

Aruguete, Mara S, Ho Huynh, Blaine L Browne, Bethany Jurs, Emilia Flint, and Lynn E McCutcheon, "How serious is the 'carelessness' problem on Mechanical Turk?," *International Journal of Social Research Methodology*, 2019, *22* (5), 441–449.

Charness, Gary, Luca Rigotti, and Aldo Rustichini, "Social surplus determines cooperation rates in the one-shot Prisoner's Dilemma," *Games and Economic Behavior*, 2016, *100*, 113–124.

Chen, Daniel L, Martin Schonger, and Chris Wickens, "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 2016, *9*, 88–97.

Chmielewski, Michael and Sarah C Kucker, "An MTurk crisis? Shifts in data quality and the impact on study results," *Social Psychological and Personality Science*, 2020, *11* (4), 464–473.

Danz, David, Neeraja Gupta, Marissa Lepper, Lise Vesterlund, and K Pun Winichakul, "Going virtual: A step-by-step guide to taking the in-person experimental lab online," *Available at SSRN 3931028*, 2021.

DellaVigna, Stefano and Devin Pope, "What Motivates Effort? Evidence and Expert Forecasts," *The Review of Economic Studies*, 06 2017, *85* (2), 1029–1069.

Eyal, Peer, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina, "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, 2021, pp. 1–20.

Fréchette, Guillaume R, Kim Sarnoff, and Leeat Yariv, "Experimental economics: Past and future," *Annual Review of Economics*, 2022, *14*, 777–794.

Goodman, Joseph K, Cynthia E Cryder, and Amar Cheema, "Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples," *Journal of Behavioral Decision Making*, 2013, *26* (3), 213–224.

Hauser, David, Gabriele Paolacci, and Jesse Chandler, "Common concerns with MTurk as a participant pool: Evidence and solutions," in "Handbook of research methods in consumer psychology," Routledge, 2019, pp. 319–337.

Horton, John J, David G Rand, and Richard J Zeckhauser, "The online laboratory: Conducting experiments in a real labor market," *Experimental Economics*, 2011, *14* (3), 399–425.

Litman, Leib, Jonathan Robinson, and Tzvi Abberbock, "TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences," *Behavior research methods*, 2017, *49* (2), 433–442.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision making*, 2010, *5* (5), 411–419.

Rapoport, Anatol, "A note on the" index of cooperation" for prisoner's dilemma," *Journal of Conflict Resolution*, 1967, *11* (1), 100–103.

Snowberg, Erik and Leeat Yariv, "Testing the waters: Behavior across participant pools," *American Economic Review*, 2021, *111* (2), 687–719.

Thomas, Kyle A and Scott Clifford, "Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments," *Computers in Human Behavior*, 2017, *77*, 184–197.

Appendix A. Results Formerly in Main Text

Table A.1. Results Summary

	Lab	**VLab**	**MTurk**	**CloudR**	**Prolific**
Panel A:	Avg	Avg	Avg	Avg	Avg
Σ-Dominated:	0.108 (0.036)	0.162 (0.043)	0.369 (0.021)	0.120 (0.014)	0.122 (0.017)
***i*-Dominant** (*DD-CC*)**:**	0.324 (0.054)	0.270 (0.052)	0.159 (0.016)	0.240 (0.018)	0.260 (0.022)
Rapoport identifier (*DC-CC*)**:**	0.189 (0.046)	0.176 (0.045)	0.093 (0.012)	0.129 (0.014)	0.106 (0.016)
Full Cooperator (*CC-CC*)**:**	0.284 (0.052)	0.338 (0.055)	0.297 (0.020)	0.447 (0.021)	0.416 (0.025)
Σ-Dominant:	0.892 (0.036)	0.838 (0.043)	0.631 (0.021)	0.880 (0.014)	0.878 (0.017)
Rapoport ordered:	0.905 (0.034)	0.919 (0.032)	0.828 (0.016)	0.917 (0.012)	0.886 (0.016)
Both:	0.797 (0.047)	0.784 (0.048)	0.549 (0.021)	0.817 (0.017)	0.782 (0.021)
Panel B:	Δ_{Frame}	Δ_{Frame}	Δ_{Frame}	Δ_{Frame}	Δ_{Frame}
Σ-Dominated:	−0.037 (0.073)	0.192 (0.090)	0.163 (0.044)	0.008 (0.030)	0.052 (0.037)
***i*-Dominant** (*DD-CC*)**:**	0.075 (0.118)	−0.092 (0.111)	0.028 (0.034)	0.042 (0.040)	0.056 (0.048)
Rapoport identifier (*DC-CC*)**:**	−0.033 (0.095)	0.048 (0.096)	−0.048 (0.024)	0.012 (0.031)	0.056 (0.047)
Full Cooperator (*CC-CC*)**:**	0.012 (0.112)	−0.192 (0.117)	−0.046 (0.041)	−0.023 (0.046)	−0.070 (0.052)
Σ-Dominant:	0.037 (0.073)	−0.192 (0.090)	−0.163 (0.044)	−0.008 (0.030)	−0.052 (0.037)
Rapoport ordered:	0.017 (0.071)	−0.065 (0.068)	0.049 (0.032)	0.025 (0.026)	−0.018 (0.035)
Both:	0.053 (0.096)	−0.235 (0.100)	−0.066 (0.045)	0.031 (0.036)	−0.040 (0.045)

Note: Standard errors for proportions in parentheses.

Table A.1 provides average outcomes across the five samples with standard errors derived from simple tests of proportion. In Panel A we first outline the proportion of individuals with particular focal behaviors over the four games (pooling data across the frame), then outline the relative effects across the re-framing in Panel B.

