American Economic Review 2021, 111(2): 687–719
https://doi.org/10.1257/aer.20181065

Testing the Waters: Behavior across Participant Pools[†]

By Erik Snowberg and Leeat Yariv*

We leverage a large-scale incentivized survey eliciting behaviors from (almost) an entire undergraduate university student population, a representative sample of the US population, and Amazon Mechanical Turk (MTurk) to address concerns about the external validity of experiments with student participants. Behavior in the student population offers bounds on behaviors in other populations, and correlations between behaviors are similar across samples. Furthermore, non-student samples exhibit higher levels of noise. Adding historical lab participation data, we find a small set of attributes over which lab participants differ from non-lab participants. An additional set of lab experiments shows no evidence of observer effects. (*JEL* C83, D90, D91)

Lab experiments have been used to amass large amounts of data on human behavior. Yet skepticism persists about whether experimental insights can be generalized beyond experimental labs and university student populations. Critics express concern that experiments are conducted on populations behaviorally distinct from those that generally interest economists, and in environments unlike those in which people actually make decisions. The paucity of data that could alleviate such concerns makes them difficult to address directly.[1]

We provide a data-driven evaluation of several external validity concerns. Leveraging unique data from the Caltech Cohort Study (CCS)—an incentivized, comprehensive behavioral survey of almost the entire undergraduate population of the California Institute of Technology (Caltech)—we shed light on the behavioral differences between undergraduates and other populations, and assess whether behavior is different in the laboratory. Specifically, we provide evidence relevant to three questions. First, are university students behaviorally different than representative populations or convenience samples, in particular Amazon's Mechanical

* Snowberg: Vancouver School of Economics, University of British Columbia and NBER (email: snowberg@mail.ubc.ca); Yariv: Department of Economics, Princeton University, Center for Economic Policy Research, and NBER (email: lyariv@princeton.edu). Stefano DellaVigna was the coeditor for this article. Snowberg gratefully acknowledges the support of NSF grants SES-1156154 and SMA-1329195. Yariv gratefully acknowledges the support of NSF grant SES-1629613 and the Gordon and Betty Moore Foundation grant 1158. We thank four anonymous reviewers for many helpful suggestions. We also thank Marina Agranov, Alessandra Casella, Armin Falk, Guillaume Fréchette, Drew Fudenberg, Johannes Haushoffer, Salvatore Nunnari, Nichole Szembrot, Emanuel Vespa, and seminar audiences at Columbia University, Cornell, NTU, UCSD, and University of Maryland for useful comments and encouragement.

[†] Go to https://doi.org/10.1257/aer.20181065 to visit the article page for additional materials and author disclosure statements.

[1] There is some useful work that examines these issues in specific settings. See the literature review for details.

Turk (MTurk)? Second, are those students who choose to participate in lab experiments different from the general student population? Third, do students change their behavior in the lab?

We show that elicited behaviors differ across our student, representative, and MTurk samples. However, comparative statics and correlations are similar. Differences in correlations can largely be accounted for by statistical insignificance in representative and MTurk samples, driven by higher levels of noise. We see some evidence of differences in observable behaviors between the general student population and self-selected lab participants, though these differences are confined to a minimal set of attributes that are easy to elicit and control for. We see no evidence of participants behaving differently inside the lab than outside of it.

Our results suggest that experiments utilizing undergraduate students, in or outside the lab, allow generalizable inferences about behavior. This is despite undergraduates differing in important ways from other populations. In addition, we document behavioral patterns that are useful both in choosing a venue and population for the execution of laboratory experiments and for the interpretation of experimental results.

To address the questions listed above, we use a large-scale, online, incentivized survey given to several different populations, as detailed in Section II. The survey is designed to elicit a battery of behavioral attributes: risk aversion, altruism, over-confidence, over-precision, implicit attitudes toward gender and race, various strategic interactions, and so on. The main results of this paper rely on the implementation of this incentivized survey using four different populations. First, in the CCS described above, 90 percent of the entire undergraduate population of Caltech participated. The second and third populations were a representative sample of the United States, and a convenience sample of US residents from MTurk, each containing approximately 1,000 participants. These datasets are unusually large for experimental work, and assure that we can detect even relatively small differences. Finally, we brought the CCS into the lab, where approximately 100 Caltech students completed the same survey, but in a substantially different environment. In addition, we wed the CCS data with historical data about lab participation in the Caltech Social Science Experimental Laboratory (SSEL). We used two further samples to address specific questions: (i) a sample of another 1,264 MTurk participants to examine the effects of different incentive levels, and (ii) a sample of 202 students from the University of British Columbia (UBC) laboratory to demonstrate the possibility of running similar surveys using different student populations.

To answer each of our main questions we employ similar analytic methods. We compare the mean levels of elicited behaviors in different samples. Then, we compare the underlying distributions. In general, statistically significant mean differences for a specific elicitation are associated with a first-order stochastic dominance relation between the different samples. Differences in behavior may be crucial for policy; for example, they may suggest when certain interventions are particularly desirable for specific groups. Nonetheless, most studies in experimental economics inspect linkages between behaviors, attributes, and treatments—in a word, correlations.[2] We

[2] A treatment effect is the correlation between a treatment and its outcomes.

therefore generate 55 correlations that we examine across datasets. We assess the degree to which these correlations coincide across datasets.

We see substantial differences in the average levels of elicited behaviors between the representative, MTurk, and student samples, as documented in Section III. Generally, behaviors in the representative and MTurk samples, with both low and high incentives, lie closer to one another than the representative sample is to either student sample. However, even the representative and MTurk samples display substantial differences. Furthermore, differences are quite apparent in the distributions of responses. For most elicitations, response distributions are ranked via first-order stochastic dominance with the CCS on one extreme and the representative sample on the other. Substantively, this means that conclusions from student populations can be useful indicators of lower or upper bounds on behavior in other populations. Descriptively, university students seem to provide upper bounds on "normative rationality" (they are less generous, more risk neutral, etc.) and on "cognitive sophistication" (they exhibit greater cognitive skills and strategic sophistication). This is the case for most of the elicitations in our UBC sample as well, though responses are somewhat closer to those of non-students.

Comparative statics and correlations between elicitations exhibit far greater similarity, and disagreement is largely consistent with the pattern of noise, or measurement error, across samples. While levels of responses may matter for various policy considerations, many social science studies focus on comparative statics or correlations. For example, risk attitudes or altruism are often elicited in experimental studies, as they are suspected to be a potential channel for explaining many observed behaviors. Most of the comparative statics in our data replicate across our samples. Furthermore, in only 7 percent of the correlations we examine do two samples have statistically significant correlations of the opposite sign. The remaining disagreement between samples is largely driven by statistical insignificance of specific correlations in the representative and MTurk samples. Both of these samples have higher levels of noise, and thus greater attenuation of correlations, than the CCS.[3] The fact that the pattern of correlations is similar across populations is encouraging—while estimated relationships may have trouble replicating from sample to sample, it is relatively unlikely that a new sample will produce an opposite result.

MTurk is widely used by economists. Presumably, this is due to MTurk enabling the collection of large volumes of data quickly and cheaply. We note, however, that representative samples have similar features in terms of both ease of access and cost. Nonetheless, the correlations between behaviors seen using the MTurk sample are fairly similar to those observed in the two other samples. A second MTurk sample, described above, shows almost no effect of halving incentives and the passage of 3.5 years. There is, thus, a cost-benefit tradeoff: while student-based studies often entail higher costs, they generally exhibit less noise.

Students who choose to participate in lab experiments differ little, behaviorally, from the overall population of students, as documented in Section IV. This addresses the concern that the same attributes that cause a student to *select* into lab experiments

[3] Noise is anything that may cause duplicate elicitations of the same behavior to be different. It includes classical measurement error, trembles, inattention, and so on. See Gillen, Snowberg, and Yariv (2019) for background on the assessment of noise, or measurement error, its effects, and statistical approaches for overcoming it.

may be driving the results observed in certain experimental settings. For example, if lab participants are motivated by altruism, lab results regarding altruism will be different from what would be observed in the general population (see Levitt and List 2007b). In order to assess these concerns, we wed data from the CCS with records of participation in lab experiments from the Social Science Experimental Laboratory (SSEL) at Caltech. As nearly all Caltech undergraduates completed the CCS, we can compare responses of those individuals who participate in lab experiments—unweighted or weighted by the number of experiments they participate in per year—to the overall population of students. Lab participants are slightly *less* generous, more risk averse, and more likely to lie. These differences, while statistically significant, are small in magnitude.

Finally, we observe behavior in the lab that is virtually identical to that on the incentivized survey, as documented in Section V. This addresses a concern that *observer effects*, reviewed in Section I, are driving experimental results. For example, experimental results showing low levels of lying and high levels of generosity compared with the "rational" benchmark may be an artifact of participants behaving differently when directly monitored by experimenters, or simply wishing to appear more ethical (Levitt and List 2007a, b). In order to examine these concerns, we invited students to the lab to take the CCS survey. We see very little difference between responses in and outside of the lab. Thus, to the extent that observer effects are important, they are not particularly sensitive to the level of monitoring by, or presence of, the researchers. Participants in our lab experiments are, however, less generous, and score higher on cognitive tasks. Yet we also find that repeated administration of the CCS is associated with reductions in generosity and increased performance on cognitive tasks.[4] While we cannot rule out some lab-specific effects on these measures, the results for generosity, at least, run counter to expressed concerns.

Taken together, our findings should be reassuring to researchers using standard student-based experiments. While we see large differences in behaviors across the vastly different populations in the CCS, the representative sample, and MTurk, these differences have limited impacts on most comparative statics and correlations between the behaviors we elicit. In addition, behavior in student populations may offer convenient bounds on behaviors in other populations. Furthermore, behavioral differences due to selection into the lab are quite limited. Last, behavior in the lab is practically indistinguishable from an experimental setting outside of the lab.

We stress that our study is unable to speak to all concerns about the experimental enterprise. For example, we do not address worries that individuals may respond differently to choices that do not mimic the somewhat artificial designs often seen in the lab. We are sympathetic to this view, and certainly believe that framing of decisions matters for choices.[5] Nonetheless, we believe that treating each specific application as *sui generis* drastically limits the generalizability of any observation made either in the lab or in the field. Instead, this paper suggests that some observations on behavioral tendencies are consistent across samples. Moreover, these observations

[4] Participants were not told ahead of time that they would be responding to the CCS in the lab in order to avoid selection effects. Repeated surveys did not impact responses to other elicitations, see Section IIIB.

[5] There are, however, several studies that illustrate the similarity of field and experimental lab data in various contexts, including peer effects on productivity (Herbst and Mas 2015), tax compliance (Alm, Bloomquist, and McKee 2015), and corruption (Armantier and Boly 2013).

can be made using standard lab or survey methodology. We hope the approach we introduce in this paper opens doors to further data-driven analyses of other aspects of external validity.

I. Related Literature

Each of the questions we address has important precedents in the literature. A small number of papers compare students to representative populations and MTurk. In line with our results, university students are less generous than representative samples of Zurich and Norway (Falk, Meier, and Zehnder 2013; Cappelen et al. 2015).[6] In addition, MTurk participants behave similarly to university students on several "heuristic and biases" experiments and non-incentivized games, as well as (incentivized) repeated public goods and Prisoner's dilemma games (Paolacci, Chandler, and Ipeirotis 2010; Horton, Rand, and Zeckhauser 2011; Berinsky et al. 2012; Goodman, Cryder, and Cheema 2013; Arechar, Gächter, and Molleman 2018).[7] Hauser, Paolacci, and Chandler (2019) offer a survey of research on the topic. We build on this work by comparing university students, a representative sample, and MTurk across a wide range of incentivized, fundamental behaviors. In addition, our university sample is (almost) exhaustive, as opposed to prior work that studies only a subset of the university population—usually those that self-select into laboratory experiments.

Several papers study whether students' self-selection into lab experiments creates bias. This work shows that selection into lab experiments from broader student populations—such as those taking introductory economics—is not related to risk aversion or generosity (Harrison, Lau, and Rutström 2009; Cleave, Nikiforakis, and Slonim 2013; Falk, Meier, and Zehnder 2013). Even so, guaranteed show-up fees yield somewhat more risk-averse lab participants (Harrison, Lau, and Rutström 2009). We add to this work by assessing selection over a large array of fundamental behaviors, and using data from (almost) the *entire* student population from which lab participants are (self-)selected.

