Testing the Waters: Behavior across Participant Pools*

Erik Snowberg
University of British Columbia, CESifo, and NBER
snowberg@mail.ubc.ca
hss.caltech.edu/∼snowberg

Leeat Yariv
Princeton University and CEPR
lyariv@princeton.edu
www.princeton.edu/yariv

July 15, 2018

Abstract

We leverage a large-scale incentivized survey eliciting behaviors from (almost) an entire university student population, a representative sample of the U.S. 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 largely similar across samples. Furthermore, non-student samples exhibit higher measurement error. Adding historical lab participation data, we find a small set of attributes over which lab participants differ from non-lab participants. Using an additional set of lab experiments, we see no evidence of observer effects.

JEL Classifications: B41, C80, C90
Keywords: Lab Selection, External Validity, Experiments

*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 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.

1 Introduction

Lab experiments have been used to amass large amounts of data on human behavior in the face of economic incentives. 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. Given the paucity of data that could alleviate such concerns, they are 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. In particular, we provide evidence relevant to 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 population? Third, do students change their behavior in the lab?

We show that correlations between elicited behaviors are similar across our student, representative, and MTurk samples, although the elicited behaviors themselves are quite different. Differences in correlations can largely be explained by statistical insignificance in representative and MTurk samples, driven by higher measurement error. 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 attributes that are easy to elicit and control for. We see no evidence of participants behaving differently inside

[1] In specific settings, there is some useful work that examines these issues. See the literature review below for details.

the lab than outside of it.

Together, our results suggest that experiments utilizing undergraduate students, in or outside of the lab, allow generalizable inferences about behavior. This is despite undergraduates differing in important ways from other populations. In addition, we document behavior 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 3. 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. This incentivized survey was run on four different populations. First, in the CCS, described above, 90% of the entire undergraduate population of Caltech participated. The second and third populations were a representative sample of the U.S., and a convenience sample of U.S. 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 use similar analytic methods to address each question. First, 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. Finally, as most experiments inspect linkages between behaviors, attributes, and treatments—in a word, correlations—we 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 CCS samples, as documented in Section 3. Generally, behaviors in the representative and MTurk samples lie closer to one another than the CCS sample is to either. 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).

Correlations, however, exhibit far greater similarity, and disagreement is largely consistent with the pattern of measurement error across samples. Most social science studies focus on correlations between attributes, rather than levels. For example, risk attitudes or altruism are often elicited, as they are suspected to be a potential channel for explaining some observed behavior. In only 7% of the correlations we examine do two samples have statistically significant correlations of the opposite sign. Moreover, 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 measurement error—and thus greater attenuation of correlations—than the CCS.[2] 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. At the time of writing, a Google Scholar search of "Mechanical Turk" and "Economics" yields over 23,000 results. Presumably, this is due to MTurk allowing the collection of large volumes of data quickly and cheaply.[3] We note,

[2]See Gillen et al. (2018) for background on the assessment of measurement error, its effects, and statistical approaches for overcoming it.

[3]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/.

however, that representative samples have similar features in terms of both ease of access and cost. Our results suggest MTurk has few advantages over other samples—it provides noisier observations than the student sample, while still suffering from concerns about representativeness. Nonetheless, the correlations between behaviors seen using the MTurk sample are fairly similar to those observed in the two other samples.[4]

Students who choose to participate in lab experiments differ little behaviorally from the overall population of students, as documented in Section 5. This addresses the concern that the same attributes that cause a student to *select* into lab experiments 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 6. This addresses a concern that *observer effects*, reviewed in the next section, are driving experimental results. For example, experimental results on the relatively 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

See also Dube et al. (2018), who estimate the low labor-supply elasticities on MTurk.

[4]There may be a use for MTurk in rapid prototyping and piloting. However, these practices are a subject of considerable debate in the experimental community.

CCS survey.[5] We see hardly any differences 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.[6] 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 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 is limited in scope. Lastly, 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.[7] 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 can be made using standard lab or survey methodology. We hope the methods we introduce in this paper open doors to further data-driven analyses of other aspects of external validity.

[5] Participants were not told ahead of time that this would be the experimental task, to avoid selection effects that would be specific to that experiment.

[6] Repeated surveys do not change other elicitations, see Subsection 4.2.

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

2 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 et al., 2013; Cappelen et al., 2015).[8] 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 et al., 2010; Horton et al., 2011; Berinsky et al., 2012; Arechar et al., 2018).[9] 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 et al., 2009; Cleave et al., 2013; Falk et al., 2013). Even so, guaranteed show-up fees yield somewhat more risk-averse lab participants (Harrison et al., 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

[8]Belot et al. (2015) compare behavior in five different games between Oxford students and local (non-representative) non-students. The results are in line with ours—non-students have more salient other-regarding preferences and exhibit less sophisticated strategic thinking. Exadaktylos et al. (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 et al. (2013) for a related study focusing on trust games. In contrast, Fosgaard (2018) illustrates 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 non-standard participants, including animals, people living in token economies, and so on.

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

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 on productivity, workers under observation seemed to out-perform those in a control group, even when nearly identical conditions were imposed.[10]

In dictator games, Hoffman et al. (1994) find an observer effect, while Bolton et al. (1998) do not.[11] 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, which 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.[12]

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, 2007b, 2008). Multiple recent papers advocate lab and experimental data's usefulness of (largely in response to Levitt and List—see, for instance, Falk and Heckman, 2009; Gächter, 2010; Kessler and Vesterlund, 2015; Camerer, 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.

[10]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).

[11]Laury et al. (1995) find no observer effect in public goods games. Anderhub et al. (2001) find few differences between online and lab participants in a game reminiscent of a consumption-saving problem.

[12]An important caveat to this statement is that we cannot test whether participating in a study in and of itself affects behavior. Given the fact that lab experiments are not naturalistic, it is difficult to envisage an experiment that could test for such an effect. Nonetheless, our results suggest that, even if some such responses are present, they are not sensitive to the level of monitoring by the researchers.

3 The Data

The foundation of our analysis is the Caltech Cohort Study (CCS), a repeated, incentivized survey covering over 90% 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.

3.1 The Student Samples

Caltech is an independent, privately-supported university located in Pasadena, California. It has approximately 900 undergraduate students, of which ∼40% 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% of the enrolled undergraduate student body (819/899) responded to the survey.[13] The average payment was $29.08 and the median time for survey completion was 35 minutes.[14] It is important to note that there is little concern about self-selection into the CCS from the participant population, due to our 90%+ response rates.

[13]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 shown in Appendix Table A.3, 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.

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

In Section 5, 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% (350/370) responded to the CCS Spring 2015 survey.[15] Conditional on participating in at least one experimental session, the median participation rate was 2 experiments per year.

In Section 6, 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 (with 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 raise concern that our student population is different from the pool utilized in most lab experiments. Such a concern should be mitigated by the following observations. First, replication 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 in Gillen et al., 2018, for details). Second, while top-10 schools account for 0.32% of the college-age population in the U.S., top-50 schools enroll only 3.77% of that population (using the *U.S. News and World Report* rankings). Thus, there seems to be little cause for concern that Caltech students are more "special" than students utilized in many other lab experiments. We do, however, believe that expanding the approach here to a wider set of universities and experimental labs, as well as experimental settings, would be of great use. We hope the methodology we offer increases the feasibility of such studies.

[15] As only enrolled students were eligible to participate, only 370 of the 403 students could participate due to early graduations or leaves of absence.

3.2 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. In Spring 2017 we utilized the SSI participant pool to run the CCS Spring 2015 survey on a representative sample of the U.S. population. We had 1,001 participants that were representative of the U.S. population across age, income, and gender. The average payment was $10.26, with an additional $3 required by SSI for each survey completion.[16] The median completion time was 33 minutes.

3.3 Amazon Mechanical Turk

In Spring 2016 we conducted our survey with a sample of 995 U.S.-based Amazon Mechanical Turk (MTurk) users. The average payment was $10.50 per participant.[17] The median completion time was 35 minutes.