The first row in Panel A of Table A.1 shows the rate at which individuals in the experiment make an obvious mistake with respect to the offered incentives i.e., choose the Σ-dominated actions. The proportion of participants choosing the σ-dominated actions is statistically inseparable between the Lab, VLab, CloudR, and Prolific samples with

approximately 12% of participants make a defect choices in the last two games.[32] In contrast, for the MTurk sample this rate grows to more than one-in-three, significantly different from all other samples.[33] Moreover, as we explain next, even this number is perhaps an underestimate of the fraction of participants making choices orthogonal to the incentives.

Where panel A in Table A.1 provides the overall average results by population sample (pooling across both the C-first and D-first frames), Panel B indicates the change in the proportion across the re-frame. The first row of Panel B shows the change in the participant proportion exhibiting a Σ-dominated choice when we move from listing C to listing D as the first action. Our results across the re-frame show that the Lab sample moves in the opposite direction from a first-option bias with a slight decrease in Σ-dominance when the D action is listed first (though this is not significant, $p = 0.640$). The first-option bias is the smallest for the CloudR sample (0.8 percentage points with $p = 7894$). The Prolific sample does show a movement 5.2 percentage point movement, where 15.6 percent of choices in the D-first sample are Σ-dominated choices. Though this difference is not significant ($p = 0.160$) but if we allowed for a one-sided test there is marginal evidence for a small first-action bias on Prolific. The largest effects though are in the VLab and MTurk sample,s where listing the D-action first leads to a 19.2 and 16.3 percentage points increase in the Σ-dominated fraction respectively ($p = 0.37$ and $p < 0.001$ on a test of proportions respectively for VLab and MTurk).[34]

In the worst-case D-first treatment 47.8 percent of the MTurk choices are Σ-dominated. Despite successfully passing the screen questions—where participants must demonstrate their understanding of the game incentives or be kicked out—approximately one half of the MTurk sample then make choices that indicate little awareness of the induced games. While approximately a third of this effect can be attributed to participants choosing the D action in games Σ-DOM1 and Σ-DOM2 simply because it is the first-listed option, the result still indicates that just under half of the sample are making choices that are orthogonal to the offered incentives. In contrast, despite similar costs per observation on CloudR and Prolific, the rates of such mistakes in these populations seems to be at most 15 percent, and we lack statistical power to say that it is even different from the laboratory.

[32]The pairwise p-values for the test of proportions for {Lab vs. VLab, Lab vs. CloudR, Lab vs. Prolific, VLab vs. CloudR, VLab vs. Prolific, CloudR vs. Prolific} are {0.3395, 0.7644, 0.735, 0.3065, 0.3465, 0.9294} respectively.

[33]$p < 0.001$ for the pair-wise tests of proportions between MTurk and all four populations

[34]The bottom section of Panel B in Table A.1 indicates that the re-framing has a consistent effect in increasing the selection of D in the VLab and MTurk samples when this action is listed first.

Appendix B. CloudResearch Approved List (CloudR) Robustness Sessions: Extended Response

Table C1. Experimental Games: Robustness Sample

PD1 game ($\rho = 0.50$):

	C	D
C	21,21	2,28
D	28,2	8,8

PD2 game ($\rho = 0.71$):

	C	D
C	19,19	8,22
D	22,8	9,9

PD3 game ($\rho = 0.05$):

	C	D
C	14,14	5,25
D	25,5	13,13

PD4 game ($\rho = 0.25$):

	C	D
C	18,18	3,27
D	27,3	12,12

Σ-DOM1 game:

	C	D
C	17,17	12,16
D	16,12	10,10

Σ-DOM2 game

	C	D
C	15,15	16,10
D	10,16	11,11

Table C2. CloudR Participants per treatment

	CloudR	CloudR-Robustness
Main	374	165
Re-frame	167	

Note: Excludes participants who did not answer the comprehension question correctly.

Table C3. Behavior Across CloudR Samples: Cooperation

Game	Rapoport	Robustness Sample	p-value	Original Sample
PD1	0.50	0.485 (0.039)	0.107	0.556 (0.021)
PD2	0.71	0.600 (0.038)	0.483	0.630 (0.021)
Σ-DOM1		0.964 (0.015)	0.174	0.935 (0.011)
Σ-DOM2		0.939 (0.019)	0.560	0.926 (0.011)
PD3	0.05	0.339 (0.037)		
PD4	0.25	0.406 (0.038)		

Note: Standard error for the proportion in parentheses. All p-values are for two-sided tests of equality between the samples.

TABLE C4. Subject Types Across CloudR Samples: Pooled Data

Type	Robustness Sample	p-value	Original Sample
Choice Profiles in Original 4 Games:			
Nash (DD-CC)	0.297 (0.036)	0.143	0.24 (0.018)
Uncond Coop (CC-CC)	0.400 (0.038)	0.284	0.447 (0.021)
Cond Coop (DC-CC & CD-CC)	0.230 (0.033)	0.286	0.192 (0.017)
Σ-dominated	0.073 (0.02)	0.087	0.12 (0.014)

Note: Standard error for the proportion in parentheses. All p-values are for two-sided tests of equality between the populations. Choice profiles are given in order of the Rapoport ratio in the PD games (so PD1, PD2-Σ-DOM1, Σ-DOM2 in this table).

TABLE C5. Additional Subject Types in CloudR Robustness Sample

Type	Robustness Sample
Σ-dominant	0.927 (0.020)
Rapoport ordered	0.812 (0.031)
Both	0.770 (0.033)
Σ-dominant profiles:	
Nash, $DDDD$-CC	0.261 (0.034)
$DDDC$-CC	0.121 (0.025)
$DDCC$-CC	0.073 (0.020)
$DCCC$-CC	0.061 (0.019)
Uncond Coop, $CCCC$-CC	0.255 (0.034)
Non-Rapoport ordered (11 profiles)	0.158 (0.028)

Note: Standard error for the proportion in parentheses. Choice profiles are given in order of the Rapoport ratio in the PD games (so PD3, PD4, PD1, PD2-Σ-DOM1, Σ-DOM2 in this table).