The literature studying "observer" (or Hawthorne) effects—the idea that the mere presence of an experimenter may change behavior—is much larger than both of the literatures reviewed above, combined. In the popular and academic imagination, this effect is tied to a series of experiments, most conducted by Elton Mayo, that took place at Western Electric's factory in Hawthorne, Illinois, in the late 1920s and early 1930s (see Mayo 1933). When studying the impacts of physical conditions

[6] Belot, Duch, and Miller (2015) compare behavior in five different games between Oxford students and local (nonrepresentative) non-students. The results are in line with ours: non-students have more salient other-regarding preferences and exhibit less sophisticated strategic thinking. Exadaktylos, Espín, and Brañas-Garza (2013) similarly examine students and non-students, and find similar results in the dictator, ultimatum, and trust games. They further report similar responses from occasional and frequent participants. See Falk, Meier, and Zehnder (2013) for a related study focusing on trust games. In contrast, Fosgaard (2018) shows substantial differences between students and lab participants drawn from the general Copenhagen population in a repeated public goods game. Fréchette (2016) reviews experiments conducted with nonstandard participants, including animals, people living in token economies, and so on.

[7] Coppock (2018) replicates the results of multiple political science experiments, originally run on representative samples, on MTurk.

on productivity, workers under observation seemed to outperform those in a control group, even when nearly identical conditions were imposed.[8]

In dictator games, Hoffman et al. (1994) find an observer effect, while Bolton, Katok, and Zwick (1998) do not.[9] Hergueux and Jacquemet (2015) do not explicitly mention the observer effect, but compare behavior in social-preference games and behavioral elicitations in the lab and online. They, too, find few significant differences in responses. These studies all use a between-participant design, with different participants in different treatments. In contrast, we use a within-participant design: we consider the same participants in different environments. Our comparison of behavior in the lab—where at least one experimenter was present throughout the experimental sessions—to an incentivized online survey— that participants took at a time and place of their choosing, with no supervision—provides a test of the presence of an observer effect across a wide range of fundamental behaviors. By and large, we find little evidence of an observer effect. An important caveat to all these papers, including ours, is that they cannot test whether participating in a study in and of itself affects behavior. Since experimental platforms are not naturalistic, it is difficult to imagine an experiment overcoming this issue. Nonetheless, our results suggest that even if some such responses are present, they are not sensitive to the level of monitoring by researchers.

To summarize, we contribute to the extant literature in three important ways. First, we compare multiple populations simultaneously: representative, MTurk, a full university population, and self-selected students that participate in laboratory experiments. Second, for these populations, we have an unusually large battery of preference elicitations. Third, we compare the populations in our data in several ways: in terms of level of behavior, in terms of comparative statics and correlations, and in terms of noise. Online Appendix Table A.2 summarizes some of the leading papers described above and our contributions.

More broadly, controversy over lab experiments' value and the use of student populations is nearly as old as the methodologies themselves, with vocal critics and defenders. Concerns about lab data's generalizability go back to at least Orne (1962), and have been discussed in various papers (see, for example, Guala and Mittone 2005, Schram 2005). They received a great deal of attention in a sequence of papers by Levitt and List (2007a, b, 2008). Multiple recent papers advocate lab and experimental data's usefulness (largely in response to Levitt and List—see, for instance, Falk and Heckman 2009, Gächter 2010, Camerer 2015, and Kessler and Vesterlund 2015). While we certainly do not speak to all concerns voiced over the experimental enterprise, we provide some data-based insights on the extent of general selection issues regarding experiments and surveys run on student populations.

[8] These studies may not actually show an observer effect (see Jones 1992, Levitt and List 2011, as well as a survey of experiments in Gillespie 1993).

[9] Laury, Walker, and Williams (1995) find no observer effect in public goods games. Anderhub, Müller, and Schmidt (2001) find few differences between online and lab participants in a game reminiscent of a consumption-saving problem.

II. The Data

The foundation of our analysis is the Caltech Cohort Study (CCS), a repeated, incentivized survey covering over 90 percent of the Caltech student body. We administered the same survey, with additional demographic questions, to two other populations: a representative sample, and a convenience sample from MTurk. The survey itself elicits an array of behavioral attributes including risk aversion, discounting, competitiveness, cognitive sophistication, implicit attitudes toward gender and race, generosity, honesty, overprecision, a probabilistic measure of lying, and so on. Here we describe the samples in more detail, before proceeding to a brief description of elicitations we use heavily throughout this paper.

A. *The Student Samples*

Caltech is an independent, privately supported university located in Pasadena, California. It has approximately 900 undergraduate students, of which ~ 40 percent are women. The Caltech Cohort Study (CCS) is comprised of various versions of an incentivized survey administered in the fall of 2013, 2014, and 2015 and the spring of 2015.

The data used in this paper come almost exclusively from the spring 2015 installment, which utilized the same version of the survey run on the other populations we inspect. Other surveys contained some, but not all, of the elicitations used here. In the spring of 2015, 91 percent of the enrolled undergraduate student body $(819/899)$ responded to the survey.[10] The average payment was \$29.08 and the median time for survey completion was 35 minutes.[11] It is important to note that there is little concern about self-selection into the CCS from the participant population, due to our 90+ percent response rates.

In Section IV, we use records from the Social Science Experimental Laboratory (SSEL) at Caltech. These records provide the number of experiments each individual in the SSEL participant pool attended. For the cohorts entering between 2011 and 2014, 403 students participated in at least one experiment held at SSEL by the summer of 2015. Of those who were eligible to participate in the CCS, 96 percent $(350/370)$ responded to the CCS spring 2015 survey.[12] Conditional on participating in at least one experimental session, the median participation rate was 2 experiments per year.

[10] To obtain such a high participation rate, we promoted the survey through multiple emails. By the third installment, the students viewed it as a well-known feature of the institute. As discussed in Section IIIA there was no correlation between how quickly students took the survey and behavioral attributes, mitigating concerns that the remaining 9 percent are substantially different. The lags between implementations of the survey were designed to mitigate concerns pertaining to a desire for consistency across surveys by participants. Conversations with multiple surveying experts suggested that for the complexity of our tasks, and the time lags we imposed, consistency biases would be nonexistent. We were unable to find data or studies that might provide information about such concerns, and hope future work will fill this gap.

[11] Similar participation rates across all our surveys limited attrition. In particular, of those who had taken the survey in the spring of 2015, 96 percent also took the survey in the fall of 2014. Similarly, of those who took the survey in 2013 and did not graduate, 89 percent also took the survey in the fall of 2014.

[12] As only enrolled students were eligible, only 370 of the 403 students could participate. The remaining 33 students either graduated early, or were on leave of absence.

In Section V, we utilize data from a series of lab experiments we conducted in summer 2015. These experiments asked participants to fill out the spring 2015 survey on SSEL's computers, with us present in the room. The experiments were advertised with a neutral name so as not to introduce selection due to the content of the experiments themselves. There were 97 participants in our lab experiment, and the average payment was $34.94 (plus a show-up fee of $10). The median completion time was 31 minutes, slightly shorter than the median completion time of the CCS when run online. Of the 97 experimental participants, 96 responded to the spring 2015 CCS survey.

Caltech is highly selective, which may lead to concerns that this population is different from those utilized in most lab experiments. Replications of standard experiments and elicitations—of risk, altruism in the dictator game, and so on—yield similar results to other student pools (see the online Appendix of Gillen, Snowberg, and Yariv 2019). However, Caltech students are known to be on the extreme edge of certain types of cognitive ability. Caltech students may also be on the far edge of other behaviors, a notion that finds some support from an additional, though smaller, sample from the University of British Columbia (UBC) lab.[13] If this is the case, our results may provide a rough guide for the maximum differences between students and the other populations we consider.

B. *The Representative Sample*

Survey Sampling International (SSI) was founded in 1977 and provides a platform for survey researchers around the world to recruit panels of respondents based on various demographic attributes.[14] In spring 2017, we utilized the SSI participant pool to run the CCS spring 2015 survey on a representative sample of the US population. We had 1,000 participants that were representative of the US population across age, income, and gender. The average payment was $10.26, with an additional $3 required by SSI for each survey completion.[15] The median completion time was 33 minutes.

C. *Amazon Mechanical Turk*

In spring 2016, we conducted our survey with a sample of 995 US-based Amazon Mechanical Turk (MTurk) users. The average payment was $10.50 per participant. The median completion time was 35 minutes. These incentives may appear high relative to those commonly used on MTurk.[16] As we discuss in Section IIIB, we also

[13] Specifically, in the summer of 2019, we ran 10 sessions at the UBC lab with a total of 202 participants. In these sessions, participants responded to a survey containing a set of elicitations, analogous to the 2015 CCS. The average payment was CDN$31.15, plus a show up fee of CDN$10. The median time to complete the survey was 48 minutes.

[14] SSI has since merged with Research Now, and the combined firm was renamed Dynata, see https://www.dynata.com/press/announcing-new-name-and-brand-research-now-ssi-is-now-dynata/.

[15] We paid SSI $3 per respondent. We do not know what fraction of that amount was passed on to the participants themselves. The payments from incentives are at least four times as large as standard participant payments through SSI. We were dissuaded from using larger amounts.

[16] While precise statistics of Mechanical Turk users are not released, estimates exist of an hourly "wage" ranging between $1–$5. See http://priceonomics.com/who-makes-below-minimum-wage-in-the-mechanical/. See also Dube et al. (2020), who estimate low labor-supply elasticities on MTurk.

conducted an auxiliary MTurk survey with a sample of 1,264 participants in summer 2019, with the incentives halved. As we show, results remained virtually unchanged. We use our original MTurk survey for most of this paper's analysis, as incentives there were comparable to those in our other samples.

With the emergence of MTurk and other convenience samples as an important resource for scholars, some recent work has already characterized the demographic profile of MTurk users, and its comparison to the US population (see, for example, Difalla, Filatova, and Ipeirotis 2018; Berinsky et al. 2012; and Huff and Tingley 2015). We find similar patterns comparing the MTurk and representative samples, which are summarized in online Appendix Table A.1.

D. *Description of Elicitations*

We examine a standard set of elicitations that we believe are particularly important for experimental work.[17] 100 survey tokens were valued at \$1 for our student sample, while 300 survey tokens were valued at \$1 for our representative and our original MTurk samples. Participants were paid for all tasks. The location of many questions within the survey was determined at random. Since we observe no order effects, we report aggregate results throughout.

1. Risk Elicitations.—We used three different risk elicitation techniques.

Risky Projects: Following Gneezy and Potters (1997), participants were asked to allocate an endowment of tokens between a safe option (keeping them), and a project that returns some multiple of the tokens with probability p, otherwise returning nothing. In spring 2015, two projects were used: the first returning 3 tokens per token invested $p = 35\%$ of the time, and the second returning 2.5 tokens $p = 50\%$ of the time.

Risky Urns: Two multiple price lists (MPLs) asked participants to choose between a lottery and sure amounts. The lottery would pay off if a ball of the color the participant chose was drawn. The first urn contained 20 balls (10 black and 10 red) and paid 100 tokens. The second contained 30 balls (15 black and 15 red) and paid 150 tokens. Taking the first MPL as an example, participants were first asked to choose the color they wanted to pay off, if drawn. They were then presented with a list of choices between a sure amount that increased in units of 10 tokens from 0 to 100, or the gamble on the urn.[18]

[17] See the online Appendix of Gillen, Snowberg, and Yariv (2019) for precise question wordings.

[18] In order to prevent multiple crossovers, the online form automatically selected the lottery over a 0-token sure amount, and 100 tokens over the lottery. In addition, participants needed to make only one choice, and all other rows were automatically filled in to be consistent with that choice. For an overview of risk elicitation techniques, see Charness, Gneezy, and Imas (2013).

Qualitative: Following Dohmen et al. (2010), participants were asked to rate themselves, on a scale of 0–10, in terms of their willingness to take risks.

2. *Monthly Discount Rate* (δ).—Participants were asked a hypothetical question about how much money they would have to be paid in 60 days to forego a \$150 payment in 30 days.[19] This was converted to a monthly discount rate (δ) using standard techniques; that is, $\delta = \$150/\text{answer}$. As this task featured hypothetical incentives, there were many extreme answers. We therefore trim the top and bottom 10 percent of answers: those that demand less than \$150, or more than \$400.

3. *Dictator Giving*.—There were four tasks that asked participants to allocate a stock of tokens between themselves and another randomly chosen, anonymous participant. In the first dictator game, participants were given a stock of 300 points, and in the second, 100 points. In a third dictator game, any amount given to the other participant was doubled; in a fourth, points allocated to the other participant were halved. In both of these latter tasks, allocations were made out of a stock of 100 points.