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 U.S. population (see, for example, Ipeirotis, 2010; Berinsky et al., 2012; Huff and Tingley, 2015). We find similar patterns comparing the MTurk and representative samples, which are summarized in Appendix Table A.1.[18]

3.4 Description of Elicitations

Throughout this paper we examine a standard set of elicitations that we believe are particularly important for experimental work. Precise question wordings can be found in the

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

[17]This is considered a fairly high wage on MTurk for a task that took about half an hour to complete. As mentioned in the Introduction, current estimates of an hourly "wage" on MTurk range between $1–$5.

[18]Overall, MTurk workers are younger, somewhat more educated, have lower incomes, and are more likely to be single than participants in the representative sample.

screenshots posted at leeatyariv.com/ScreenshotsSpring2015.pdf. Throughout, 100 survey tokens were valued at $1 for our student sample, while 300 survey tokens were valued at $1 for our representative and MTurk samples. Participants were paid for all tasks. The location of questions within the survey was determined at random. Since we observe no order effects, we report aggregate results throughout.

3.4.1 Risk Elicitations

We used three different risk elicitation techniques.[19]

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 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 certainty equivalent that increased in units of 10 tokens from 0 to 100 or the gamble on the urn.[20]

Qualitative: Following Dohmen et al. (2011), participants were asked to rate themselves, on a scale of 0–10, in terms of their willingness to take risks.

[19] For an overview of risk elicitation techniques, see Charness et al. (2013).

[20] In order to prevent multiple crossovers, the online form automatically selected the lottery over a 0 token certainty equivalent, 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.

3.4.2 Discounting (δ)

Participants were asked a hypothetical question about how much money we would have to pay them in 60 days to forego a $150 payment delayed by only 30 days after the completion of the survey.[21] 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% of answers—those that demand less than $150, or more than $400.

3.4.3 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.

3.4.4 Prisoner's Dilemma

There were two symmetric Prisoner's Dilemma games with different payoffs. However, payoffs were scaled by a common factor to keep the same relative incentives across the two games. 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.

3.4.5 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.

[21] Having both payoffs in the future removes any effects of present bias.

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.

3.4.6 Cognitive Tasks

We used two types of cognitive tasks.

Raven's Matrices: Participants were asked to complete five Raven's Matrices, which are commonly used for assessing abstract reasoning, see Raven (1936). Each Raven's Matrix consisted of a 3x3 matrix with eight of the nine cells featuring a geometric design. Participants had to choose the correct geometric pattern to complete the matrix out of six possibilities. 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 thus are designed to measure individuals' ability to reflect on problems and override immediate intuitions.[22] 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.

3.4.7 Confidence in Guesses

Following 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.

[22] We used variations on the original questions, as some responders may have been exposed to the originals.

3.4.8 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 in the first one, or whether they preferred to receive 10 tokens per correct sum regardless of others' performance. This second task provides an elicitation of willingness to compete.[23]

3.4.9 Implicit Association Tests (IAT)

We assessed implicit attitudes toward gender and race separately using two Implicit Attitude Tests (IATs; Greenwald et al., 1998).[24] Although controversial, IATs are often viewed as measures of discriminatory attitudes.

4 Comparison of Different Samples

We begin our analysis by comparing different participant pools: university students (from the CCS), a convenience sample (from MTurk), and a representative sample of the U.S. (from SSI). 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 behave in a more normatively rational and cognitively sophisticated way. These mean differences do not tend to lead to disagreement in the signs

[23] For further details, see Gillen et al. (2018).

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

of correlations between behaviors. The differences in the levels of statistical significance associated with those correlations are broadly consistent with differences in measurement error across the samples. In particular, the CCS exhibits more significant correlations and also the lowest level of measurement error.

4.1 Differences in Behavior

The average measures of each behavior are quite different across the three samples, as shown in Table 1. 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). This implies that a population with higher than average cognitive ability should exhibit different behaviors.

It is also clear from Table 1 that the representative and MTurk samples are closer to each other than either is to the CCS.[25] Moreover, the average levels of behavioral measures in the MTurk sample are usually between those in the representative sample and 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 vary across measures. Differences are the most substantial for the Risky Projects 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 mean differences we report are not only statistically significant, they are also substantially 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 re-

[25]The differences we see are not due to differences in gender and race composition, as can be seen from Appendix Table A.6. This table re-weights the CCS sample to match the gender and racial composition of the representative samples.

[26]Gillen et al. (2018) also show that these elicitations are relatively stable over time, exhibit less measurement error, and generate different responses across genders.

Table 1: Differences in choices: three samples

	Samples			Differences		
	Rep.	MTurk	CCS	Rep.−MTurk	Rep.−CCS	MTurk−CCS
First Risky Project	46	44	59	2.7**	−13***	−16***
(out of 100)	(.89)	(.85)	(1.2)	(1.2)	(1.5)	(1.4)
Second Risky Project	95	98	143	−2.7	−48***	−45***
(out of 200)	(1.8)	(1.7)	(2.1)	(2.5)	(2.8)	(2.7)
First Risky Urn	49	56	59	−7.3***	−10***	−3.2***
(20 balls)	(.76)	(.63)	(.52)	(.99)	(.96)	(.84)
Second Risky Urn	67	78	86	−11***	−19***	−8.0***
(30 balls)	(1.2)	(.96)	(.74)	(1.6)	(1.5)	(1.3)
Qualitative Risk Aversion	5.0	4.9	5.8	0.11	−0.76***	−0.87***
	(.08)	(.08)	(.08)	(.11)	(.11)	(.11)
Discounting (δ)	0.67	0.67	0.77	0.00	−0.10***	−0.10***
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
First Dictator Game	39	26	14	14***	25***	12***
(given out of 100)	(.58)	(.71)	(.84)	(.91)	(1.0)	(1.2)
Second Dictator Game	115	74	38	41***	77***	36***
(given out of 300)	(1.7)	(2.0)	(2.4)	(2.7)	(2.9)	(3.1)
Dictator, Tokens Given	39	30	26	8.9***	12.8***	3.8***
are Doubled	(.62)	(.79)	(1.2)	(1.0)	(1.3)	(1.4)
Dictator, Tokens Given	39	25	9.0	14***	30***	16***
are Halved	(.61)	(.74)	(.69)	(.95)	(.91)	(1.0)
Prisoner's Dilemma	46	57	68	−11***	−22***	−11***
(% dominant strat.)	(1.2)	(1.3)	(1.5)	(1.8)	(1.9)	(2.0)
Reported Heads	2.9	3.0	3.3	−0.14***	−0.41***	−0.28***
(out of 5)	(.03)	(.03)	(.04)	(.05)	(.05)	(.05)
Reported Switches	4.4	4.5	5.5	−0.18**	−1.1***	−0.96***
(out of 9)	(.06)	(.06)	(.07)	(.08)	(.09)	(.09)
Raven's Matrices	1.2	1.3	1.8	−0.17***	−0.62***	−0.46***
(out of 5)	(.03)	(.04)	(.04)	(.05)	(.05)	(.06)
CRT	0.46	1.4	1.7	−0.89***	−1.2***	−0.31***
(out of 3)	(.03)	(.04)	(.04)	(.04)	(.04)	(.05)
Confidence in Guesses	2.9	2.9	3.1	−0.05	−0.25***	−0.20***
	(.03)	(.03)	(.03)	(.05)	(.05)	(.05)
Competition	40	29	33	11***	6.8***	−3.8*
(% competing)	(1.6)	(1.5)	(1.7)	(2.1)	(2.3)	(2.2)
IAT Race	59	68	81	−8.9	-23**	−14*
	(8.2)	(4.8)	(5.6)	(9.5)	(10)	(7.3)
IAT Gender	104	90	94	13*	9.4	−4.0
	(5.9)	(4.8)	(5.9)	(7.6)	(8.4)	(7.5)
Percent Male	47	50	62	−3.5	−13***	−10***
	(1.6)	(1.6)	(1.7)	(2.2)	(2.3)	(2.3)

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses.

maining variables. 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.