Appendix C. Prolific Robustness Sessions: Extended Response

Table D1. Prolific Participants per treatment

	Prolific	Prolific-Robustness
Main	250	125
Re-frame	135	

Note: Excludes participants who did not answer the comprehension question correctly.

Table D2. Behavior Across Prolific Samples: Cooperation

Game	Rapoport	Robustness Sample	p-value	Original Sample
PD1	0.50	0.488 (0.045)	0.127	0.566 (0.025)
PD2	0.71	0.584 (0.044)	0.885	0.577 (0.025)
Σ-DOM1		0.904 (0.026)	0.865	0.909 (0.015)
Σ-DOM2		0.952 (0.019)	0.623	0.94 (0.012)
PD3	0.05	0.320 (0.042)		
PD4	0.25	0.328 (0.042)		

Note: Standard error for the proportion in parentheses. All p-values are for two-sided tests of equality between the samples.

Table D3. Subject Types Across Prolific Samples: Pooled Data

Type	Robustness Sample	p-value	Original Sample
Choice Profiles in Original 4 Games:			
Nash (DD-CC)	0.312 (0.042)	0.255	0.260 (0.022)
Uncond Coop (CC-CC)	0.352 (0.043)	0.208	0.416 (0.025)
Cond Coop (DC-CC & CD-CC)	0.208 (0.036)	0.897	0.203 (0.021)
Σ-dominated	0.128 (0.030)	0.862	0.122 (0.017)

Note: Standard error for the proportion in parentheses. All p-values are for two-sided tests of equality between the populations. Choice profiles are given in order of the Rapoport ratio in the PD games (so PD1, PD2-Σ-DOM1, Σ-DOM2 in this table).

TABLE D4. Additional Subject Types in Prolific Robustness Sample

Type	Robustness Sample
Σ-dominant	0.872 (0.030)
Rapoport ordered	0.808 (0.035)
Both	0.728 (0.040)
Σ-dominant profiles:	
Nash, *DDDD-CC*	0.272 (0.040)
DDDC-CC	0.120 (0.029)
DDCC-CC	0.088 (0.025)
DCCC-CC	0.072 (0.023)
Uncond Coop, *CCCC-CC*	0.176 (0.034)
Non-Rapoport ordered (11 profiles)	0.144 (0.032)

Note: Standard error for the proportion in parentheses. Choice profiles ar given in order of the Rapoport ratio in the PD games (so PD3, PD4, PD1, PD2-Σ-DOM1, Σ-DOM2 in this table).

Appendix D. Instructions for Laboratory Experiment

D.1. **Instructions for Main Treatment.**

Welcome and thank you for participating in this study. This is an experiment on decision making. Please turn off your cell phones and similar devices now and place them on the top shelf of your station. Please do not talk to or in any way try to communicate with other participants in the room. Your earnings in today's experiment will depend on your decisions, the decisions of others in the room, and on chance. Any money you make will be paid privately and in cash at the end of the experiment. We will start with a brief description of your task today. If you have any questions, please raise your hand and we will come to answer you in private.

Explanation of your task

There are four rounds in today's study, each consisting of a decision table. Your task will be to choose one option from two alternatives for each decision table. A round will end when all participants submit their choices.

At the end of the fourth round, the computer will randomly and anonymously pair you with another participant in the room. Next, the computer will randomly select one of your four rounds. You will be paid for that round based on you and the matched participant's choices in that round. Your final earnings will then consist of payoff from this one round and a participation fee of $6.

Every round is equally likely to be selected for payment, so you should treat each round as if it determines your final payment. Also, there are only four decisions in this study, so you should consider them carefully.

Description of a Decision Table

Below is an example decision table: Both you and the matched participant make choices

Your Decision	Other's Decision	Your Payoff	Other's Payoff
A	A	$18	$18
A	B	$6	$15
B	A	$15	$6
B	B	$10	$10

between Option A and Option B. The decision table indicates the payout for you and the other participant for each possible combination of choices.

Suppose this decision table was selected for payment, then in addition to the participation fee:

(1) if both participants choose A, they each receive $18;

(2) if you choose A and the matched participant chooses B, then you receive $6, and they receive $15;

(3) Vice versa if you choose B and the matched participant chooses A, then you receive $15, and they receive $6.

(4) if both participants choose B, they each receive $10;

We will begin the study with a few questions about your understanding of the decision table and then proceed to the first round.

D.2. **Instructions for Re-framed Treatment.** [Introductory instructions and section with "Explanation of your task" were identical to D.1]

Description of a Decision Table
Below is an example decision table: Both you and the matched participant make choices

Your Decision	Other's Decision	Your Payoff	Other's Payoff
A	A	$10	$10
A	B	$15	$6
B	A	$6	$15
B	B	$18	$18

between Option A and Option B. The decision table indicates the payout for you and the other participant for each possible combination of choices.

Suppose this decision table was selected for payment, then in addition to the participation fee:

(1) if both participants choose A, they each receive $10;

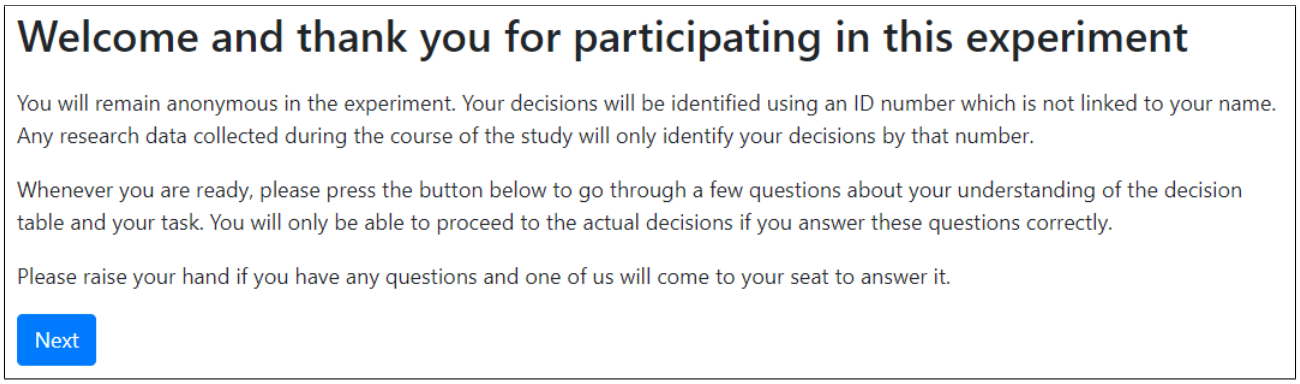
(2) if you choose A and the matched participant chooses B, then you receive $15, and they receive $6;