4. *Prisoner's dilemma*.—There were two symmetric Prisoner's dilemma games with different payoffs. Across the two games, payoffs were scaled by a common factor to keep the same relative incentives. Participants were told that, for each game, they would be randomly matched with another participant at the end of the survey and paid based on their own choice and the choice of the other participant.

5. *Lying*.—Two questions were meant to (probabilistically) measure the willingness of participants to lie. Both asked participants to toss a coin some fixed number of times and report an outcome. In the first task, participants were asked to report the number of heads they got in 5 coin tosses, knowing they would be paid 30 tokens for each. In the second, participants were asked to report the number of switches (or number of runs minus one) they got in a sequence of 10 coin tosses, knowing they would, again, be paid 30 tokens for each.

6. *Cognitive Tasks*.—We used two types of cognitive tasks.

Raven's Matrices: Participants were asked to complete five Raven's (1936) matrices, which are commonly used for assessing abstract reasoning. Each Raven's matrix consisted of a 3×3 matrix with eight of the nine cells featuring a geometric design. Participants had to choose, out of six possibilities, the correct geometric pattern to complete the matrix. Participants were given 30 seconds to complete each task, and were paid 20 tokens for each correctly completed matrix.

Cognitive Reflection Test (CRT): Participants responded to variations on the three questions from Frederick (2005). These questions have an "obvious" wrong answer, and are thus designed to measure individuals' ability to reflect on problems

[19] The fact that both payoffs are in the future removes any effects of present bias.

and override immediate intuitions.[20] As in the Raven's matrices task, participants were given 30 seconds to complete each question, and paid 20 tokens for each question answered correctly.

7. Confidence in Guesses.—Similar to the over-precision task of Ortoleva and Snowberg (2015), participants were asked to guess the number of jellybeans in (a picture of) a jar, and then rate how confident they were about their guess. Ratings were on a six point scale, ranging from "not confident at all" to "certain." Participants repeated this task three times, and we averaged their responses.

8. Competition.—The essential elements of Niederle and Vesterlund (2007) were presented to participants. First, they had three minutes to solve as many sums of five two-digit numbers as they could. They were told they would be grouped with three other participants at random. If they solved the most sums correctly within their group of four, they would receive 40 tokens per correct sum. Next, participants repeated the task, but were asked before whether they preferred to be paid the same way as the first time, or whether they preferred to receive 10 tokens per correct sum regardless of others' performance. This second question provides an elicitation of willingness to compete.[21]

9. Implicit Association Tests (IAT).—We assessed implicit attitudes toward gender and race separately using two implicit association tests (Greenwald, McGhee, and Schwartz 1998).[22] Although controversial, IATs are often viewed as measures of discriminatory attitudes.

III. Comparison of Different Samples

We begin our analysis by comparing different participant pools: university students (from the CCS and UBC), a convenience sample (from MTurk), and a representative sample of the United States (from SSI).[23] We see large differences in the average levels of the behaviors we examine. For most behaviors there are clear first-order stochastic dominance relationships between the samples, with the CCS on one extreme and the representative sample on the other. This implies that results from experiments on university students may usefully bound behaviors in representative samples. In particular, students, especially Caltech students, behave in a more normatively rational and cognitively sophisticated way. These mean differences do not tend to lead to disagreement in comparative statics or the signs of correlations between behaviors. The differences in the levels of statistical significance associated with those correlations are broadly consistent with differences in noise across the

[20] We used variations on the original questions, as some participants may have been exposed to those.

[21] For further details, see Gillen, Snowberg, and Yariv (2019).

[22] These scores are derived from the differences between mean latencies across the two combined classification stages in each of the IATs, see Greenwald, Nosek, and Banaji (2003). The gender task measured the implicit association between gender and the sciences or humanities.

[23] The UBC data was collected in a lab environment rather than through an online survey. As we report in Section V, we observe few differences between online and lab responses to the CCS.

samples. In particular, the CCS exhibits more significant correlations and also the lowest level of noise.

A. *Differences in Behavior*

The average measures of each behavior are quite different across the samples, as shown in Table 1.[24] In the case of the CCS, this should be unsurprising, and perhaps even reassuring, as students at elite universities are an extraordinary set of individuals. Moreover, prior research has established links between intellectual ability and various behaviors such as risk aversion and discounting (Dohmen et al. 2010, 2018).

It is also clear from Table 1 that the representative and MTurk samples appear closer to each other than the representative sample is to either of the student samples.[25] The average levels of behavioral measures in the MTurk sample are usually between those in the representative sample and the CCS. In fact, several measures—reflecting risk attitudes, discounting, confidence, and attitudes towards race—show no significant difference across the MTurk and representative samples. For risk aversion, the magnitude of the differences between the CCS and other samples varies across measures. Differences are the most substantial for the risky project measures. These elicitations mimic a stock/bond portfolio choice (or risky/safe assets) that resemble investment decisions, and are therefore particularly important for many economic settings.[26]

The average response in the UBC sample is usually between the representative and CCS samples, and, in most cases, closer to the CCS than the representative sample. There are some notable exceptions: on discounting, competition, cognitive tasks, and IAT gender the UBC sample is closer to the representative sample than the CCS. Thus, behavior on the CCS may offer a bound on differences between student samples and other, commonly used, samples. A more comprehensive comparison of student samples would be of great use for establishing this point conclusively.

The mean differences we report are not only statistically significant, they are also substantively large. A summary measure of statistical difference between two samples is the number of control variables needed in order for the two samples to be balanced on the remaining variables. That is, we look for the minimal number of variables such that, controlling for those, differences between the two samples on the remaining variables are not jointly significant (using an F-test). For the representative and MTurk samples, one would need to control for 10 of the variables in Table 1 for those samples to be statistically balanced on the other 10. For the representative and CCS sample, one would need 12 controls. Finally, for the MTurk and CCS samples, one would need 9 controls.[27]

[24] Throughout, fractions reported for the Prisoner's dilemma are averaged across our two games as those led to statistically indistinguishable results.

[25] The differences we find are not due to differences in gender and race composition, as can be seen from online Appendix Table A.7. This table re-weights the CCS sample to match the gender and racial composition of the representative samples.

[26] Gillen, Snowberg, and Yariv (2019) also show that the risky project measures are relatively stable over time, exhibit less measurement error than other risk-attitude elicitations, and have different responses across genders. We replicate this difference between genders across all three of our main samples in Table 2.

[27] Since our UBC sample is far smaller, the analogous calculations should be interpreted with care, as they are not directly comparable. Nonetheless, for theUBC and representative samples, one would need to control for 11 of

TABLE 1—DIFFERENCES IN CHOICES: FOUR SAMPLES

	Samples				Differences			
	Rep.	MTurk	CCS	UBC	Rep. − MTurk	Rep. − CCS	MTurk − CCS	CCS − UBC
First risky project (out of 100)	46 (0.89)	44 (0.85)	59 (1.2)	52 (2.0)	2.7 (1.2)	−13 (1.5)	−16 (1.4)	7.5 (2.6)
Second risky project (out of 200)	95 (1.8)	98 (1.7)	143 (2.1)	111 (3.7)	−2.7 (2.5)	−48 (2.8)	−45 (2.7)	31 (4.6)
First risky urn (20 balls)	49 (0.76)	56 (0.63)	59 (0.52)	60 (1.3)	−7.3 (0.99)	−10 (0.96)	−3.2 (0.84)	−1.1 (1.2)
Second risky urn (30 balls)	67 (1.2)	78 (0.96)	86 (0.73)	84 (2.0)	−11 (1.6)	−19 (1.5)	−8.1 (1.3)	2.3 (1.8)
Qualitative risk aversion	5.0 (0.08)	4.9 (0.08)	5.8 (0.08)	5.3 (0.15)	0.11 (0.11)	−0.76 (0.11)	−0.87 (0.11)	0.44 (0.17)
Monthly discount rate (δ)	0.67 (0.01)	0.67 (0.01)	0.77 (0.01)	0.66 (0.01)	−0.00 (0.01)	−0.10 (0.01)	−0.10 (0.01)	0.11 (0.02)
First dictator game (given out of 100)	39 (0.58)	26 (0.71)	14 (0.84)	16 (1.4)	14 (0.91)	25 (1.0)	12 (1.1)	−2.4 (1.8)
Second dictator game (given out of 300)	115 (1.7)	74 (2.0)	38 (2.4)	47 (4.1)	41 (2.7)	77 (2.9)	36 (3.1)	−9.5 (5.2)
Dictator, tokens given are doubled	39 (0.62)	30 (0.79)	26 (1.2)	24 (1.8)	8.9 (1.0)	13 (1.3)	3.7 (1.4)	1.9 (2.6)
Dictator, tokens given are halved	39 (0.61)	25 (0.74)	9.0 (0.68)	17 (1.5)	14 (0.95)	30 (0.91)	16 (1.0)	−7.7 (1.6)
Prisoner's dilemma (percent dominant strategy)	46 (1.2)	57 (1.3)	68 (1.5)	72 (2.9)	−11 (1.8)	−22 (1.9)	−11 (2.0)	−4.0 (3.3)
Reported heads (out of 5)	2.9 (0.03)	3.0 (0.03)	3.3 (0.04)	3.4 (0.08)	−0.14 (0.05)	−0.41 (0.05)	−0.28 (0.05)	−0.08 (0.10)
Reported switches (out of 9)	4.4 (0.06)	4.5 (0.06)	5.5 (0.07)	5.5 (0.14)	−0.18 (0.08)	−1.1 (0.09)	−0.95 (0.09)	−0.04 (0.16)
Raven's matrices (out of 5)	1.2 (0.03)	1.3 (0.04)	1.8 (0.04)	1.3 (0.08)	−0.17 (0.05)	−0.63 (0.05)	−0.46 (0.06)	0.46 (0.10)
CRT (out of 3)	0.46 (0.03)	1.4 (0.04)	1.7 (0.04)	1.1 (0.07)	−0.89 (0.04)	−1.2 (0.04)	−0.31 (0.05)	0.57 (0.08)
Confidence in guesses	2.9 (0.03)	2.9 (0.03)	3.1 (0.03)	3.3 (0.06)	−0.05 (0.05)	−0.25 (0.05)	−0.20 (0.04)	−0.14 (0.07)
Competition (percent competing)	40 (1.6)	29 (1.5)	33 (1.7)	45 (3.5)	11 (2.1)	65 (2.3)	−4.0 (2.2)	−12 (3.8)
IAT race	59 (8.2)	68 (4.8)	81 (5.6)	105 (14)	−8.9 (9.5)	−23 (10)	−14 (7.3)	−24 (13)
IAT gender	104 (5.9)	90 (4.8)	95 (5.9)	106 (13)	13 (7.6)	8.8 (8.4)	−4.6 (7.5)	−11 (13)
Percent male	47 (1.6)	50 (1.6)	62 (1.7)	47 (3.5)	−3.5 (2.2)	−15 (2.3)	−11 (2.3)	15 (3.8)
Observations	1,000	995	819	202	–	–	–	–

Note: Standard errors in parentheses.

The differences between samples are also large in terms of magnitude. For example, differences in the amount allocated in the risky project measures are between 15–25 percent of the maximum possible differences, or around 50–75 percent of

the variables for those samples to be statistically balanced. For the UBC sample and the MTurk sample, or the UBC sample and the CCS, one would need to control for 10 of the variables.

a standard deviation of these measures within a given sample. The differences in discount rates are similarly substantial in terms of sample standard deviations.[28] Differences in giving in dictator games are also large, corresponding to approximately 10 percent of the budget. This last point is important as it suggests that generosity in student populations, sometimes viewed as a student or lab artifact, may actually offer a lower bound of generosity in the overall population.[29] Nonetheless, we note that comparative statics across our variants of the dictator game are similar across our samples and identical for our MTurk and representative samples. We return to a general analysis of the connections between elicited behaviors in Section IIC.

The distributions of various behavioral elicitations, in Figure 1, display a clear pattern: there is a first-order stochastic dominance relationship between the CCS and the representative sample in most behaviors.[30] Additionally, the distribution in the MTurk sample is generally closer to the representative sample, as implied by Table 1. In some cases, we also see a clear first-order stochastic dominance relationship between the MTurk distributions and those of the CCS and the representative sample. These observations also largely hold when considering the UBC instead of the MTurk sample. Overall, these facts imply that the differences in means in Table 1 are not driven by small groups of people with extreme behaviors, but are, instead, population-level shifts.