In magnitude terms, the differences between samples are large. For example, differences in the amount allocated in the Risky Project measures are between 15–25% of the maximum possible differences, or around 50–75% of a standard deviation of these measures within a given sample. The differences in discount rates are similarly substantial in terms of sample standard deviations.[27] Differences in giving in dictator games are also large, corresponding to approximately 10% 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.[28] 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 4.3.

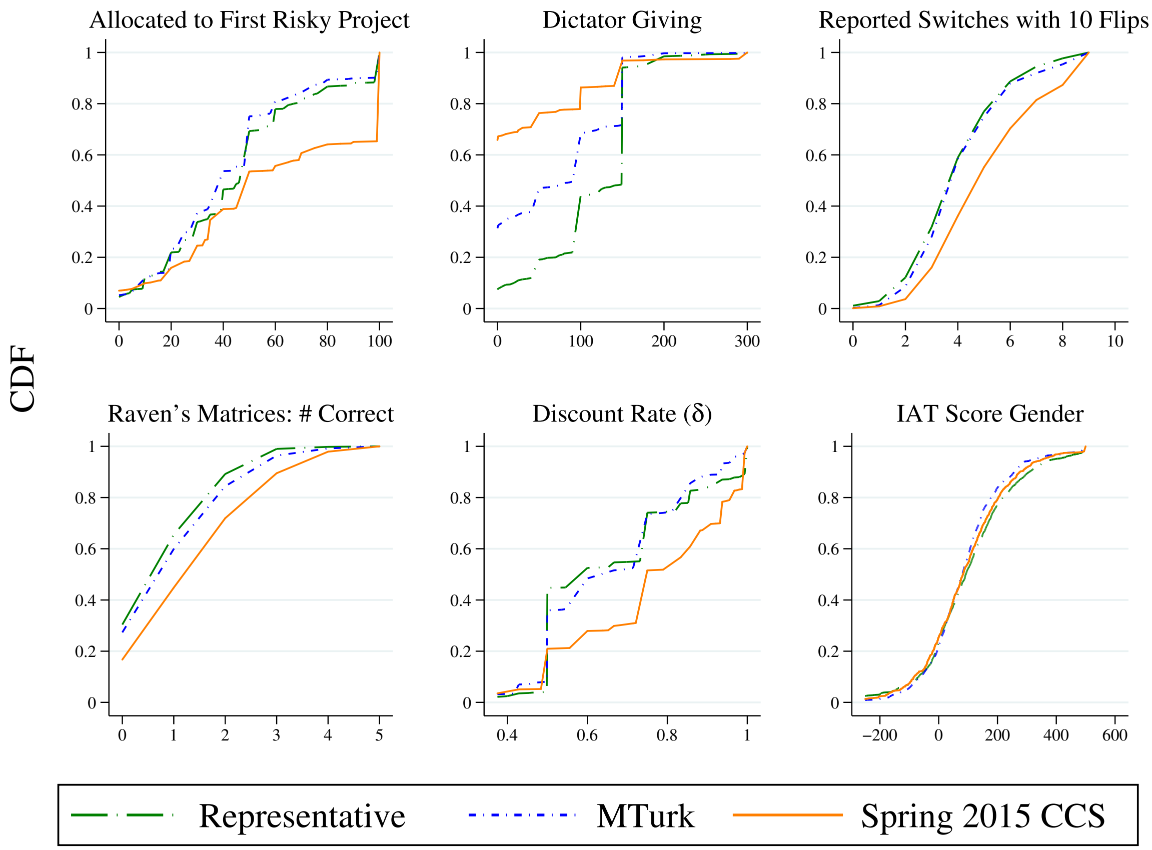
The distributions of various behavioral elicitations, in Figure 1, display a clear pattern: there is a first-order stochastic relationship between the CCS and the representative sample in most behaviors.[29] Additionally, the distribution in the MTurk sample is generally closer to the representative sample, as implied by Table 1. In some cases, we also see a clear first-order stochastic relationship between the MTurk distributions and those of the CCS and the representative sample, although this is less common. Overall, these facts imply that the differences in means in Table 1 are not driven by small groups of people with extreme

[27] 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. With a discount rate of 0.77 (CCS), the future wage y would need to surpass only $7,200 to justify the investment in training.

[28] This observation is in line with prior work on this elicitation, see Falk et al. (2013) and Cappelen et al. (2015), as well as the discussion of the literature in Section 2.

[29] Figure A.1 in the 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



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 student population is closer to the ideal of normative rationality and exhibits 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.

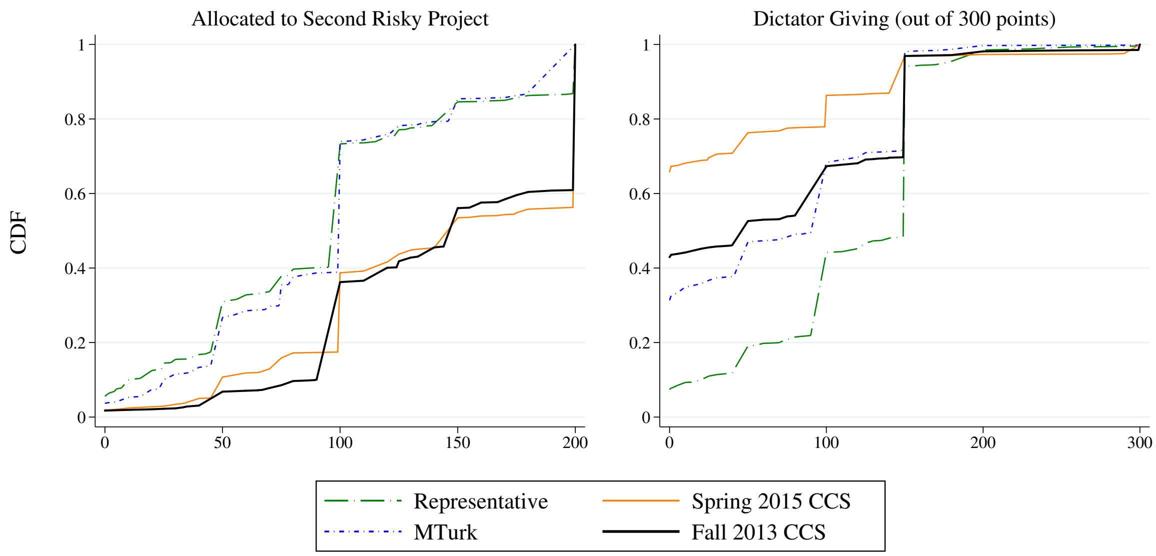
Participants in the CCS exhibited wide variation in their enthusiasm to complete the surveys, with some completing the survey as soon as it was announced, and some waiting several weeks. We find no correlation between the speed at which participants responded to the survey and the behavioral proxies we measure. Specifically, while surveys of the MTurk and representative samples were completed within a couple of hours, and a few days, respectively, it took multiple weeks and reminders to obtain the same level of participation in the CCS. As shown in Appendix Table A.3, 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. We see one exception, detailed in Table A.4: those 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 further suggests that the fact that we were unable to survey $\sim$9% of the Caltech undergraduate student body does not significantly impact the results discussed here, and previews the results in Sections 5 and 6.

4.2 Learning and the CCS

A final difference between samples is that 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 incentivized tasks, or about their own preferences. Changes in distributions of responses over time could then imply that responses in later installments of the CCS survey are an artifact of repetition, rather than a reflection of 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

Figure 2: Repetition does not alter most behaviors, except Dictator Giving.



pattern of responses comparing the first installment of the 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.[30] 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.[31]

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

[30]The Fall 2013 survey contained only one risky-project task, which was identical to the Second Risky Project task on the Spring 2015 CCS. The description of these two tasks (first, second) was chosen without regard to their longevity on the survey.