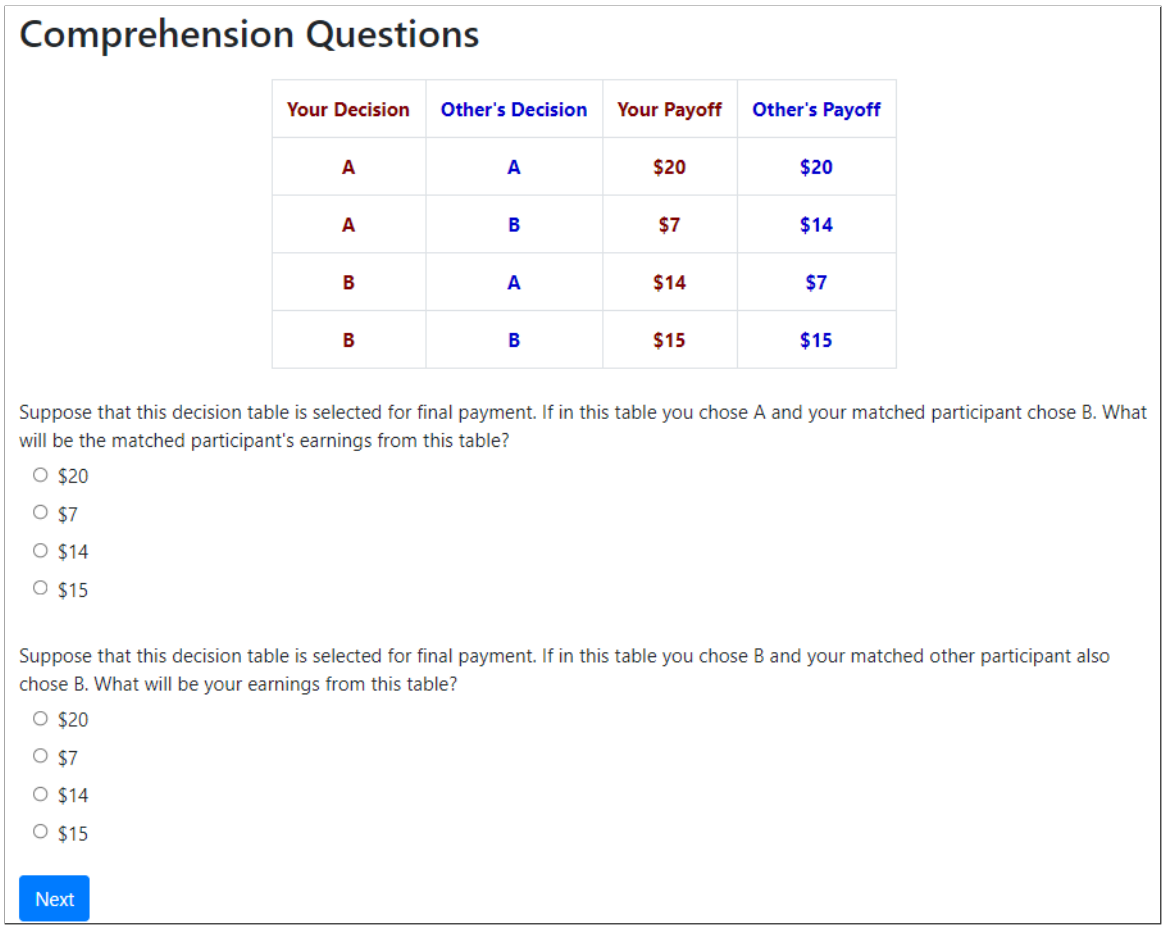
(3) Vice versa if you choose B and the matched participant chooses A, then you receive $6, and they receive $15.

(4) if both participants choose B, they each receive $18;

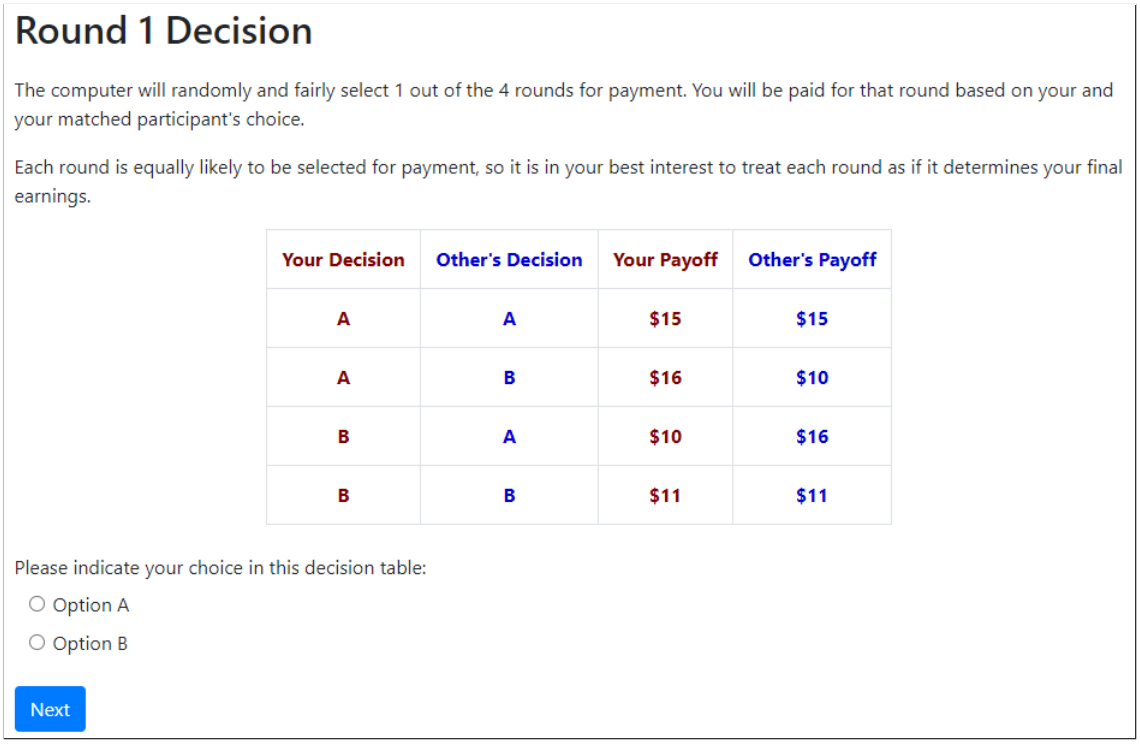
We will begin the study with a few questions about your understanding of the decision table and then proceed to the first round.