As the CCS exhibits first-order stochastic dominance relationships with the other samples, it offers bounds on the behaviors we measure. Observations in the CCS serve as an upper-bound on the distribution of risk aversion, discounting, lying, and performance on intellectual tasks, and a lower bound on generosity, as shown in Figure 1. Overall, the Caltech and UBC student populations are closer to the ideal of normative rationality and exhibit greater cognitive sophistication than the other populations we examine. This might be useful for experiments mimicking certain economic environments—say, ones where participants stand for professional investors in large firms. However, it may also lead to muted "behavioral" aspects of choice when using university students as a sample population.

We find no correlation between the time difference (from announcement of the CCS survey to when a participant completed it) and the behavioral proxies we measure. While CCS participants exhibited wide variation in their enthusiasm, with some completing the survey as soon as it was announced, and some waiting several weeks, this time difference was uncorrelated with behavioral proxies, as shown in online Appendix Table A.4. There are no statistically significant differences in behavioral measures between the overall population and the 374 people who took the survey after a single reminder, the 530 that took it within a week of launch, and the remaining 289 people who took it after a week had passed. There is one exception, not measured on the survey: as detailed in online Appendix Table A.5, those

[28] This can be directly calibrated to real-world examples. Suppose someone has a monthly salary of \$6,000. Assume that a one-month training decreases income in that month by \$4,000, but increases it to some amount y thereafter. With a monthly discount rate of 0.67 (MTurk and representative), the future wage y would need to be at least \$8,000 to justify the investment in training. With a discount rate of 0.77 (CCS), the future wage y would need to surpass only \$7,200.

[29] This observation is in line with prior work on this elicitation, see Falk, Meier, and Zehnder (2013) and Cappelen et al. (2015), as well as the discussion of the literature in Section I.

[30] Figure A.1 in the online Appendix contains the cumulative differences of features described in Table 1 that are not in Figure 1.

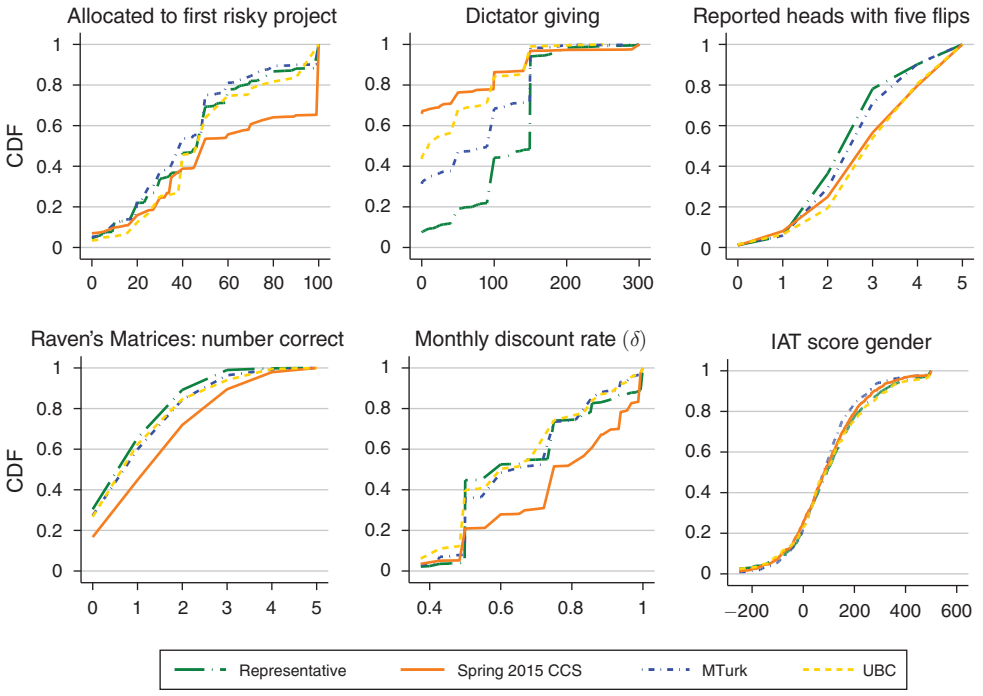

FIGURE 1. DISTRIBUTION OF RESPONSES IN REPRESENTATIVE SAMPLE, MTURK, AND CCS

that took more than a week to take the CCS were much less likely to participate in experiments in the Caltech Social Science Experimental Laboratory (SSEL). This suggests that the fact that we were unable to survey about 9 percent of the Caltech undergraduate student body does not significantly impact the results discussed here, and previews the results in Sections IV and V.[31]

B. *Learning and Time Stability*

The CCS was repeated multiple times with many of the same participants. This repetition may have affected responses in a number of ways, including through participants learning about how to respond to the tasks or about their own preferences. Changes in distributions of responses over time would imply that responses in later installments of the CCS survey are an artifact of repetition, rather than what one would see in a standard lab setting. Fortunately, we find few differences in the distribution of responses on the CCS over time.

Of the few tasks that were repeated across multiple versions of the survey, there were only two classes of elicitations with strong variation over time: cognitive tasks (CRT and Raven's matrices), and giving in the dictator game. The first panel of Figure 2 shows the typical pattern of responses comparing the first installment of the

[31] These observations are reminiscent of those noted by Arechar, Kraft-Todd, and Rand (2017). They document that the time of day, as well as the day of the week, when responses are submitted on MTurk do not strongly associate with various participant attributes.

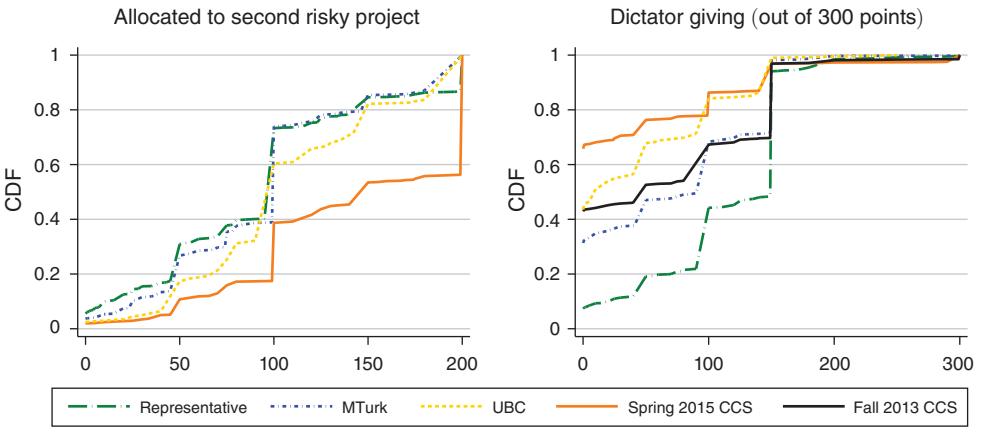

FIGURE 2. REPETITION DOES NOT ALTER MOST BEHAVIORS, EXCEPT DICTATOR GIVING

CCS, in the fall of 2013, with the installment we focus on here, conducted 1.5 years later. The distributions of responses on the two installments of the CCS for the second risky project are nearly identical, and clearly different than those emerging from MTurk and the representative sample.[32] The second panel shows the atypical pattern in dictator giving, where participants showed far more generosity initially: so much so that the distribution is quite similar to that observed on MTurk. Thus, it appears that although the CCS features a less generous participant pool than the representative sample, this difference is exaggerated by some participants changing their behavior as they complete the task multiple times.[33]

The two cognitive exercises were introduced on the spring 2015 survey, and the same questions were repeated in the fall of 2015. Focusing on those participants who took both surveys, and *did not* participate in our lab experiment in the summer of 2015 (500 people), the mean CRT score was 1.67 in the spring and 1.95 in the fall. Similarly, the mean number of Raven's matrices was 1.85 in the Spring and 1.91 in the fall. Thus, repetition of these tasks may widen the gap in performance between the CCS and other pools. However, the difference we observe with first-time responders is clear.

To further understand the stability of responses in different participant pools and the effect of different incentive levels, we ran our MTurk study a second time in the summer of 2019, and collected an entirely new participant sample. This version also included attention screeners in order to weed out "bots" and low-attention participants. We received a total of 1,264 responses, of which 212 failed at least one attention screener. The conversion rate between tokens and money was reduced by one-half, resulting in an average payment of $5.21.[34]

[32] The fall 2013 survey contained only one risky-project task, which was identical to the second risky-project task on the spring 2015 CCS.

[33] Those who first participated in the CCS after fall 2013 also seem less generous, suggesting that some behavioral change is due to social interactions, rather than interactions with the task itself.

[34] This level of compensation is standard, see, for example, DellaVigna and Pope (2018).

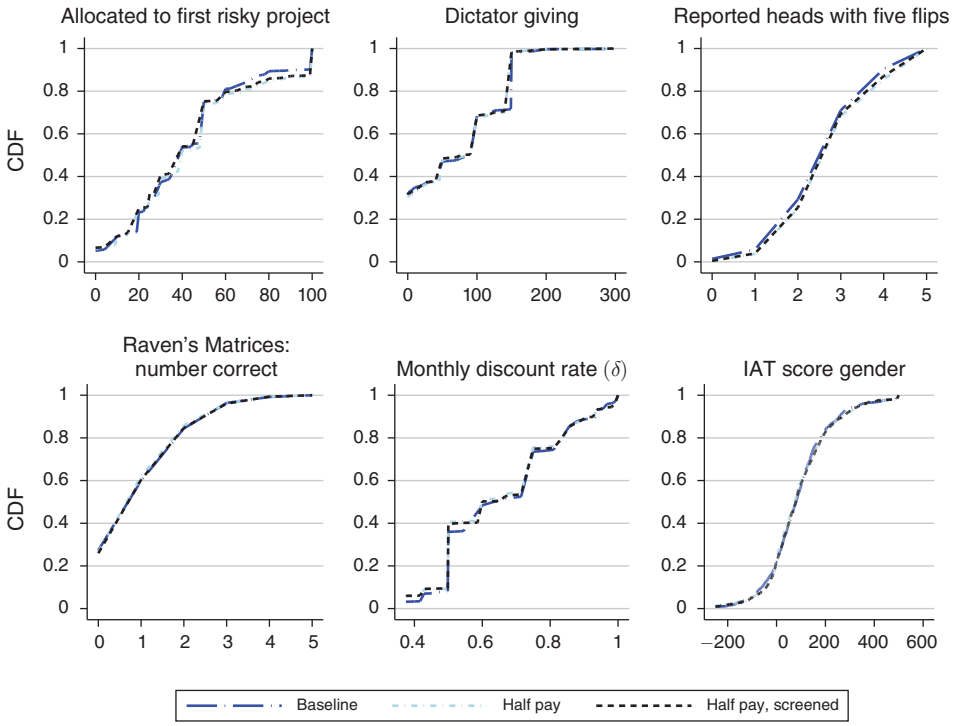

FIGURE 3. DISTRIBUTION OF RESPONSES IN DIFFERENT MTURK SAMPLES

The results from our original MTurk sample and the new sample, with and without low-attention participants, were, quite surprisingly, nearly identical, as shown in Figure 3. Online Appendix B provides more details of our implementation, and an analysis of the additional elicitations from Table 1. The main differences are that the new sample was slightly more dishonest, slightly less competitive, and somewhat less implicitly biased on race. Taken together, these results suggest that responses within participant pools are quite stable, and that the biggest difference in stability may come from learning within participants.

C. *Comparative Statics and Correlations*

Experimental work is often concerned with how varying experimental parameters or participant's attributes changes behavior (*comparative statics*), and how different behaviors and attributes relate to one another (*correlations*). Our experimental elicitations featured one set in which parameters varied: the dictator game. As can be seen from Table 1, doubling the value to the receiver of allocated tokens roughly doubles the contributions of CCS participants and increases them substantially in the UBC sample. In contrast, the average behavior across those elicitations changes very little—if at all—in the representative and MTurk samples. In the parlance of the literature concerned with the reliability of social science findings, the result we find in the CCS sample does not replicate in the MTurk and representative samples (Camerer et al. 2016, Open Science Collaboration 2015).