[31]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.

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.

4.3 Similarity of Correlations

Experimental work is often concerned with the correlations between attributes. Attempting to connect behaviors with each other, with demographics, or with a treatment, boils down to an examination of correlations. Thus, in this subsection, we examine the correlations between different behaviors in our data, with the results shown in Figure 3. Our method of comparison is motivated by a very simple notion of replication. In particular, if a study documents a particular statistically significant correlation in a particular sample, would a statistically significant correlation of the same sign be found in another sample? Although this notion is quite simple, it is close to those used in recent multi-study replication exercises (Camerer et al., 2016; Open Science Collaboration, 2015).[32]

Figure 3 displays the sign and significance (at the 10% level) of correlations in the three samples: first the representative sample, then MTurk, then the CCS. When there are multiple elicitations of an attribute, we use the first principal component of these elicitations. 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. The threshold of $p < 0.1$ is chosen conservatively, although many readers may prefer a cutoff of $p < 0.05$. Appendix Figures A.2 and A.3 are similar to Figure 3, but use p-value cutoffs of 0.05 and 0.01, respectively.[33]

[32]Statistical comparisons of correlation matrices are complicated by the fact that 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.

[33]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 3: Correlations across the representative sample, MTurk, and CCS.

	Risk Aversion	Discounting (δ)	Dictator	Prisoner's Dilemma	Lying	Cognitive	Confidence	Compete	IAT Race	IAT Gender	Male
Risk Aversion		0+−	+0+	−00	−0−	00−	−	−	0	0+0	−
Discounting (δ)	0+−		−−0	0	0+0	+	0−0	0	+00	00+	+00
Dictator	+0+	−−0		−	−	−−0	0+−	00−	0	0	0−−
Prisoner's Dilemma	−00	0	−		0++	+	00+	0++	0	0	00+
Lying	−0−	0+0	−	0++		−++	+0+	0++	0	0	0++
Cognitive	00−	+	−−0	+	−++		−−+	00+	00−	0	+
Confidence	−	0−0	0+−	00+	+0+	−−+		+	0	00+	+
Compete	−	0	00−	0++	0++	00+	+		0+0	00+	0++
IAT Race	0	+00	0	0	0	00−	0	0+0		+	0
IAT Gender	0+0	00+	0	0	0	0	00+	00+	+		00+
Male	−	+00	0−−	00+	0++	+	+	0++	0	00+	

Notes: indicates complete agreement, complete disagreement, and two out of three samples agreeing.

There is a lot of consistency between the correlations observed in the three samples. In only four out of the 55 correlations we examine (7%) 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%) 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.

Much of the "moderate" disagreement is driven by statistically insignificant results in either the MTurk or representative sample (or both). Indeed, of the 28 moderate disagreements, 20 feature statistically insignificant correlations in the representative sample, 16 in the MTurk sample, and only 9 in the CCS, which has an ∼18% smaller sample size. As correlations may be attenuated by noise, or measurement error, we next turn to an examination of the extent of measurement error across these three samples.

4.4 Measurement Error

We use a simple method to ascertain the extent of measurement error in a sample, building on Gillen et al. (2018). 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 measurement error, 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.[34] Then, we have that:

$$1 - \widehat{\mathrm{Corr}}[X^a, X^b] \to_p 1 - \mathrm{Corr}[X^a, X^b] = \frac{\sigma^2_{\nu_X}}{\sigma^2_{X^*} + \sigma^2_{\nu_X}}. \quad (1)$$

Thus, $1 - \widehat{\mathrm{Corr}}[X^a, X^b]$ is an estimate of the proportion of variation of an elicitation that is due to measurement error.

[34]This implies that $\mathbb{E}[\nu_X^a \nu_X^b] = 0$ and $\frac{\mathrm{Var}[\nu_X^a]}{\mathrm{Var}[X^a]} = \frac{\mathrm{Var}[\nu_X^b]}{\mathrm{Var}[X^b]} := \frac{\mathrm{Var}[\nu_X]}{\mathrm{Var}[X]}$.

Table 2: Percent of Variation due to Measurement Error

	Rep.	MTurk	CCS
Risky Projects	59%	47%	43%
	(2.9%)	(2.7%)	(2.9%)
Risky Urns	35%	32%	25%
	(2.4%)	(2.3%)	(2.3%)
Lottery Menu	49%	33%	28%††
	(2.7%)	(2.4%)	(2.4%)
Ambiguous Urn	30%	31%	22%
	(2.3%)	(2.3%)	(2.1%)
Compound Urn	31%	26%	26%†
	(2.3%)	(2.1%)	(2.2%)
Dictator Giving	37%	18%	15%
	(2.5%)	(1.8%)	(1.8%)
IAT Race	36%	46%	42%
	(2.4%)	(2.7%)	(2.8%)
IAT Gender	45%	46%	39%
	(2.6%)	(2.7%)	(2.8%)
N	1,000	995	819

Notes: † indicates figure is from the Fall 2014 CCS (N=893), and †† indicates figure is from the Fall 2015 CCS ($N = 863$).

Using this relationship, Table 2 shows that the CCS has the lowest measurement error of the three samples in all elicitations except for IAT Race. These differences are often significant when comparing the CCS and the representative sample. As greater measurement error leads to greater attenuation of estimated correlations, variations in noise across our samples can help explain the patterns we identified in Figure 3.

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. 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 et al., 2018).[35] Furthermore, our analysis suggests a natural cost-benefit tradeoff: while student-based studies often entail higher costs, they imply lower noise. Such cost-benefit analysis also suggests that MTurk may have few advantages over representative samples. Indeed, the costs of running a study on a representative or MTurk sample are comparable. So is the level of measurement error. However, MTurk is not representative and is associated with greater levels of measurement error than those seen in our student sample.

5 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 measurement error. This comparison was done using a survey that covered nearly the entire population of Caltech students. In contrast, lab experiments include non-random, and possibly non-representative, samples of university students: those who *select* to go to the lab. In this section we ask whether selection into the lab results in non-representative 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. This difference 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).

[35] The incentives used for our student population are on par with standard payments in experimental labs. It would be interesting to investigate whether student studies run with lower incentives, comparable to those used on MTurk and on our representative sample, yield comparable behaviors and measurement errors. On MTurk, DellaVigna and Pope (2017) suggest that incentives have a substantial impact on performance in a simple-effort task.

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 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.[36] The averages for each population are displayed in the first three columns of Table 3, while the final two columns compare the two lab-going populations to the overall population that participated in the CCS.

We see little 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 explored in the previous section. The average 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

[36]There are two caveats to this exercise. First, because we do not observe all cohorts in the Spring 2015 survey for the full time they are at Caltech, we effectively have a slice of their overall participation records. Second, the number of experiments run at SSEL fluctuates over the years, which impacts the number of experiments *available* to students at different times.

Table 3: Those who participate in lab experiments are not substantially different from the overall population.