D.3. **Screenshots of the Laboratory Experiment.** Following are the screenshots of the lab experiment for the main sample. The screens for the re-framed sample were identical except that the labels of options on the decision table reversed.

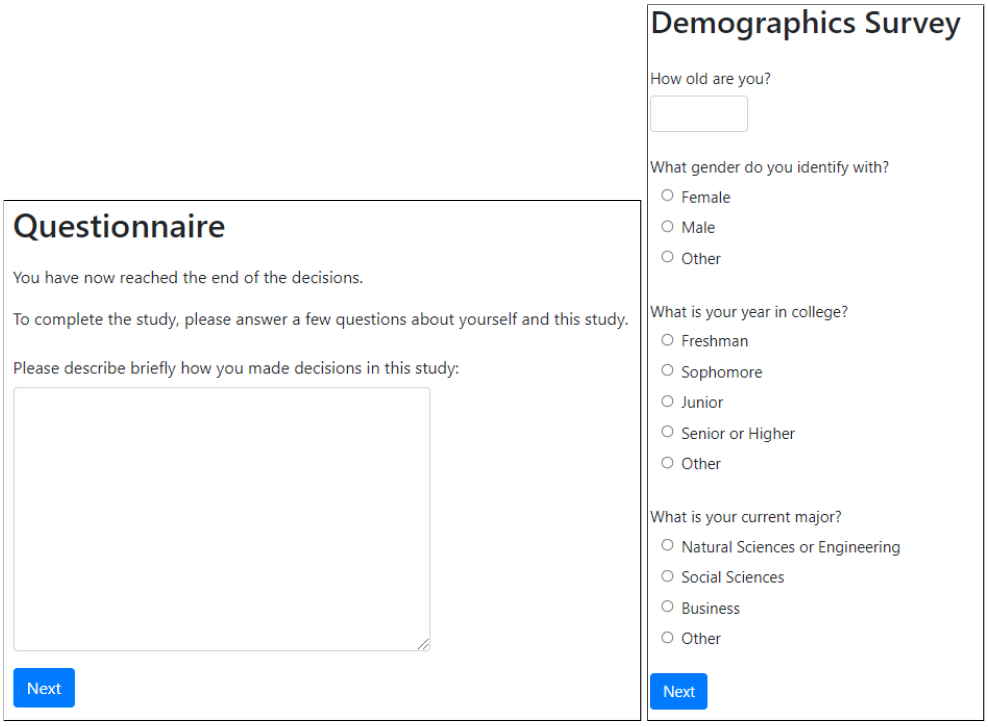


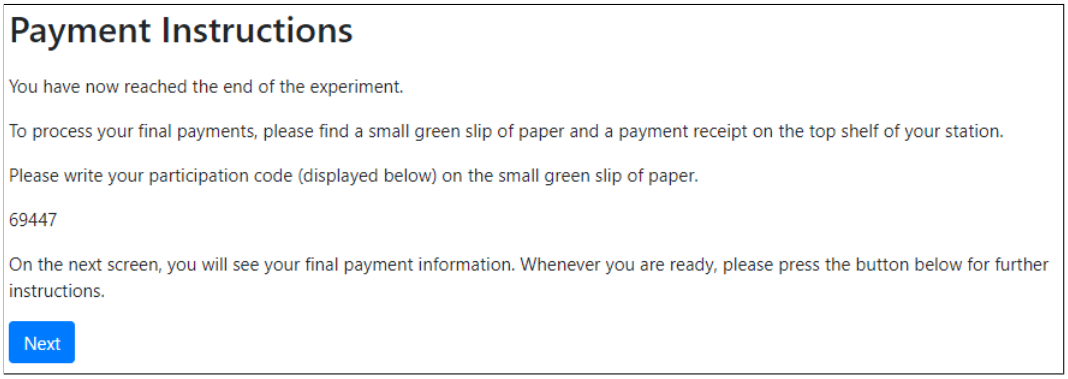


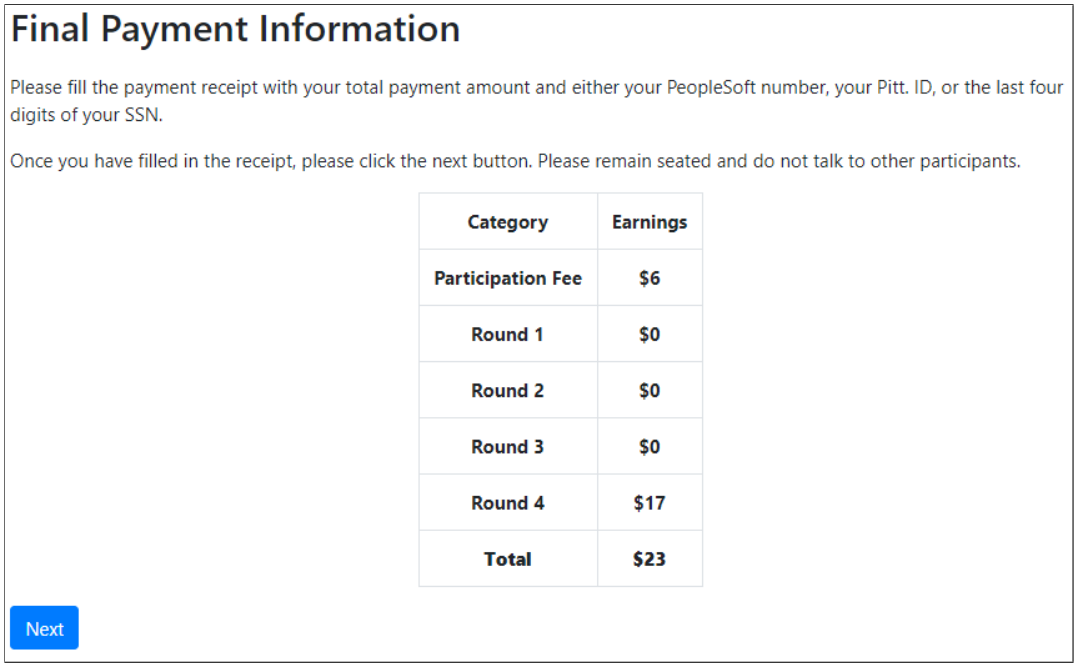
[For the re-framed sample option *A* corresponded to *D* and option *B* corresponded to *C*. The answers to the comprehension questions changed accordingly. Participants couldn't move forward without answering these questions correctly.]



[Rounds 2, 3 and 4 screens were the same as round 1 with different decision tables. For the re-framed sample option *A* corresponded to *D* and option *B* corresponded to *C*, the screens were otherwise the same as the main sample. The four decision tables were presented to the participants in random order.]