Table 2—Comparative Statics on Gender

	Rep.	MTurk	CCS	UBC
Panel A. Competition (*percent competing*)				
Male	42	34	41	55
	(2.3)	(2.1)	(2.2)	(5.1)
Female	38	25	21	36
	(2.1)	(2.0)	(2.3)	(4.7)
Difference	3.2	8.9	21	18
	(3.1)	(2.9)	(3.3)	(6.9)
Panel B. First risky project (*out of 100*)				
Male	49	46	67	56
	(1.3)	(1.3)	(1.6)	(3.2)
Female	44	41	48	48
	(1.2)	(1.1)	(1.7)	(2.3)
Difference	4.5	4.7	19	7.8
	(1.8)	(1.7)	(2.4)	(3.9)
Panel C. Dictator, tokens given are halved				
Male	38	21	6.6	14
	(0.94)	(1.0)	(0.79)	(2.1)
Female	39	29	13	19
	(0.78)	(1.0)	(1.2)	(2.1)
Difference	−0.72	−8.2	−6.3	−5.2
	(1.2)	(1.5)	(1.4)	(2.9)
Observations	1,000	995	819	202

Note: Standard errors in parentheses.

Comparative statics with respect to fixed attributes, like gender, can be compared across all three samples as well, as shown in Table 2. The first panel examines the classic finding of Niederle and Vesterlund (2007), that women are less likely to select into competition. This finding replicates well in the CCS, UBC, and MTurk samples, although point estimates of the levels of, and differences between, competitiveness of men and women vary. It does not, however, replicate in the representative sample, where the point estimate of the difference in competition between men and women is statistically indistinguishable from zero.

Different patterns of replication occur across different measures. Panel B of Table 2 examines another common finding in the literature, that women are more risk averse (see Byrnes, Miller, and Schafer 1999 for a review of the relevant experimental work in psychology; see Croson and Gneezy 2009 and Eckel and Grossman 2008 for reviews of related experimental work in economics). This finding replicates across all four samples for the risky project measure, although, again, the point estimate of the difference between the genders is larger in the student samples than in the other two samples. Panel C examines whether women give more in the dictator game, and finds this is the case in both the student and the MTurk samples, but not the representative sample. Women's tendency towards generosity replicates previous findings, see Klinowski (2018) and references therein.

Comparisons of correlations across our samples can be carried out for all the major behaviors and attributes in our data, whether or not those relationships have

been inspected in prior literature. This exercise gives a sense of how likely it is that a finding in one sample will replicate in others, and thus a sense of how likely it is that experimental results will replicate across samples. When there are multiple elicitations of an attribute, we use the first principal component of these elicitations.[35] Our findings are summarized in Figure 4, which displays the sign and significance (at the 10 percent level) of correlations in three samples: first the representative sample, then MTurk, then the CCS.[36]

In Figure 4, a positive and significant correlation is denoted with a "+," a negative and significant correlation is denoted with a "−," and an insignificant correlation is denoted with a "0." When all three samples agree, we use a single symbol in that cell. Thus, the first panel of Table 2 is represented in the appropriate cell in Figure 4 by 0++. The threshold of $p < 0.1$ is chosen conservatively, although many readers may prefer a cutoff of $p < 0.05$. Online Appendix Figures A.2 and A.3 are similar to Figure 4, but use p-value cutoffs of 0.05 and 0.01, respectively.[37]

There is a lot of consistency between the comparative statics observed in the three samples. In only 4 out of the 55 correlations we examine (7 percent) is there a strong disagreement: with a positive and statistically significant correlation in one sample, and a negative and statistically significant correlation in another. In 23 cases (42 percent) there is complete agreement between the three samples. The remaining cases show an agreement in the sign of a correlation (if significant), and disagreement is simply due to one or two of the samples exhibiting statistically insignificant correlations. The most noticeable disagreement occurs in correlations involving the dictator game, lying, confidence, competitiveness, and gender.

Much of the partial agreement is driven by statistically insignificant results in either the MTurk or representative samples (or both), as in Table 2. Indeed, of the 28 partial agreements, 20 feature statistically insignificant correlations in the representative sample, 16 in the MTurk sample, and only 9 in the CCS, which has an 18 percent smaller sample size. As correlations may be attenuated by measurement error, we next turn to an analysis of noise across all three samples.

D. *Noise and Measurement Error*

We use a simple method to ascertain the extent of noise in a sample, building on Gillen, Snowberg, and Yariv (2019). Noise corresponds to anything that may cause differences between responses to duplicate elicitations of the same behavior. In particular, it includes classical measurement error, trembles, inattention, differences

[35] We use principal components both to simplify the presentation and to reduce measurement error (see Gillen, Snowberg, and Yariv 2019). Reducing measurement error lessens attenuation in correlations, and thus lowers the appearance of agreement due to measurement-error-induced statistical insignificance.

[36] We do not include the UBC sample in this comparison due to its smaller size. Many correlations in that sample are consequently insignificant, and could paint a misleading picture of disagreement with our other samples. Our exercise is informed by the fact that statistical comparisons of correlations are inherently difficult. In particular, a random perturbation in one variable affects its correlations with all other variables in the matrix. Most methods model the joint distribution of variables as a multi-variate normal, and test for differences between estimated distributions. For a useful overview, see Diedenhofen and Musch (2015). As many of our variables are clearly not normally distributed, and statistical differences are less important than how differences would manifest themselves in substantive conclusions pertaining to the relationships between variables, we use a different approach.

[37] Those figures lead to similar conclusions. However, as significance restrictions become more demanding, fewer correlations are significant, which mechanically causes the appearance of more agreement.

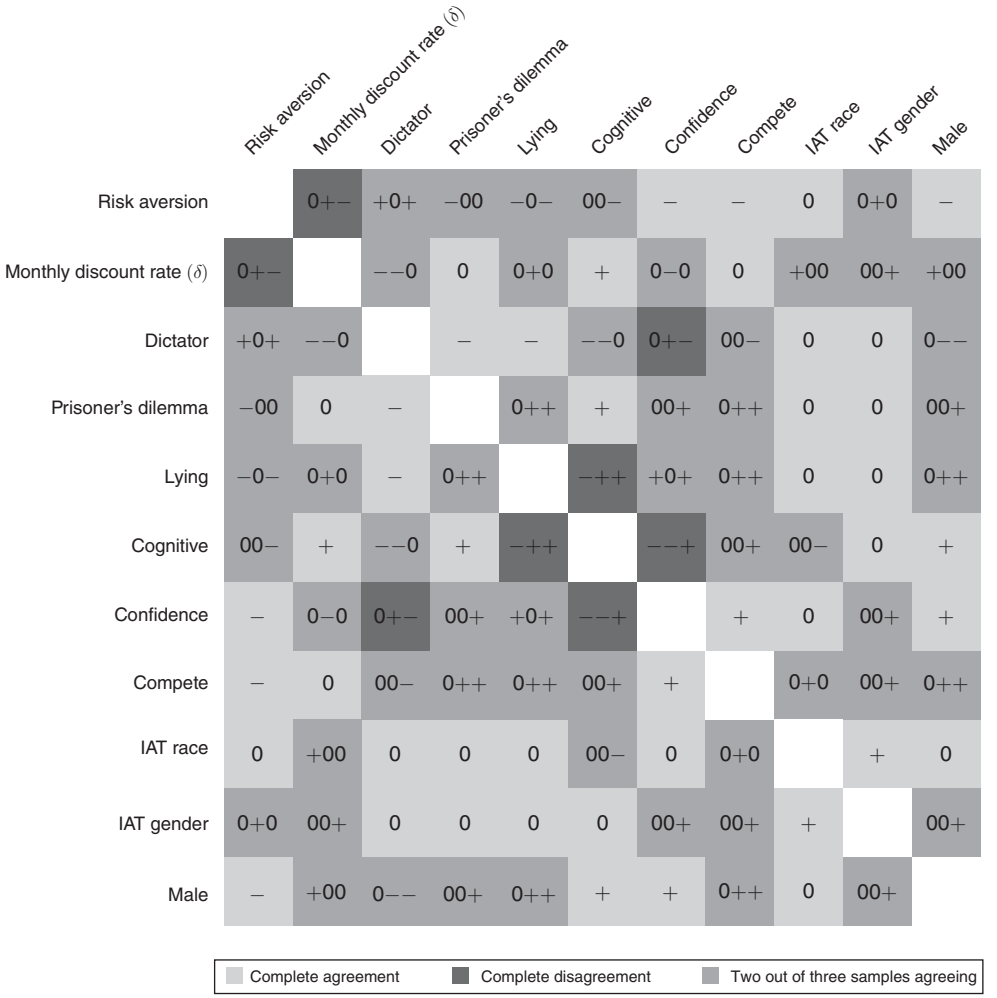

FIGURE 4. CORRELATIONS ACROSS THE REPRESENTATIVE SAMPLE, MTURK, AND CCS

in the grid design between multiple price lists, and other potential causes of inconsistency that we may not be able to fully enumerate, and that we are unable to disentangle. Our method relies on the inclusion of several duplicate elicitations in our survey(s), such as the first and second risky project, the two risky urns (MPLs), and so on. To understand how these are used to assess noise, consider two elicitations of the same underlying parameter X^*. In particular, $X^a = X^* + \nu_X^a$ and $X^b = X^* + \nu_X^b$, with ν_X^a, ν_X^b i.i.d., mean zero, random variables.[38] Then, we have

$$1 - \widehat{\text{corr}}\left[X^a, X^b\right] \to_p 1 - \text{corr}\left[X^a, X^b\right] = \frac{\sigma^2_{\nu_X}}{\sigma^2_{X^*} + \sigma^2_{\nu_X}}. \tag{1}$$

[38] This implies that $E[\nu_X^a \nu_X^b] = 0$ and $\text{var}[\nu_X^a]/\text{var}[X^a] = \text{var}[\nu_X^b]/\text{var}[X^b] := \text{var}[\nu_X]/\text{var}[X]$.

Table 3—Percent of Variation Due to Noise

	Rep.	MTurk	CCS	UBC
Risky projects	59	47	43	53
	(2.9)	(2.7)	(2.9)	(6.2)
Risky urns	35	32	25	31
	(2.4)	(2.3)	(2.3)	(5.1)
Lottery menu	49	33	28[a]	34
	(2.7)	(2.4)	(2.4)	(5.3)
Ambiguous urn	30	31	22[b]	29
	(2.3)	(2.3)	(2.1)	(5.0)
Compound urn	31	26	26[b]	22
	(2.3)	(2.1)	(2.2)	(4.4)
Dictator giving	37	18	15	26
	(2.5)	(1.8)	(1.8)	(4.8)
IAT race	36	46	42	44
	(2.4)	(2.7)	(2.8)	(5.9)
IAT gender	45	46	39	58
	(2.6)	(2.7)	(2.8)	(6.4)
Observations	1,000	995	819	202

Notes: [a] indicates figure is from the fall 2015 CCS (observations = 863), and [b] indicates figure is from the fall 2014 CCS (observations = 893).

Thus, $1 - \widehat{\text{corr}}\,[X^a, X^b]$ is an estimate of the proportion of variation of an elicitation that is due to noise.[39]

Using this relationship, Table 3 shows that the CCS exhibits the lowest noise level of the four samples in all elicitations except for IAT race. These differences are often significant when comparing the CCS and the representative sample. As greater noise leads to greater attenuation of estimated correlations, variations in noise across our samples can help explain the patterns identified in Figure 4. Additionally, reducing incentives by a factor of two, as we did in our second MTurk sample, discussed above, does not noticeably increase noise; see online Appendix Table B.2. While the UBC sample seems to generate noisier observations than the CCS, it is important to note that, even in the CCS, lab responses are far noisier than online survey responses. Indeed, as shown in online Appenix Table A.3, the noise levels in lab responses at Caltech are comparable to those seen in the UBC sample.

The differences in the amount of noise across our samples raise caution on certain conclusions derived from comparing correlations using student data and data from other populations. Noise, or measurement error, could cause significant correlations in student samples not to replicate in other samples. In view of recent concerns about the lack of reproducibility of experimental results (see Ioannidis 2005, and references that follow), our observations emphasize the importance of techniques to deal with measurement error in experiments, especially when using non-student samples (Gillen, Snowberg, and Yariv 2019). Furthermore, our analysis suggests a

[39] In principle, estimated correlations could be negative. This would generally be evidence that either the elicitations are measuring distinct attributes, with differences not due to noise, or that one of the elicitations has an inverted scale.

natural cost-benefit tradeoff: while student-based studies often entail higher costs, they imply lower noise, at least when administered using an online survey.