	Samples			Differences	
	Everyone (E)	Participant (P)	Weighted Participant (WP)	E−P	E−WP
First Risky Project	59	55	52	4.8**	7.3***
(out of 100)	(1.2)	(1.8)	(1.8)	(2.2)	(2.2)
Second Risky Project	143	139	132	4.2	11***
(out of 200)	(2.1)	(3.2)	(3.3)	(3.8)	(3.9)
First Risky Urn	59	58	58	0.82	1.0
(20 balls)	(.52)	(.77)	(.74)	(.93)	(.90)
Second Risky Urn	86	86	85	0.06	0.89
(30 balls)	(.73)	(1.1)	(.99)	(1.3)	(1.2)
Qualitative Risk Aversion	5.8	5.7	5.7	0.05	0.09
	(.08)	(.12)	(.12)	(.15)	(.15)
Discounting (δ)	0.77	0.78	0.77	-0.01	-0.01
	(.01)	(.01)	(.01)	(.01)	(.01)
First Dictator Game	14	12	9.2	2.2	4.7***
(given out of 100)	(.84)	(1.1)	(1.0)	(1.4)	(1.3)
Second Dictator Game	38	32	26	6.1	12***
(given out of 300)	(2.4)	(3.2)	(2.8)	(3.9)	(3.7)
Dictator, Tokens Given	26	26	26	-0.00	-0.10
are Doubled	(1.2)	(1.8)	(1.8)	(2.2)	(2.2)
Dictator, Tokens Given	9.0	7.8	6.0	1.2	2.9***
are Halved	(.68)	(.94)	(.84)	(1.2)	(1.1)
Prisoner's Dilemma	68	67	69	0.68	-1.4
(% dominant strat.)	(1.5)	(2.3)	(2.3)	(2.8)	(2.7)
Reported Heads	3.3	3.4	3.5	-0.11	-0.18**
(out of 5)	(.04)	(.06)	(.06)	(.08)	(.08)
Reported Switches	5.5	5.5	5.8	-0.01	-0.34**
(out of 9)	(.07)	(.11)	(.11)	(.13)	(.13)
Raven's Matrices	1.8	1.8	1.8	-0.01	-0.02
(out of 5)	(.04)	(.07)	(.07)	(.08)	(.08)
CRT	1.7	1.7	1.7	-0.03	-0.07
(out of 3)	(.04)	(.06)	(.06)	(.07)	(.07)
Confidence in Guesses	3.1	3.1	3.1	0.09	0.06
	(.03)	(.05)	(.05)	(.06)	(.06)
Competition	33	34	33	-0.26	0.16
(% competing)	(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)

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses.

one-fourth and one-sixth, respectively, of the corresponding differences when comparing the representative sample and the CSS.[37]

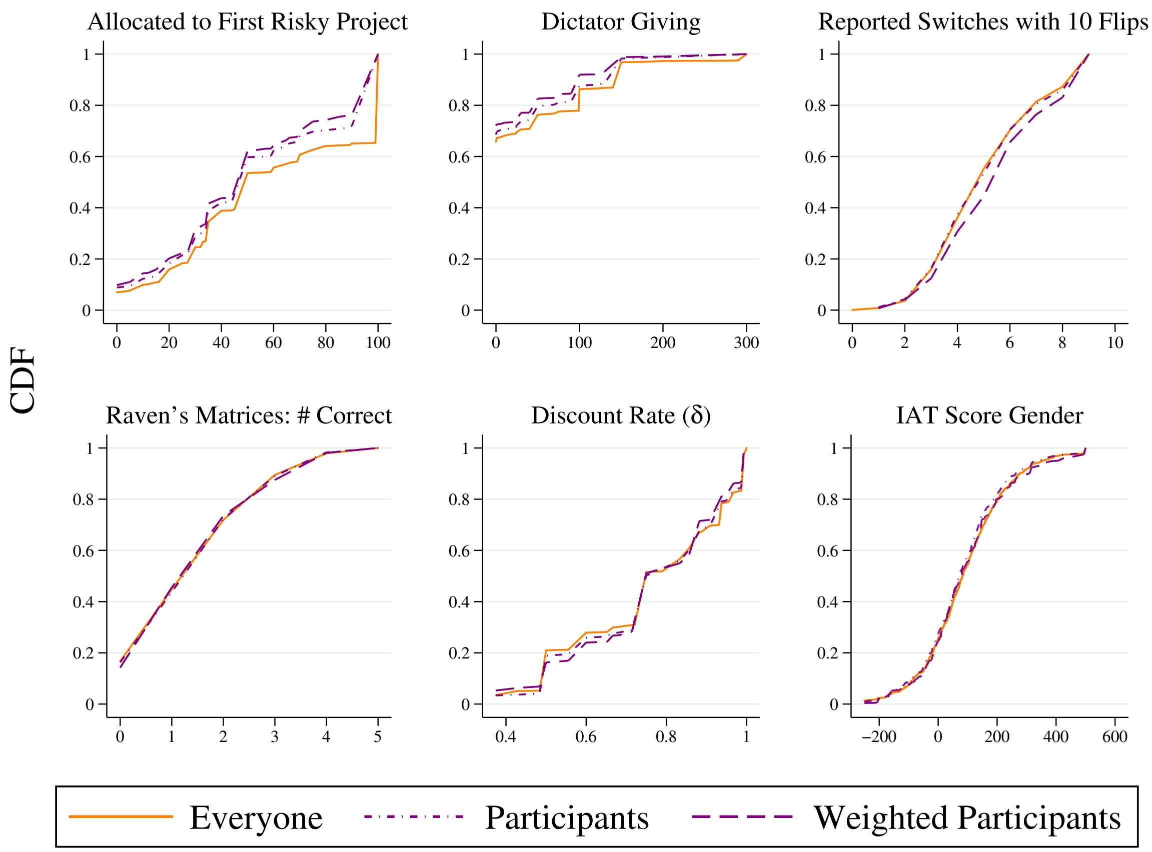
The average differences we observe are again indicative of first-order stochastic dominance relations in the underlying distributions, as shown in Figure 4. The panels of this figure display the cumulative distribution functions corresponding to the same selected set of elicitations depicted in Figure 1 for the overall and lab-going subpopulations of the CCS. These images echo the message emerging from Table 3—lab participants are similar to the underlying population, with some small differences for certain elicitations when weighting the set of participants by lab experience.[38]

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, and papers that followed) are not a consequence of selection into the lab. This is a particularly important counter-factual to

[37]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 Appendix Table A.5.

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

Figure 4: Distribution of responses in the CCS: Overall, Participants, and Weighted Participants



the literature that suggests that 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 correlation patterns across subsamples, although smaller subsamples produce more statistically significant results. Figure 5 is an analogue of Figure 3, where each entry corresponds, from left to right, to the sign and significance of the correlation in the CCS overall population, the subset of CCS participants who showed up to at least one SSEL experiment, and that same subset weighted by lab experience. Once again, a positive and significant correlation (at the 10% 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.

Figure 5: Correlations across everyone, participants, and weighted participants

	Risk Aversion	Discounting (δ)	Dictator	Prisoner's Dilemma	Lying	Cognitive	Confidence	Compete	IAT Race	IAT Gender	Male
Risk Aversion		−	+00	−0−	−−0	−10	−	−	0	0	−
Discounting (δ)	−		0	0	0	+	0	0	0	+00	0
Dictator	+00	0		−	−	−	−	−	00+	0−0	−00
Prisoner's Dilemma	−0−	0	−		+	++0	+	+00	0	0+0	+
Lying	−−0	0	−	+		+00	+00	+	0	0++	+00
Cognitive	−−0	+	−	++0	+00		+	+	−00	0	+
Confidence	−	0	−	+	+00	+		+	0	+00	+
Compete	−	0	−	+00	+	+	+		0	+00	+
IAT Race	0	0	00+	0	0	−00	0	0		+	0
IAT Gender	0	+00	0−0	0+0	0++	0	+00	+00	+		+
Male	−	0	−00	+	+00	+	+	+	0	+	

Notes: indicates complete agreement, complete disagreement, and two out of three samples agreeing.

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 statistically significant correlation in another. In 37 of the 55 (67%) correlations, there is full agreement. In 40 of the 55 (73%), 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 3 and Figure 4. In only four of the cells (7%) 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.[39] In only 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% level) is between the correlation found in the overall CCS population and that found when participants are weighted by lab experience.

In summary, we see some selection effects in lab participation, though the lab-going subsample is non-representative 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 by our data. Lastly, 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—one risk elicitation, one dictator game, and one lying task—are jointly statistically significant, while others are not.[40] That is, if one controlled for only three

[39] Once again, analogous figures for 5% and 1% significance levels are shown in the Appendix, in particular Appendix Figures A.5 and A.6. Those figures generate similar conclusions. However, as before, when significance restrictions become more demanding, fewer correlations are significant, which mechanically causes the appearance of more agreement.