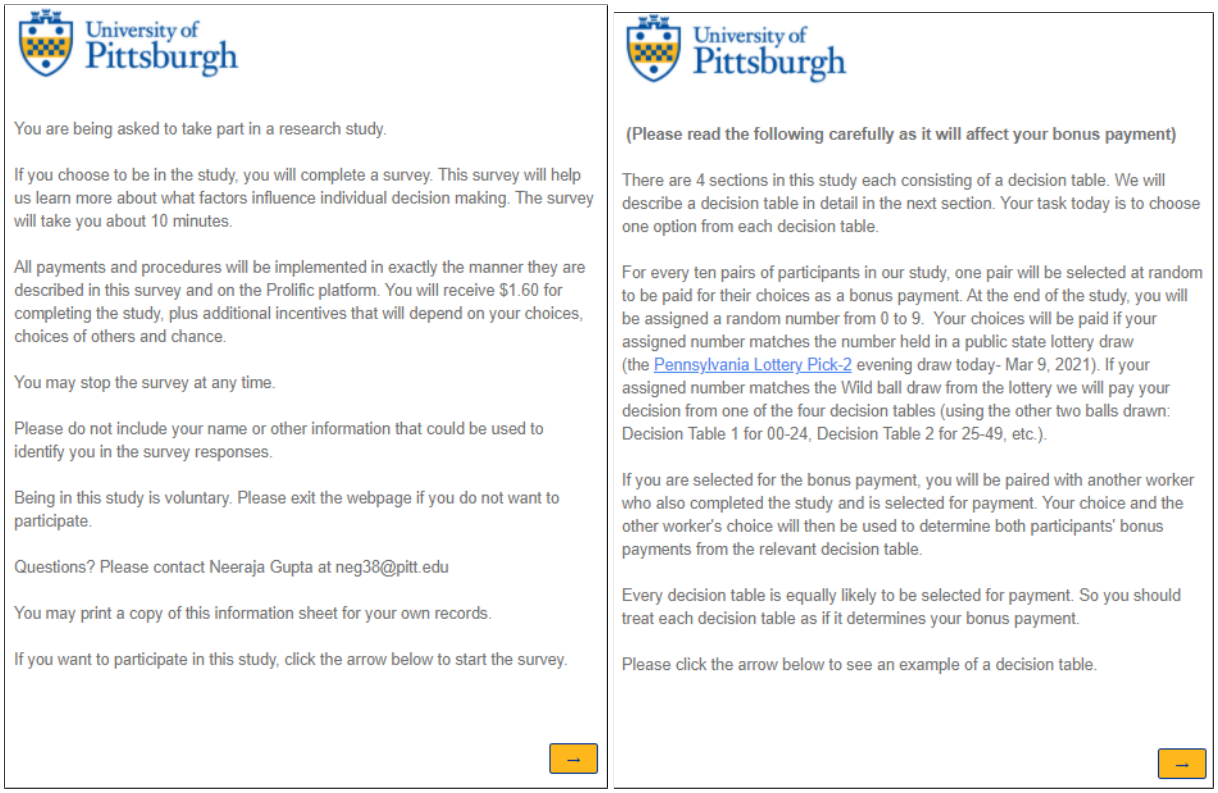




[Participants were then invited to the payment room one by one and paid in cash in private.]

Appendix E. Instructions for Online Experiment

Following are the screenshots of the online experiment for the main Prolific sample. The screens for the MechTurk sample were the same as the Prolific sample.



University of Pittsburgh

You are being asked to take part in a research study.

If you choose to be in the study, you will complete a survey. This survey will help us learn more about what factors influence individual decision making. The survey will take you about 10 minutes.

All payments and procedures will be implemented in exactly the manner they are described in this survey and on the Prolific platform. You will receive $1.60 for completing the study, plus additional incentives that will depend on your choices, choices of others and chance.