IV. Selection into the Lab

The prior section indicated that representative and student populations yield similar correlations between elicitations, despite differences in levels of elicited behaviors and levels of noise. This comparison was done using a survey that covered nearly the entire population of Caltech students. In contrast, lab experiments include nonrandom, and possibly nonrepresentative, samples of university students: those who *select* to go to the lab. In this section we ask whether selection into the lab results in nonrepresentative behaviors in that population. Broadly speaking, we find very few differences between the population that goes to the lab and the overall university population.

In principle, participants who select into experiments may have different attributes than those who do not. Such differences would reduce our ability to extrapolate from lab experiments, even to the population of students from which participants are drawn. This is especially relevant for particular classes of experiments. For example, generosity is commonly observed in the lab (see, for example, Roth 1995 for references). However, if individuals who contribute to others' research by showing up to the lab are more generous than the overall population, these conclusions might lack external validity (see Levitt and List 2007b).

The CCS offers a unique opportunity to examine selection into lab experiments. Given the high response rate, the surveys provide an array of attributes of the underlying population of potential participants. Data from the Caltech Social Science Experimental Laboratory (SSEL) supply the full record of participation in lab experiments for each student. We can therefore identify lab participants in the CCS data, and compare the patterns of their elicited behaviors to those of the underlying population of students.

We examine two ways of characterizing the population that goes to the lab. The first simply compares responses, on the CCS, of the population that has participated in at least one experiment in SSEL with the entire population in the CCS. The second compares responses *weighted by participation* with the entire CCS population. For this second comparison, we weight responses on the CCS of those who go to the lab by their *lab experience*—that is, the average number of times per year a CCS participant went to the lab. Behavior measures weighted by participation mimic behavior (on the CCS) of the average population one would see across all lab experiments. The averages for each population are displayed in the first three columns of Table 4, while the final two columns compare the two lab-going populations to the overall population that participated in the CCS.

We see little behavioral difference between the population that goes to the lab and the overall population. Indeed, the only statistically significant difference in behavior is in the amount allocated in the first risky project. In addition, the subsample that goes to the lab has a significantly greater proportion of females.

The difference between the average lab population (the lab population weighted by lab experience) and the overall population is more significant, but small relative to the differences between the samples examined in the previous section. The aver-

Table 4—Lab Participants Are Not Substantially Different from the Overall Population

	Samples			Differences	
	Everyone	Participant	Weighted participant		
	(E)	(P)	(WP)	E − P	E − WP
First risky project (out of 100)	59	55	52	4.8	7.3
	(1.2)	(1.8)	(1.8)	(2.2)	(2.2)
Second risky project (out of 200)	143	139	132	4.2	11
	(2.1)	(3.2)	(3.3)	(3.8)	(3.9)
First risky urn (20 balls)	59	58	58	0.82	1.0
	(0.52)	(0.77)	(0.74)	(0.93)	(0.90)
Second risky urn (30 balls)	86	86	85	0.06	0.89
	(0.73)	(1.1)	(0.99)	(1.3)	(1.2)
Qualitative risk aversion	5.8	5.7	5.7	0.05	0.09
	(0.08)	(0.12)	(0.12)	(0.15)	(0.15)
Monthly discount rate (δ)	0.77	0.78	0.77	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
First dictator game (given out of 100)	14	12	9.2	2.2	4.7
	(0.84)	(1.1)	(1.0)	(1.4)	(1.3)
Second dictator game (given out of 300)	38	32	26	6.1	12
	(2.4)	(3.2)	(2.8)	(3.9)	(3.7)
Dictator, tokens given are doubled	26	26	26	−0.00	−0.10
	(1.2)	(1.8)	(1.8)	(2.2)	(2.2)
Dictator, tokens given are halved	9.0	7.8	6.0	1.2	2.9
	(0.68)	(0.94)	(0.84)	(1.2)	(1.1)
Prisoner's dilemma (percent dominant strategy)	68	67	69	0.68	−1.4
	(1.5)	(2.3)	(2.3)	(2.8)	(2.7)
Reported heads (out of 5)	3.3	3.4	3.5	−0.11	−0.18
	(0.04)	(0.06)	(0.06)	(0.08)	(0.08)
Reported switches (out of 9)	5.5	5.5	5.8	−0.01	−0.34
	(0.07)	(0.11)	(0.11)	(0.13)	(0.13)
Raven's matrices (out of 5)	1.8	1.8	1.8	−0.01	−0.02
	(0.04)	(0.07)	(0.07)	(0.08)	(0.08)
CRT (out of 3)	1.7	1.7	1.7	−0.03	−0.07
	(0.04)	(0.06)	(0.06)	(0.07)	(0.07)
Confidence in guesses	3.1	3.1	3.1	0.09	0.06
	(0.03)	(0.05)	(0.05)	(0.06)	(0.06)
Competition (percent competing)	33	34	33	−0.26	0.16
	(1.7)	(2.5)	(2.5)	(3.0)	(3.0)
IAT race	81	87	81	−6.0	0.32
	(5.6)	(8.5)	(8.5)	(10)	(10)
IAT gender	95	85	103	9.8	−7.7
	(5.9)	(8.5)	(9.5)	(10)	(11)
Percent male	62	55	57	6.2	5.2
	(1.7)	(2.7)	(2.7)	(3.2)	(3.2)
Observations	819	350	350	—	—

Note: Standard errors in parentheses.

age lab participant is more risk averse, more willing to lie, and less generous than the overall university population. The largest differences, in the second risky project and dictator giving, are less than one-fourth and one-sixth, respectively, of the corresponding differences when comparing the representative sample and the CCS.[40]

The average differences we observe are again indicative of first-order stochastic dominance relations in the underlying distributions, as shown in Figure 5. The panels of this figure display the cumulative distribution functions corresponding to the same selected set of elicitations depicted in Figure 1 for the overall and lab-going subpopulations of the CCS. These images echo the message emerging from Table 4. Lab participants are similar to the underlying population, with some small differences for certain elicitations when weighting the set of participants by lab experience.[41]

Overall, lab participants are more risk averse, less generous, and more willing to lie on the spring 2015 CCS. The previous section documented that the CCS sample is less risk averse than the representative or MTurk samples. As lab participants are *more* risk averse than their underlying population, risk behaviors for the lab-going population are slightly closer to the representative and MTurk samples. Nevertheless, the Caltech lab participants are still significantly and substantially less risk averse than participants in the other two samples. Generosity, as reflected by dictator giving, displays the opposite pattern: lab participants are even less generous than the underlying student population, which increases the difference with the other samples. This is particularly interesting in view of a frequent concern that generosity in experiments is an artifact of behavior in the lab due to selection of participants who are willing to spend time helping researchers and therefore more likely to be generous in general. While the differences in reported heads or reported switches are small, lab participants are, if anything, more likely to lie than their underlying population, and certainly relative to the other two samples. This implies that conclusions about reluctance to lie in experiments (see Gneezy 2005, Erat and Gneezy 2012) are not a consequence of selection into the lab. This is a particularly important counterfactual to the literature that suggests lab participants choose actions that make them look more "moral" or "ethical" (see, for example, Levitt and List 2007a, b).

Finally, we see largely similar comparative statics and correlation patterns across subsamples, although smaller subsamples produce less statistically significant results. Table 4 shows that comparative statics on dictator giving as the conversion rate changes replicate across the main sample and the two lab-going subsamples. Two of the three comparative statics examined in Table 2—between gender and competition or risk aversion—also replicate across all three samples, as shown in Figure 6. This figure is an analog of Figure 4, in which each entry corresponds, from left to right, to the sign and significance of the correlation in the CCS, the subset of CCS participants who participated in at least one SSEL experiment, and that same subset weighted by lab experience. Once again, a positive and significant correlation

[40] Comparing responses of those who participate in experiments more than the median number per year to those who participate less than the median number produces no statistically significant differences, see online Appendix Table A.6.

[41] The cumulative distributions for the remaining set of elicitations is in online Appendix Figure A.4.

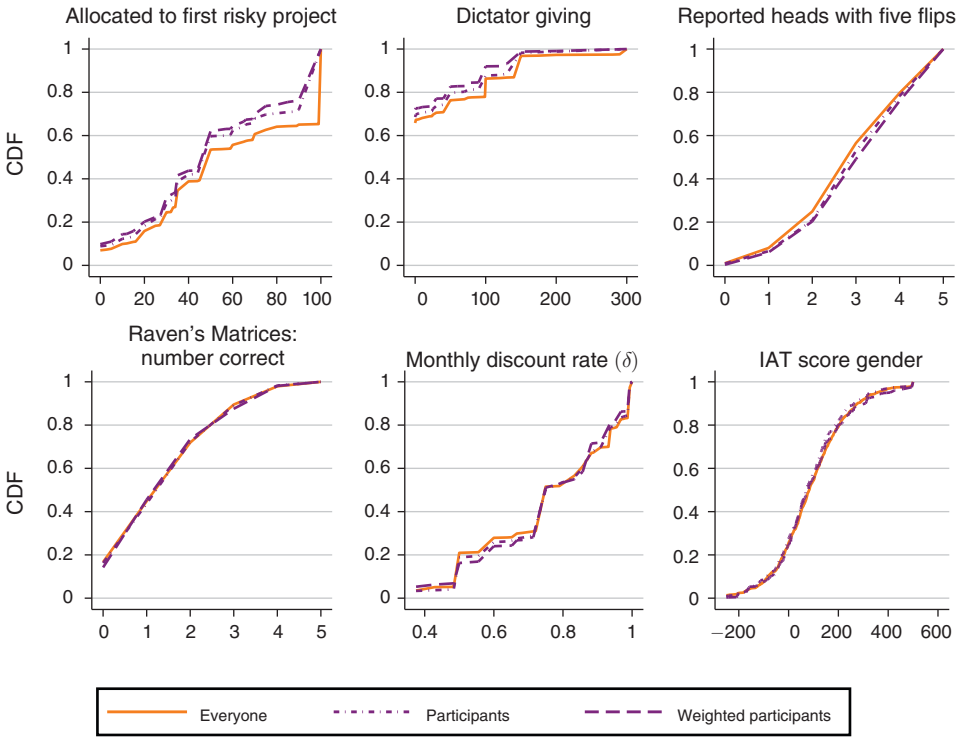

FIGURE 5. DISTRIBUTION OF RESPONSES IN THE CCS AND LAB-GOING SUBSAMPLES

(at the 10 percent level) is denoted with a "+," a negative and significant correlation is denoted with a "−," and an insignificant correlation is denoted with a "0." When all three samples agree, we use a single symbol in that cell. The final comparative static we examine in Table 2—dictator giving and gender—is statistically insignificant in the two subsamples, similar to most other cases of partial agreement in Figure 6.

There is substantial agreement between the signs of correlations in the full CCS sample and the subset of lab participants, whether or not they are weighted by lab experience. None of the cells exhibit complete disagreement: a positive and significant correlation in one sample, and a negative and significant correlation in another. In 37 of the 55 (67 percent) correlations, there is full agreement. In 40 of the 55 (73 percent), correlations have the same sign when considering the overall CCS population and the subset of lab participants weighted by lab experience: the subsample with the largest differences in Table 4 and Figure 5. In only four of the cells (7 percent) in which there is some disagreement is the correlation within the overall CCS population insignificant. Thus, most disagreement is due to one of the smaller samples exhibiting an insignificant correlation. This is unsurprising; smaller samples have larger standard errors, and thus lower levels of significance.[42] In only

[42] Figures for 5 percent and 1 percent significance levels are shown in the online Appendix, see Figures A.5 and A.6. Those figures lead to similar conclusions. As before, when significance restrictions become more demanding, fewer correlations are significant, which mechanically causes the appearance of more agreement.