[40] Specifically, the First Risky Project, Dictator Giving, Doubled, and Reported Heads, are jointly statistically significant. All other elicitations are jointly statistically insignificant.

variables in Table 3, 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.

6 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 profound changes in behavior. Indeed, the *Hawthorne* or *observer effect*, reviewed above, has been a topic of discussion for over 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 in the lab.

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 invited students from the cohorts covered by the survey to participate.[41] We were present in the lab for all sessions. In total, 97 students participated, 55% of which were women. This is in agreement with, but somewhat more extreme, than the over-representation of women in lab experiments observed in the previous section. Lab participants retook the CCS survey from the Spring of 2015, which allows us to compare whether the lab environment itself changes participants' responses. Ninety-nine percent (96/97) of the students participating in our lab experiments 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. The first column of Table 4 shows the average responses on the Spring 2015 CCS of those who came to our Summer 2015 experiments. Consistent with the results in the previous section, respondents are more

[41] 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 4: Differences in choices: Lab versus Survey

	Survey	Lab	Difference
First Risky Project	54	54	0.50
(out of 100)	(3.2)	(3.3)	(4.6)
Second Risky Project	134	138	−3.9
(out of 200)	(5.8)	(5.8)	(8.2)
First Risky Urn	60	56	4.0**
(20 balls)	(1.5)	(1.3)	(2.0)
Second Risky Urn	87	84	3.0
(30 balls)	(2.0)	(1.7)	(2.6)
Qualitative Risk Aversion	5.4	5.4	−0.01
	(.21)	(.21)	(.29)
Discounting (δ)	0.78	0.78	−0.01
	(.02)	(.02)	(.03)
First Dictator Game	13	10	3.1
(given out of 100)	(2.3)	(2.0)	(3.0)
Second Dictator Game	35	24	11
(given out of 300)	(6.5)	(6.0)	(8.8)
Dictator, Tokens Given	29	29	−0.42
are Doubled	(3.4)	(3.7)	(5.1)
Dictator, Tokens Given	7.1	5.3	1.9
are Halved	(1.7)	(1.5)	(2.2)
Prisoner's Dilemma	68	70	−1.6
(% dominant strat.)	(4.2)	(4.5)	(6.1)
Reported Heads	3.3	3.3	0.07
(out of 5)	(.12)	(.11)	(.16)
Reported Switches	5.3	5.2	0.04
(out of 9)	(.20)	(.17)	(.26)
Raven's Matrices	1.9	2.5	−0.58***
(out of 5)	(.13)	(.13)	(.19)
CRT	1.6	2.1	−0.48***
(out of 3)	(.11)	(.11)	(.16)
Confidence in Guesses	2.9	2.9	−0.11
	(.08)	(.09)	(.12)
Competition	32	29	3.1
(% competing)	(4.8)	(4.7)	(6.7)
IAT Race	109	84	25
	(18)	(13)	(22)
IAT Gender	99	65	34
	(15)	(15)	(21)
Percent Male	45	45	0.00
	(5.1)	(5.1)	(7.2)

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses.

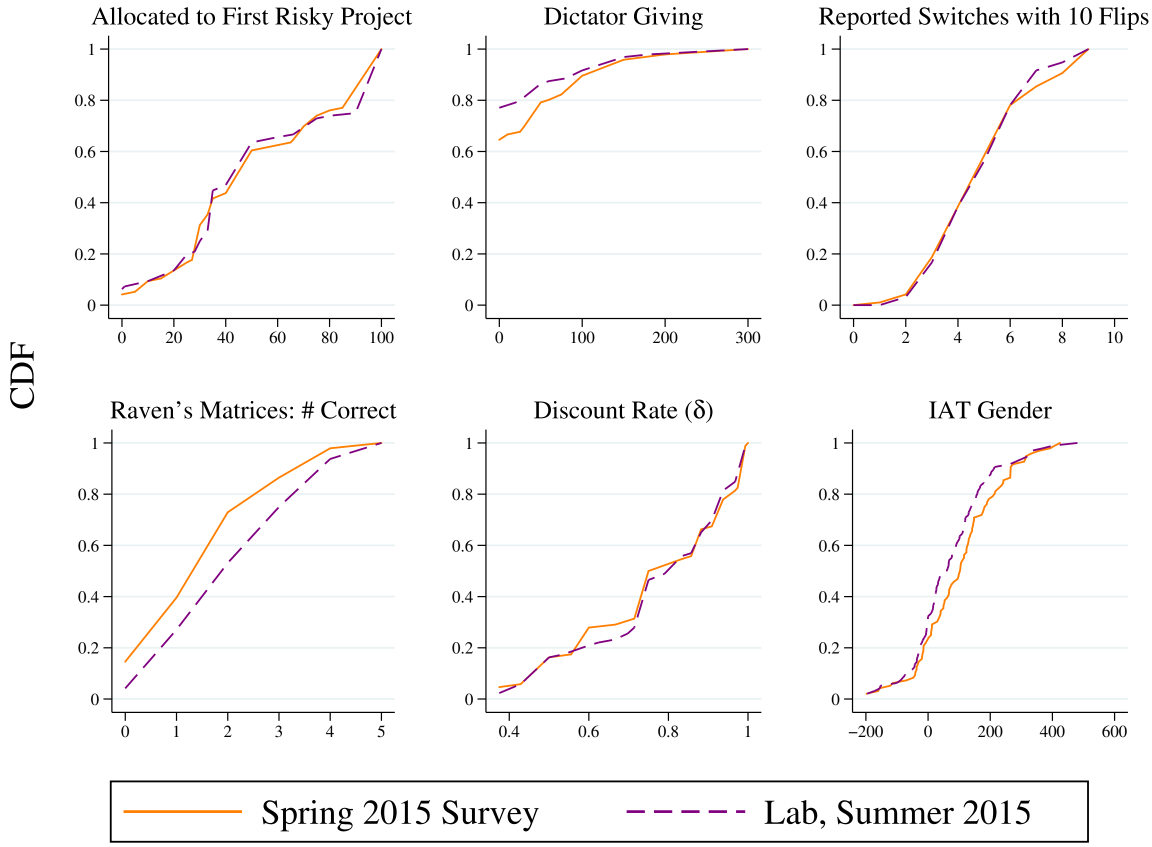
likely to be female, less altruistic, and more risk averse. The second column displays the same behaviors elicited from the same population, but this time in the lab. As can be seen, on most behaviors the differences are small in magnitude and statistically insignificant.

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% 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 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 Subsection 4.2. 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.

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 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 6, which depicts the analogous cumulative distributions to Figures 1 and 4 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,

Figure 6: Distribution of responses in the Spring 2015 survey vs. the Lab ($N = 96$)



the distributions in the lab and on the survey are nearly identical.

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.[42] We therefore include the analogue to Figures 3 and 5 in Appendix Figure A.7.

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, for example, Laury et al., 1995; Bolton et al., 1998).[43]

[42]Indeed, 45 of the 55 correlations we inspected in our SSEL data and in our CCS data—restricted to participants who participated in our lab experiment—were insignificant.

[43]These results are also in line with Anderhub et al. (2001), who compared behavior in the lab and online,

7 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?

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 explained by statistical insignificance in the representative and MTurk samples, driven by higher measurement error. 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.

Taken together, our findings should be reassuring for researchers using standard student-based experiments, and those who might like to rely on their results. In particular, this study provides evidence that generalizable inferences about human behavior are possible from lab experiments. There are other 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.[44] In practice, it would be quite unusual for a practitioner to take evidence from a student-based lab sample directly to public policy.

albeit with different participants in each, and in only one game.

[44] 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.

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 will 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 lab). We hope the methodology we introduce opens the door to future data-driven studies of other facets of external validity.