You may stop the survey at any time.

Please do not include your name or other information that could be used to identify you in the survey responses.

Being in this study is voluntary. Please exit the webpage if you do not want to participate.

Questions? Please contact Neeraja Gupta at neg38@pitt.edu

You may print a copy of this information sheet for your own records.

If you want to participate in this study, click the arrow below to start the survey.

→

University of Pittsburgh

(Please read the following carefully as it will affect your bonus payment)

There are 4 sections in this study each consisting of a decision table. We will describe a decision table in detail in the next section. Your task today is to choose one option from each decision table.

For every ten pairs of participants in our study, one pair will be selected at random to be paid for their choices as a bonus payment. At the end of the study, you will be assigned a random number from 0 to 9. Your choices will be paid if your assigned number matches the number held in a public state lottery draw (the Pennsylvania Lottery Pick-2 evening draw today- Mar 9, 2021). If your assigned number matches the Wild ball draw from the lottery we will pay your decision from one of the four decision tables (using the other two balls drawn: Decision Table 1 for 00-24, Decision Table 2 for 25-49, etc.).

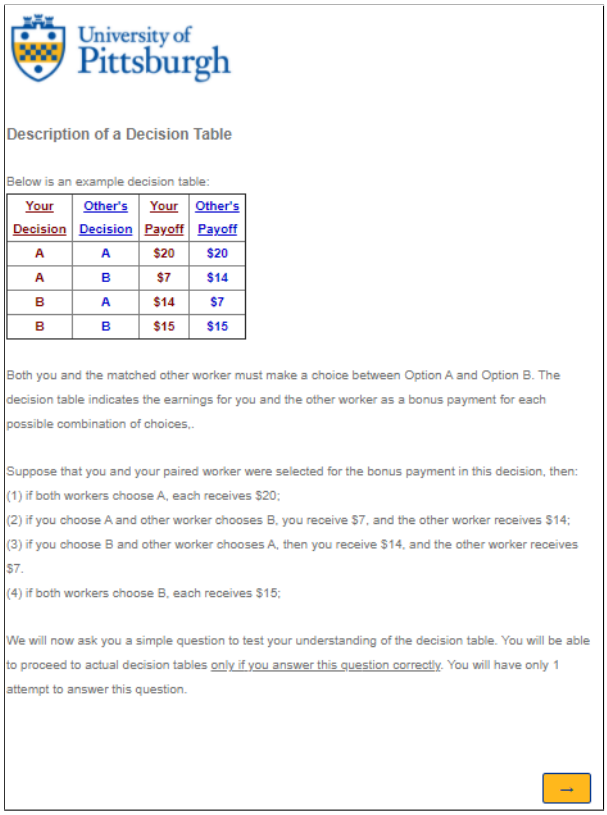
If you are selected for the bonus payment, you will be paired with another worker who also completed the study and is selected for payment. Your choice and the other worker's choice will then be used to determine both participants' bonus payments from the relevant decision table.

Every decision table is equally likely to be selected for payment. So you should treat each decision table as if it determines your bonus payment.

Please click the arrow below to see an example of a decision table.

→

[For the re-framed sample option *A* corresponded to *D* and option *B* corresponded to *C*. The answers to the comprehension questions changed accordingly. Participants were dismissed with the show-up of $1.60 for answering the comprehension question incorrectly on Prolific ($0.50 on MechTurk). On MechTurk, participants who answered the comprehension question correctly were offered additional $0.50.]


University of Pittsburgh

Description of a Decision Table

Below is an example decision table:

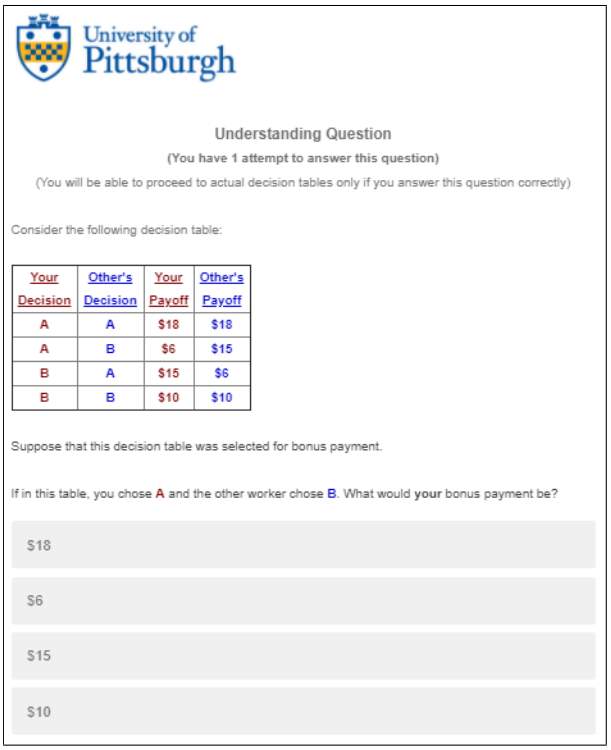
Your Decision	Other's Decision	Your Payoff	Other's Payoff
A	A	$20	$20
A	B	$7	$14
B	A	$14	$7
B	B	$15	$15

Both you and the matched other worker must make a choice between Option A and Option B. The decision table indicates the earnings for you and the other worker as a bonus payment for each possible combination of choices,.

Suppose that you and your paired worker were selected for the bonus payment in this decision, then:
(1) if both workers choose A, each receives $20;
(2) if you choose A and other worker chooses B, you receive $7, and the other worker receives $14;
(3) if you choose B and other worker chooses A, then you receive $14, and the other worker receives $7.
(4) if both workers choose B, each receives $15;

We will now ask you a simple question to test your understanding of the decision table. You will be able to proceed to actual decision tables only if you answer this question correctly. You will have only 1 attempt to answer this question.