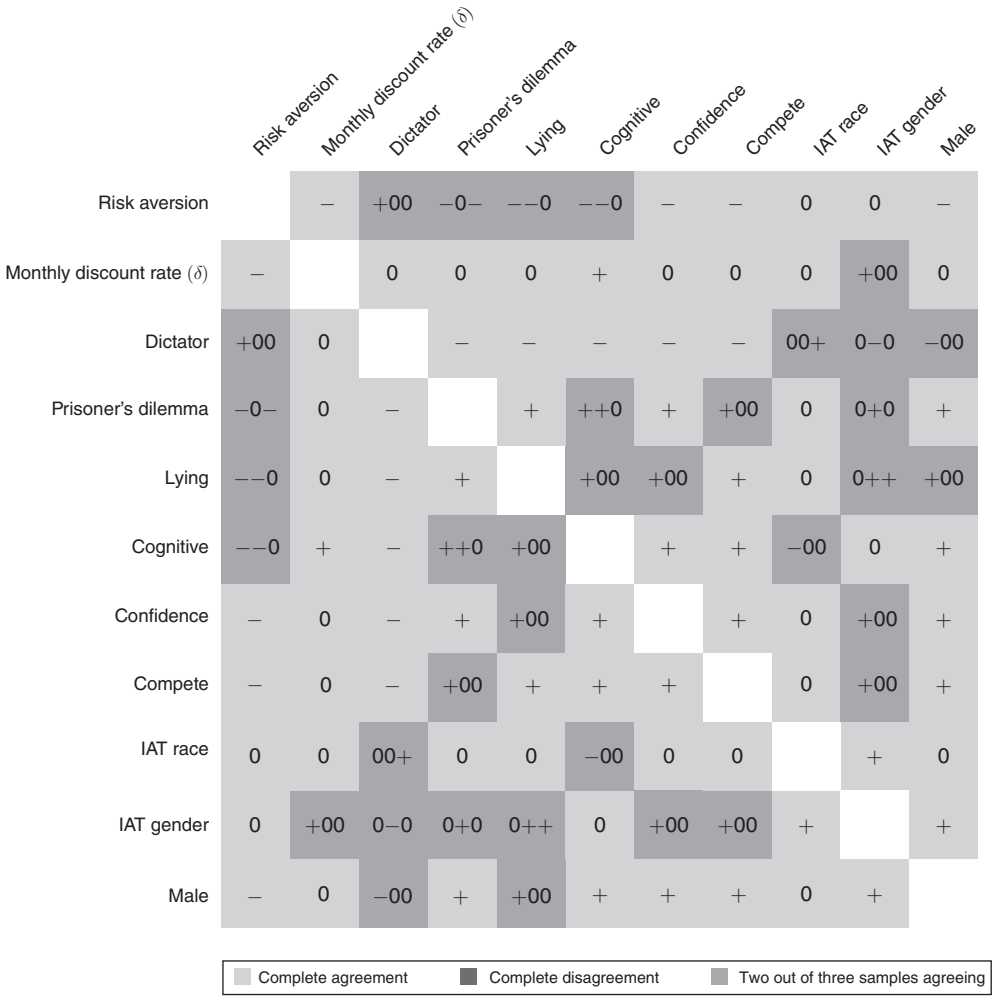

Figure 6. Correlations across Everyone in the CCS, Participants, and Weighted Participants

one of the 18 cells in which there is some disagreement, corresponding to the correlation between gender and lying, is there a statistically significant difference across samples. In that cell, the only statistically significant difference (at the 10 percent level) is between the correlation found in the overall CCS population and that found when participants are weighted by lab experience.[43]

In summary, we see some selection effects in lab participation, though the lab-going subsample is nonrepresentative in terms of only a few behaviors. In fact, several concerns voiced in the literature about selection into the lab driving experimental results—for example, in the context of social preferences—are not borne out

[43] To gauge how likely this is to occur by chance, we can randomly assign 350 Caltech students to constitute the sample that goes to the lab with rates given by the empirical distribution, and compute Figure 6 under this simulated draw. With 100 simulations, 33 of the 55 cells, on average, show complete agreement. Thus, the rate of agreement in Figure 6 is higher than expected by chance, albeit, not by much: it is at the seventy-fifth percentile of our simulation results.

by our data. Finally, correlations between attributes appear remarkably similar for lab participants and the population from which they are drawn.

It would be fairly easy to control for the selection effects we identify. Only a small set of attributes (the first risky project and reported heads) are jointly statistically significant, while others are not. That is, if one controlled for only three variables in Table 4, the sample would be statistically balanced on the other 17. A particular implication of this fact is that the difference in the gender composition of lab participants does not explain much of the average differences we observe.

V. Behavior in the Lab

Although the lab-going subpopulation exhibits similar behaviors to the overall university sample, it is still possible that being in the lab, or being observed by experimenters, would result in changes in behavior. Indeed, the Hawthorne or "observer" effect, reviewed in Section I, has been a topic of discussion for more than 80 years. In this final section of analysis, we show that, to the extent that an observer effect plays a role, its impacts are not sensitive to the level of monitoring of participants. We see few differences between behavior in and out of the lab, and what differences do exist are in line with learning documented in Section IIIB.

In order to compare responses in the lab to responses in the online CCS, we conducted a sequence of experiments at SSEL in the summer of 2015. We were present in the lab for all sessions. We invited students from the cohorts covered by the survey to participate.[44] In total, 97 students participated. Lab participants retook the CCS survey from the spring of 2015, which allows us to compare whether the lab environment itself changes participants' responses. Of the 97 students participating in our lab sessions, 96 (99 percent) also participated in the spring 2015 CCS survey. On average, participants spent a comparable amount of time filling out the survey online (35 minutes) and in the lab (31 minutes).

Average responses in the lab and on the survey are very similar, as illustrated by Table 5. The first column shows the average responses of the full population to the spring 2015 CCS, reproduced from Table 1. By comparing these with the CCS responses of the 96 who also attended a lab session (the second column), it can be seen that the only major difference is the gender of the lab participants, similar to what we found in the prior section.

The third column of Table 5 displays the same behaviors elicited from the same population, but this time in the lab. As shown in the fifth column, the differences between behavioral elicitations on the survey and in the lab are small and statistically insignificant. Three elicitations exhibit statistically significant differences. In the first risky urn, students are more risk averse in the lab, though the difference is small in magnitude (less than 4 percent of the maximum allowed willingness to pay). In the two cognitive tasks (Raven's matrices and CRT), lab participants significantly outperform survey participants. At face value, this suggests that lab participants might be somewhat more attentive or more willing to exert effort than

[44] Experimental sessions were separated by a few months from the spring and fall installments of the survey. The name of the experiment was intentionally not indicative of its content, so participants were not aware they would be completing the CCS survey in the lab when signing up.

Table 5—Differences in Choices: Lab versus Survey

	Samples			Differences	
		Lab sample			
	CCS	Survey (S)	Lab (L)	CCS − S	S − L
First risky project (out of 100)	59	54	54	6.1	0.50
	(1.2)	(3.2)	(3.3)	(3.8)	(4.6)
Second risky project (out of 200)	143	134	138	10	−3.9
	(2.1)	(5.8)	(5.8)	(6.4)	(8.2)
First risky urn (20 balls)	59	60	56	−0.51	4.0
	(0.52)	(1.5)	(1.3)	(1.6)	(2.0)
Second risky urn (30 balls)	86	87	84	−0.63	3.0
	(0.73)	(2.0)	(1.7)	(2.3)	(2.6)
Qualitative risk aversion	5.8	5.4	5.4	0.39	−0.01
	(0.08)	(0.21)	(0.21)	(0.24)	(0.29)
Monthly discount rate (δ)	0.77	0.78	0.78	−0.01	−0.01
	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)
First dictator game (given out of 100)	14	13	9.8	1.1	3.1
	(0.84)	(2.3)	(2.0)	(2.6)	(3.0)
Second dictator game (given out of 300)	38	35	24	3.1	11
	(2.4)	(6.5)	(6.0)	(7.4)	(8.8)
Dictator, tokens given are doubled	26	29	29	−2.7	−0.42
	(1.2)	(3.4)	(3.7)	(3.7)	(5.1)
Dictator, tokens given are halved	9.0	7.1	5.3	2.0	1.8
	(0.68)	(1.7)	(1.5)	(2.1)	(2.2)
Prisoner's dilemma (percent dominant strategy)	68	68	70	−0.46	−1.6
	(1.5)	(4.2)	(4.5)	(4.7)	(6.1)
Reported heads (out of 5)	3.3	3.3	3.3	−0.04	0.07
	(0.04)	(0.12)	(0.11)	(0.13)	(0.16)
Reported switches (out of 9)	5.5	5.2	5.2	0.27	0.04
	(0.07)	(0.20)	(0.17)	(0.22)	(0.26)
Raven's matrices (out of 5)	1.8	1.9	2.5	−0.11	−0.58
	(0.04)	(0.13)	(0.13)	(0.14)	(0.19)
CRT (out of 3)	1.7	1.6	2.1	0.06	−0.48
	(0.04)	(0.11)	(0.11)	(0.12)	(0.16)
Confidence in guesses	3.1	2.9	3.0	0.32	−0.11
	(0.03)	(0.08)	(0.09)	(0.10)	(0.12)
Competition (percent competing)	34	32	29	1.3	3.1
	(1.7)	(4.8)	(4.7)	(5.1)	(6.7)
IAT race	81	109	84	−31	25
	(5.6)	(18)	(13)	(17)	(22)
IAT gender	95	99	65	−4.6	34
	(5.9)	(15)	(15)	(18)	(21)
Percent male	62	45	45	19	0
	(1.7)	(5.1)	(5.1)	(5.3)	(−)
Observations	819	96	96	—	—

Note: Standard errors in parentheses.

they are outside of the lab. However, it is also consistent with what we see for CCS participants in general when they take an additional survey, as discussed in Section IIIB. Despite these differences, when controlling for only one of these variables—the number of correct Raven's matrices—the samples are balanced on the other 19.

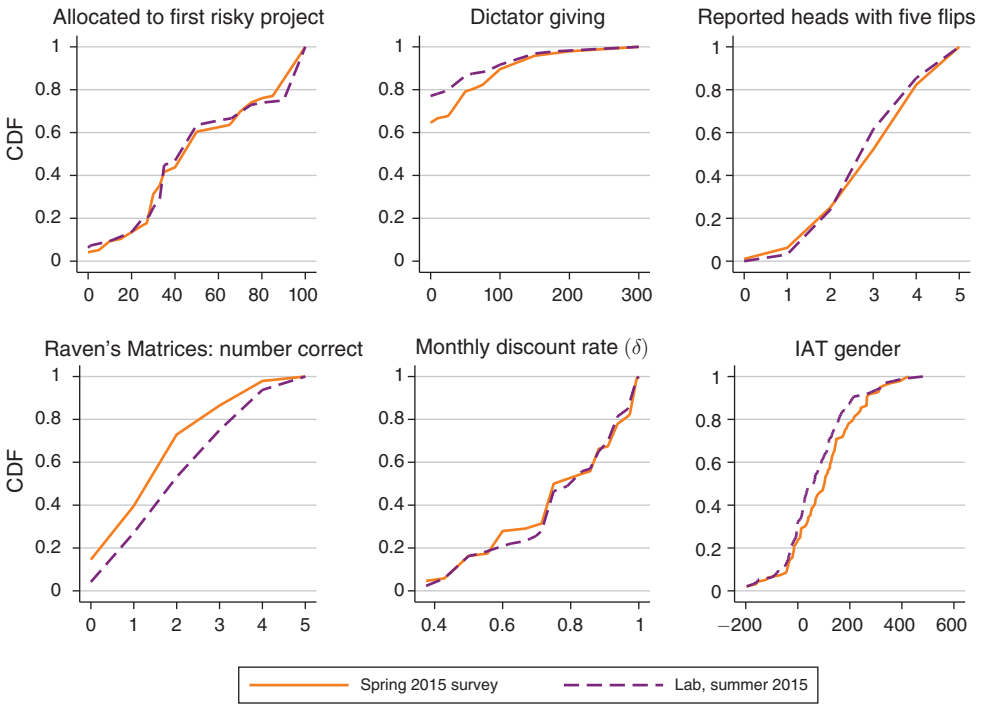

FIGURE 7. DISTRIBUTION OF RESPONSES IN THE SPRING 2015 SURVEY VERSUS THE LAB (OBSERVATIONS = 96)

A careful inspection of our results for the cognitive tasks does not allow us to reject either a learning or a lab-based performance effect. Recall that for participants who took the spring and fall 2015 survey, and *did not* participate in our lab experiment in the summer of 2015 (500 people), the mean CRT was 1.67 in the spring and 1.95 in the fall. Similarly, the mean number of correct Raven's matrices was 1.85 in the spring, and 1.91 in the fall. The increase for these participants is not as substantial as for those who participated in the lab experiment, indicating a lab-based effect. On the other hand, 90 of the 96 lab participants took the fall 2015 survey, and their average scores were 2.19 for the CRT, and 2.79 for the five Raven's matrices. This persistent improvement in performance is consistent with a learning effect.