References

Alm, James, Kim M. Bloomquist, and Michael McKee, "On the External Validity of Laboratory Tax Compliance Experiments," *Economic Inquiry*, Jan 2015, *53* (2), 1170–1186.

Anderhub, Vital, Rudolf Müller, and Carsten Schmidt, "Design and Evaluation of an Economic Experiment via the Internet," *Journal of Economic Behavior & Organization*, 2001, *46* (2), 227–247.

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

Armantier, Olivier and Amadou Boly, "Comparing Corruption in the Laboratory and in the Field in Burkina Faso and in Canada," *The Economic Journal*, Apr 2013, *123* (573), 1168–1187.

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

Berinsky, Adam J. Gregory A. Huber, and Gabriel S. Lenz, "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Political Analysis*, 2012, *20* (3), 351–368.

Bolton, Gary E., Elena Katok, and Rami Zwick, "Dictator Game Giving: Rules of Fairness Versus Acts of Kindness," *International Journal of Game Theory*, 1998, *27* (2), 269–299.

Camerer, Colin F., "The Promise and Success of Lab–Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List," in Guillaume R. Fréchette and Andrew Schotter, eds., *The Handbook for Experimental Economic Methodology*, Oxford, UK: Oxford University Press, 2015, chapter 14.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu, "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 2016, *351* (6280), 1433–1436.

Cappelen, Alexander W., Knut Nygaard, Erik Ø. Sørensen, and Bertil Tungodden, "Social Preferences in the Lab: A Comparison of Students and a Representative Population," *The Scandinavian Journal of Economics*, 2015, *117* (4), 1306–1326.

Charness, Gary, Uri Gneezy, and Alex Imas, "Experimental Methods: Eliciting Risk Preferences," *Journal of Economic Behavior & Organization*, 2013, *87* (1), 43–51.

Cleave, Blair L, Nikos Nikiforakis, and Robert Slonim, "Is there Selection Bias in Laboratory Experiments? The Case of Social and Risk Preferences," *Experimental Economics*, 2013, *16* (3), 372–382.

Coppock, Alexander, "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach," *Political Science Research and Methods*, 2018, *forthcoming.*

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

Diedenhofen, Birk and Jochen Musch, "cocor: A Comprehensive Solution for the Statistical Comparison of Correlations," *PlOS ONE*, 2015, *10* (4), e0121945. https://doi.org/10.1371/journal.pone.0121945.

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

_ , _ , _ , _ et al., "On the Relationship Between Cognitive Ability and Risk Preference," 2018. Rationality and Competition Discussion Paper 76.

_ , _ , _ , _ , Jürgen Schupp, and Gert G Wagner, "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association*, 2011, *9* (3), 522–550.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri, "Monopsony in Online Labor Markets," Technical Report, NBER Working Paper #24416 2018.

Erat, Sanjiv and Uri Gneezy, "White Lies," *Management Science*, 2012, *58* (4), 723–733.

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

Falk, Armin and James J Heckman, "Lab Experiments are a Major Source of Knowledge in the Social Sciences," *Science*, 2009, *326* (5952), 535–538.

_ , Stephan Meier, and Christian Zehnder, "Do Lab Experiments Misrepresent Social Preferences? The Case of Self-selected Student Samples," *Journal of the European Economic Association*, 2013, *11* (4), 839–852.

Fosgaard, Toke R., "Cooperation Stability: A Representative Sample in the Lab," 2018. IFRO Working Paper 2018/08.

Fréchette, Guillaume R., "Laboratory Experiments: Professionals versus Students," in Guillaume R. Fréchette and Andrew Schotter, eds., *Handbook of Experimental Economic Methodology*, Oxford University Press, 2015, chapter 17, pp. 360–390.

_ , "Experimental Economics Across Subject Populations," in John H. Kagel and Alvin E. Roth, eds., *The Handbook of Experimental Economics*, Vol. 2, Princeton, New Jersey: Princeton University Press, 2016, chapter 7, pp. 435–480.

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

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

Gillen, Ben, Erik Snowberg, and Leeat Yariv, "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study," *Journal of Political Economy*, 2018, *Forthcoming.*

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

Gneezy, Uri, "Deception: The Role of Consequences," *American Economic Review*, 2005, *95* (1), 384–394.

_ **and Jan Potters**, "An Experiment on Risk Taking and Evaluation Periods," *The Quarterly Journal of Economics*, 1997, *112* (2), 631–645.

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

_ **, Debbie E McGhee, and Jordan LK Schwartz**, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology*, 1998, *74* (6), 1464–1480.

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

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

Herbst, D. and A. Mas, "Peer Effects on Worker Output in the Laboratory Generalize to the Field," *Science*, Oct 2015, *350* (6260), 545–549.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith, "Preferences, Property Rights, and Anonymity in Bargaining Games," *Games and Economic Behavior*, 1994, *7* (3), 346–380.

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

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

Ioannidis, John P.A., "Why most Published Research Findings are False," *PLOS Medicine*, 2005, *2* (8), e124: https://doi.org/10.1371/journal.pmed.0020124.

Ipeirotis, Panagiotis G, "Demographics of Mechanical Turk," 2010. NYU Center for Digital Economy Research, *mimeo.*

Jones, Stephen RG, "Was there a Hawthorne Effect?," *American Journal of Sociology*, 1992, *98* (2), 451–468.

Kessler, Judd and Lise Vesterlund, "The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects," *Handbook of Experimental Economic Methodology, Oxford University Press, Oxford, UK*, 2015, *20.*

Laury, Susan K., James M. Walker, and Arlington W. Williams, "Anonymity and the Voluntary Provision of Public Goods," *Journal of Economic Behavior & Organization*, 1995, *27* (3), 365–380.

Levitt, Steven D. and John A. List, "Viewpoint: On the Generalizability of Lab Behaviour to the Field," *Canadian Journal Of Economics*, 2007, *40* (2), 347–370.

Levitt, Steven D and John A List, "What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?," *Journal of Economic Perspectives*, 2007, *21* (2), 153–174.

_ **and** _ , "Homo economicus Evolves," *Science*, 2008, *319* (5865), 909–910.

_ **and** _ , "Was there Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments," *American Economic Journal: Applied Economics*, 2011, *3* (1), 224–38.

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

Niederle, Muriel and Lise Vesterlund, "Do Women Shy Away from Competition? Do Men Compete too Much?," *Quarterly Journal of Economics*, August 2007, *122* (3), 1067–1101.

Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, 2015, *349* (6251), 943.

Orne, Martin T, "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and their Implications," *American Psychologist*, 1962, *17* (11), 776–783.

Ortoleva, Pietro and Erik Snowberg, "Overconfidence in Political Behavior," *American Economic Review*, February 2015, *105* (2), 504–535.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis, "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, August 2010, *5* (5), 411–419.

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

Roth, Alvin E., "Bargaining Experiments," in John H. Kagel and Alvin E. Roth, eds., *Handbook of Experimental Economics*, Princeton, New Jersey: Princeton University Press, 1995, chapter 4, pp. 253–348.

Schram, Arthur, "Artificiality: The Tension between Internal and External Validity in Economic Experiments," *Journal of Economic Methodology*, 2005, *12* (2), 225–237.

Online Appendix—Not Intended for Publication

A Additional Tables and Figures

Table A.1: Demographic attributes of MTurk compared to a representative sample

	Representative	MTurk
Age		
18–25	16%	23%
26–54	53%	70%
55–64	18%	6%
65+	13%	1%
Race / Ethnicity		
White	71%	74%
Black	12%	8%
Hispanic	8%	6%
Asian	5%	7%
Education		
High School or Less	20%	10%
Some College	23%	30%
Associates Degree	10%	11%
Bachelors Degree	31%	38%
Post Graduate Degree	16%	12%
Employment Status		
Employed	54%	67%
Unemployed	8%	10%
Out of Labor Force	14%	11%
Online Worker	6%	10%
Retired	18%	2%
Income		
Less than $20K	17%	32%
Between $20K and $30K	14%	16%
Between $30K and $50K	19%	23%
Between $50K and $70K	19%	13%
Between $70K and $150K	25%	14%
More than $150K	6%	2%
Marital Status		
Single	32%	50%
Partnered	53%	42%
Seperated / Divorced / Widowed	14%	9%
N	1,000	995

Table A.2: Percent of Variance due to Measurement Error in Different Samples.