→


University of Pittsburgh

Understanding Question

(You have 1 attempt to answer this question)

(You will be able to proceed to actual decision tables only if you answer this question correctly)

Consider the following decision table:

Your Decision	Other's Decision	Your Payoff	Other's Payoff
A	A	$18	$18
A	B	$6	$15
B	A	$15	$6
B	B	$10	$10

Suppose that this decision table was selected for bonus payment.

If in this table, you chose A and the other worker chose B. What would **your** bonus payment be?

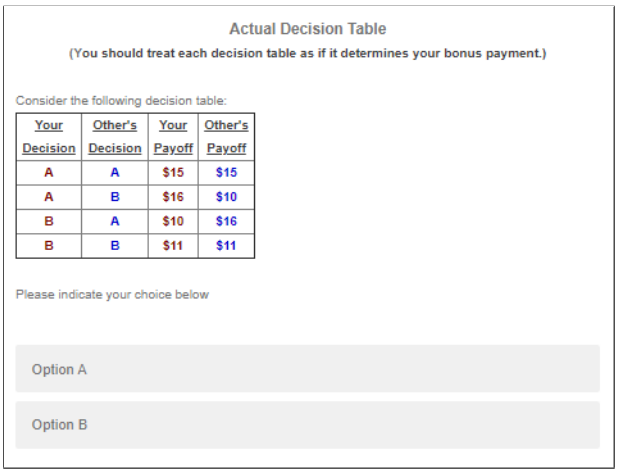
$18

$6

$15

$10

[Next, the four decision tables were presented to the participants in random order. For the re-framed sample option *A* corresponded to *D* and option *B* corresponded to *C*, the screens were otherwise the same as the main sample.]


Actual Decision Table

(You should treat each decision table as if it determines your bonus payment.)

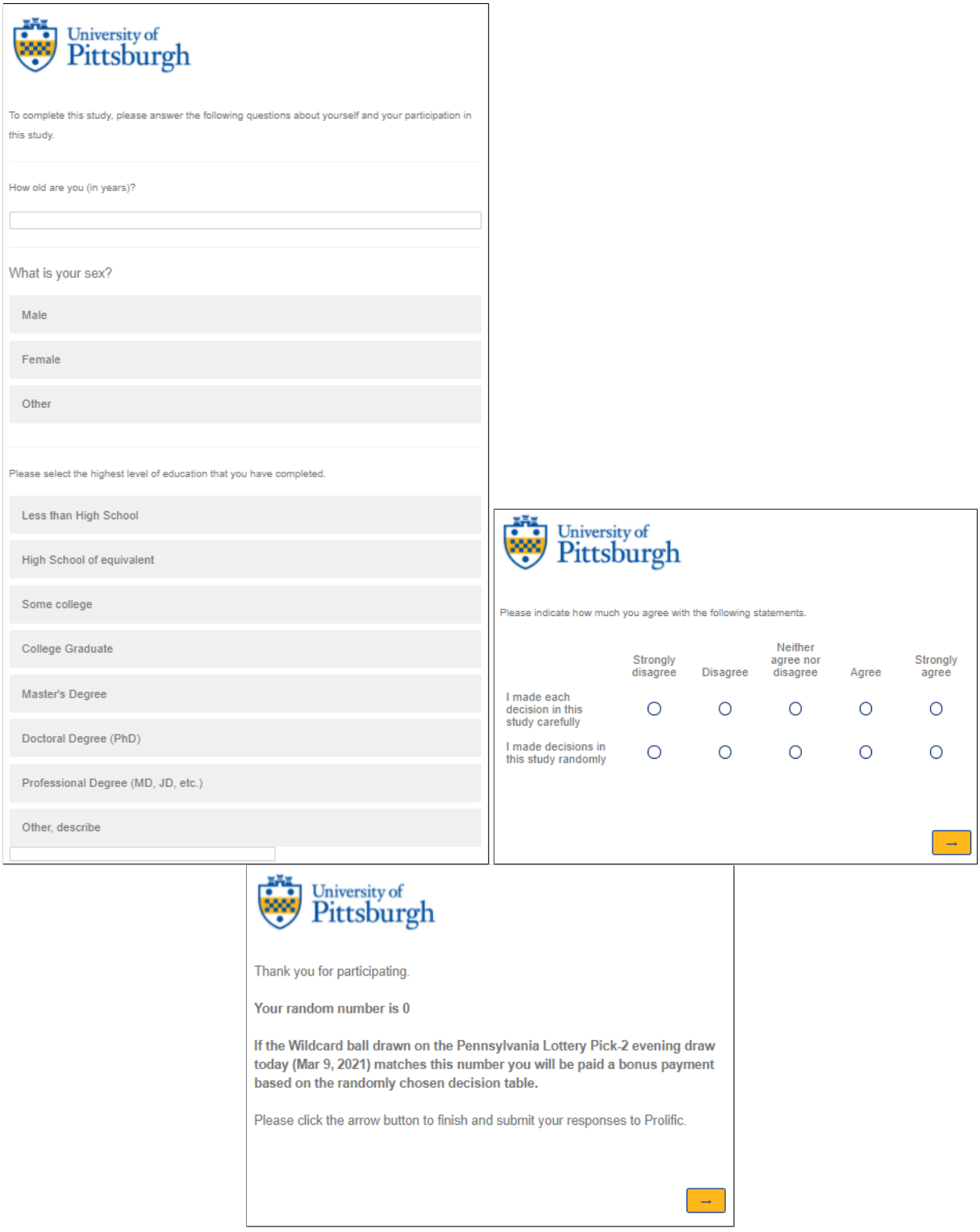
Consider the following decision table:

Your Decision	Other's Decision	Your Payoff	Other's Payoff
A	A	$15	$15
A	B	$16	$10
B	A	$10	$16
B	B	$11	$11

Please indicate your choice below

Option A

Option B



[Fixed fees were credited to the participants immediately upon approval of the submission and the bonus payments were made within 24 hours of completion.]