Also consistent with some learning are the distributions of dictator giving shown in Figure 7, which depicts the analogous cumulative distributions to Figures 1 and 5 across the lab and survey environment. Although there is no statistically significant difference in means for this elicitation, the distribution in the lab first-order stochastically dominates that of the survey. This is consistent with the learning effect in the dictator game illustrated in Figure 2. For all other distributions, except for those of the number of correct Raven's matrices, the distributions in the lab and on the survey are nearly identical.

Comparative statics and correlations between measures are, almost uniformly, statistically indistinguishable between the lab and the survey environment. However, as our sample of lab participants is relatively small, many correlations are insignificant simply due to high standard errors. As shown in online Appendix Figure A.7, the analog of Figures 4 and 6, 45 of the 55 correlations we inspected in our SSEL

data and in our CCS data—restricted to participants in our lab experiment—were insignificant.

In summary, we see little difference between behavior in the lab and outside. The overall similarity between results generated through lab experiments and online surveys is in line with previous results suggesting a lack of observer effect in particular settings like the dictator game or public goods games (see Laury, Walker, and Williams 1995; and Bolton, Katok, and Zwick 1998).[45]

VI. Discussion

In this paper, we leverage a large-scale survey run on multiple populations to answer three questions. First, are university students behaviorally different than representative populations or convenience samples, specifically Amazon's Mechanical Turk (MTurk)? Second, are those students who choose to participate in lab experiments different from the general student population? Third, do students change their behavior when they are in the lab?

Comparative statics and correlations between behaviors are similar across the student, representative, and MTurk samples, although the distributions of individual behaviors are quite different. Differences in correlations can largely be accounted for by statistical insignificance in the representative and MTurk samples, driven by greater noise. We see some evidence for differences in observable behaviors between the general student population and self-selected lab participants, though these differences are confined to a minimal set of behaviors that are easy to elicit and control for. We see no evidence of observer effects—differences in behavior when completing tasks in the lab while being observed by experimenters.

There are other potential advantages of lab experiments that our study does not speak to directly. Indeed, the lab is often said to enable more intricate experimental designs, by allowing experimenters to provide detailed instructions and monitor participants' attention.

We caution that we cannot address the concern that different framings of problems, or different backgrounds or experiences of participants, would not affect behavior. We expect they would. For example, a participant in an online or lab auction may behave differently from a seasoned bidder in FCC auctions.[46] In practice, it would be quite unusual for a practitioner to take evidence from a student-based lab sample directly to public policy. Instead, a scholar might design a mechanism on the basis of lab insights. The resulting policy would then be field tested to ensure insights carry over, and to allow fine-tuning of the policy's details. It is precisely this sort of protocol that our study supports.

In general, concerns about external validity focus on how findings extend to different people, different environments, and/or different choices. Our study has much to say about external validity concerns due to different participant populations and provides insights on the effects of particular environments (incentivized survey or

[45] These results are also in line with Anderhub, Müller, and Schmidt (2001), who compared behavior in the lab and online, albeit with different participants in each, and in only one game.

[46] See, however, Fréchette (2015) for a comparison of several experiments run on students and professionals. By and large, he reports similar results across the two types of participants.

lab). Still, more could be learned by cataloging the differences in fundamental attributes between other samples and platforms. We hope the methodology we introduce opens the door to future data-driven studies of these facets of external validity.

REFERENCES

Alm, James, Kim M. Bloomquist, and Michael McKee. 2015. "On the External Validity of Laboratory Tax Compliance Experiments." *Economic Inquiry* 53 (2): 1170–86.

Anderhub, Vital, Rudolf. 2001. "Design and Evaluation of an Economic Experiment via the Internet." *Journal of Economic Behavior and Organization* 46 (2): 227–47.

Arechar, Antonio A., Simon Gächter, and Lucas Molleman. 2018. "Conducting Interactive Experiments Online." *Experimental Economics* 21 (1): 99–131.

Arechar, Antonio A., Gordon Kraft-Todd, and David G. Rand. 2017. "Turking Overtime: How Participant Characteristics and Behavior Vary Over Time and Day on Amazon Mechanical Turk." *Journal of the Economic Science Association* 3 (1): 1–11.

Armantier, Olivier, and Amadou Boly. 2013. "Comparing Corruption in the Laboratory and in the Field in Burkina Faso and in Canada." *Economic Journal* 123 (573): 1168–87.

Belot, Michèle, Raymond Duch, and Luis Miller. 2015. "Who Should Be Called to the Lab? A Comprehensive Comparison of Students and Non-Students in Classic Experimental Games." *Journal of Economic Behavior & Organization* 113: 26–33.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (3): 351–68.

Bolton, Gary E., Elena Katok, and Rami Zwick. 1998. "Dictator Game Giving: Rules of Fairness versus Acts of Kindness." *International Journal of Game Theory* 27 (2): 269–99.

Byrnes, James P., David C. Miller, and William D. Schafer. 1999. "Gender Differences in Risk Taking: A Meta-analysis." *Psychological Bulletin* 125 (3): 367–83.

Camerer, Colin F. 2015. "The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter, 249–295. Oxford, UK: Oxford University Press.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Cappelen, Alexander W., Knut Nygaard, Erik Ø. Sorensen, and Bertil Tungodden. 2015. "Social Preferences in the Lab: A Comparison of Students and a Representative Population." *Scandinavian Journal of Economics* 117 (4): 1306–26.

Charness, Gary, Uri Gneezy, and Alex Imas. 2013. "Experimental Methods: Eliciting Risk Preferences." *Journal of Economic Behavior and Organization* 87: 43–51.

Cleave, Blair L., Nikos Nikiforakis, and Robert Slonim. 2013. "Is There Selection Bias in Laboratory Experiments? The Case of Social and Risk Preferences." *Experimental Economics* 16 (3): 372–82.

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7 (3): 613–28.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2): 448–74.

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69.

Diedenhofen, Birk, and Jochen Musch. 2015. "cocor: A Comprehensive Solution for the Statistical Comparison of Correlations." *PLOS ONE* 10 (4): e0121945.

Difallah, Djellel, Elena Filatova, Panos Ipeirotis. 2018. "Demographics and Dynamics of Mechanical Turk." *WSDM '18: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining,* 135-43.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2010. "Are Risk Aversion and Impatience Related to Cognitive Ability?" *American Economic Review* 100 (3): 1238–60.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2018. "On the Relationship Between Cognitive Ability and Risk Preference." *Journal of Economic Perspectives* 32 (2): 115–34.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–50.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri. 2020. "Monopsony in Online Labor Markets." *American Economic Reveiw: Insights* 2 (1): 33–46.

Eckel, Catherine C., and Philip J. Grossman. 2008. "Men, Women and Risk Aversion: Experimental Evidence." In *Handbook of Experimental Economics Results*. Vol. 1, edited by Charles R. Plott, and Vernon L. Smith, 1061–73. Amsterdam: North-Holland.

Erat, Sanjiv, and Uri Gneezy. 2012. "White Lies." *Management Science* 58 (4): 723–33.

Exadaktylos, Filippos, Antonio M. Espín, and Pablo Brañas-Garza. 2013. "Experimental Subjects are not Different." *Scientific Reports* 3 (1).

Falk, Armin, and James J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326 (5952): 535–38.

Falk, Armin, Stephan Meier, and Christian Zehnder. 2013. "Do Lab Experiments Misrepresent Social Preferences? The Case of Self-Selected Student Samples." *Journal of the European Economic Association* 11 (4): 839–52.

Fosgaard, Toke R. 2018. "Cooperation Stability—A Representative Sample in the Lab." Unpublished.

Fréchette, Guillaume R. 2015. "Laboratory Experiments: Professionals versus Students." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter, 360–90. Oxford: Oxford University Press.

Fréchette, Guillaume R. 2016. "Experimental Economics Across Subject Populations." In *The Handbook of Experimental Economics*. Vol. 2, edited by John H. Kagel and Alvin E. Roth, 435–80. Princeton, NJ: Princeton University Press.

Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19 (4): 25–42.

Gächter, Simon. 2010. "(Dis)advantages of Student Subjects: What is Your Research Question?" *Behavioral and Brain Sciences* 33 (2-3): 92–93.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy* 127 (4): 1826–63.

Gillespie, Richard. 1993. *Manufacturing Knowledge: A History of the Hawthorne Experiments*. Cambridge, UK: Cambridge University Press.

Gneezy, Uri. 2005. "Deception: The Role of Consequences." *American Economic Review* 95 (1): 384–94.

Gneezy, Uri, and Jan Potters. 1997. "An Experiment on Risk Taking and Evaluation Periods." *Quarterly Journal of Economics* 112 (2): 631–45.

Goodman, Joseph K., Cynthia E. Cryder, and Amar Cheema. 2013. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26 (3): 213–24.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74 (6): 1464–80.

Greenwald, Anthony G., Brian A. Nosek, and Mahzarin R. Banaji. 2003. "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm." *Journal of Personality and Social Psychology* 85 (2): 197–216.

Guala, Francesco, and Luigi Mittone. 2005. "Experiments in Economics: External Validity and the Robustness of Phenomena." *Journal of Economic Methodology* 12 (4): 495–515.

Harrison, Glenn W., Morten I. Lau, and E. Elisabet Rutström. 2009. "Risk Attitudes, Randomization to Treatment, and Self-Selection into Experiments." *Journal of Economic Behavior and Organization* 70 (3): 498–507.

Hauser, David, Gabriel Paolacci, and Jesse Chandler. 2019. "Common Concerns with MTurk as a Participant Pool: Evidence and Solutions." In *Handbook of Research Methods in Consumer Psychology*, edited by Kardes Frank R. Paul M. Herr, and Norbert Schwarz, 318–45. New York: Routledge.

Herbst, Daniel, and Alexandre Mas. 2015. "Peer Effects on Worker Output in the Laboratory Generalize to the Field." *Science* 350 (6260): 545–49.

Hergueux, Jerome, and Nicolas Jacquemet. 2015. "Social Preferences in the Online Laboratory: A Randomized Experiment." *Experimental Economics* 18 (2): 251–83.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior* 7 (3): 346–80.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14 (3): 399–425.

Huff, Connor, and Dustin Tingley. 2015. "'Who are these people?' Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research & Politics* 2 (3): 1–12.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8): e124.

Jones, Stephen R. G. 1992. "Was there a Hawthorne Effect?" *American Journal of Sociology* 98 (2): 451–68.

Kessler, Judd B., and Lise Vesterlund. 2015. "The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter. Oxford, UK: Oxford University Press.

Klinowski, David. 2018. "Gender Differences in Giving in the Dictator Game: The Role of Reluctant Altruism." *Journal of the Economic Science Association* 4 (5): 110–22.

Laury, Susan K., James M. Walker, and Arlington W. Williams. 1995. "Anonymity and the Voluntary Provision of Public Goods." *Journal of Economic Behavior and Organization* 27 (3): 365–80.

Levitt, Steven D., and John A. List. 2007a. "Viewpoint: On the Generalizability of Lab Behaviour to the Field." *Canadian Journal of Economics* 40 (2): 347–70.

Levitt, Steven D., and John A. List. 2007b. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21 (2): 153–74

Levitt, Steven D., and John A. List. 2008. "Homo Economicus Evolves." *Science* 319 (5865): 909–10.

Levitt, Steven D., and John A. List. 2011. "Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments." *American Economic Journal: Applied Economics* 3 (1): 224–38.

Mayo, Elton. 1933. *The Human Problems of an Industrial Civilization*. New York: Routledge.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122 (3): 1067–101.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): 943.

Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and their Implications." *American Psychologist* 17 (11): 776–83.

Ortoleva, Pietro, and Erik Snowberg. 2015. "Overconfidence in Political Behavior." *American Economic Review* 105 (2): 504–35.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5 (5): 411–19.

Raven, James C. 1936. "Mental Tests Used in Genetic Studies: The Performance of Related Individuals on Tests Mainly Educative and Mainly Reproductive." PhD diss. University of London.

Roth, Alvin E. 1995. "Bargaining Experiments." In *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth, 253–348. Princeton, NJ: Princeton University Press.

Schram, Arthur. 2005. "Artificiality: The Tension between Internal and External Validity in Economic Experiments." *Journal of Economic Methodology* 12 (2): 225–37.

Snowberg, Erik, and Leeat Yariv. 2021. "Replication Data for "Testing the Waters: Behavior across Participant Pools." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E120227V1.

This article has been cited by:

1. Stefano DellaVigna, Devin Pope. 2022. Stability of Experimental Results: Forecasts and Evidence. *American Economic Journal: Microeconomics* **14**:3, 889-925. [Abstract] [View PDF article] [PDF with links]