	Spring 2015 CCS				
	Everyone	Participant	Weighted Participant	SSEL Participant	SSEL Participant (In Lab)
Risky Projects	43%	48%	47%	41%	45%
	(2.9%)	(4.6%)	(6.6%)	(8.3%)	(8.6%)
Risky Urns	25%	21%	19%	24%	52%
	(2.3%)	(3.3%)	(6.5%)	(6.7%)	(9.0%)
Dictator Giving	15%	15%	14%	18%	23%
	(1.8%)	(2.8%)	(4.4%)	(5.9%)	(6.5%)
IAT Race	42%	41%	46%	30%	40%
	(2.8%)	(4.3%)	(12.4%)	(7.3%)	(8.3%)
IAT Gender	39%	40%	37%	54%	58%
	(2.8%)	(4.3%)	(10.7%)	(9.2%)	(9.4%)
N	819	350	350	96	96

Figure A.1: Distribution of responses in representative sample, MTurk, and CCS

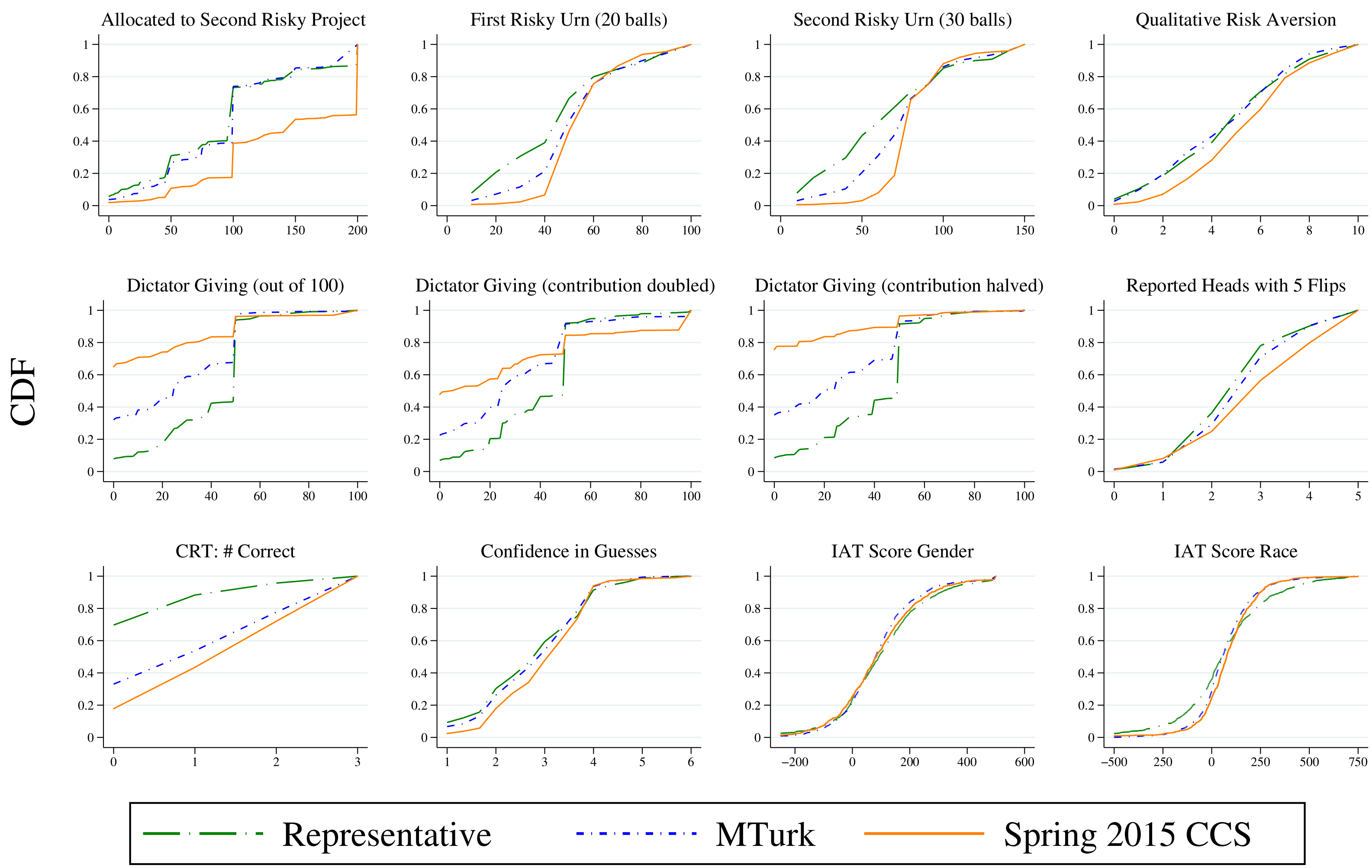


Figure A.2: Correlations across the representative sample, MTurk, and CCS (5% level).

	Risk Aversion	Discounting (δ)	Dictator	Prisoner's Dilemma	Lying	Cognitive	Confidence	Compete	IAT Race	IAT Gender	Male
Risk Aversion		0+−	+0+	0	−0−	00−	−	−	0	0	−
Discounting (δ)	0+−		0−0	0	0+0	+	0−0	0	0	00+	+00
Dictator	+0+	0−0		−	−	−−0	0+0	00−	0	0	0−0
Prisoner's Dilemma	0	0	−		0++	+	00+	0++	0	0	00+
Lying	−0−	0+0	−	0++		0++	+0+	0++	0	0	0++
Cognitive	00−	+	−−0	+	0++		−−+	00+	00−	0	+
Confidence	−	0−0	0+0	00+	+0+	−−+		+	0	00+	0++
Compete	−	0	00−	0++	0++	00+	+		0+0	0++	0++
IAT Race	0	0	0	0	0	00−	0	0+0		+	0
IAT Gender	0	00+	0	0	0	0	00+	00+	+		00+
Male	−	+00	0−0	00+	0++	+	0++	0++	0	00+	

Notes: indicates complete agreement, complete disagreement, and two out of three samples agreeing.

Figure A.3: Correlations across the representative sample, MTurk, and CCS (1% level).

	Risk Aversion	Discounting (δ)	Dictator	Prisoner's Dilemma	Lying	Cognitive	Confidence	Compete	IAT Race	IAT Gender	Male
Risk Aversion		0+−	0	0	−0−	00−	−	0−−	0	0	0−−
Discounting (δ)	0+−		0−0	0	0	+	0	0	0	0	+00
Dictator	0	0−0		−	−	0−0	0+0	00−	0	0	0−0
Prisoner's Dilemma	0	0	−		0++	++0	0	0	0	0	00+
Lying	−0−	0	−	0++		0+0	+0+	0++	0	0	0++
Cognitive	00−	+	0−0	++0	0+0		0−0	00+	0	0	+
Confidence	−	0	0+0	0	+0+	0−0		0++	0	00+	0++
Compete	0−−	0	00−	0	0++	00+	0++		0	00+	0++
IAT Race	0	0	0	0	0	0	0	0		+	0
IAT Gender	0	0	0	0	0	0	00+	00+	+		00+
Male	0−−	+00	0−0	00+	0++	+	0++	0++	0	00+	

Notes: [light grey] indicates complete agreement, [dark grey] complete disagreement, and [medium grey] two out of three samples agreeing.

Figure A.4: Distribution of responses in the CCS

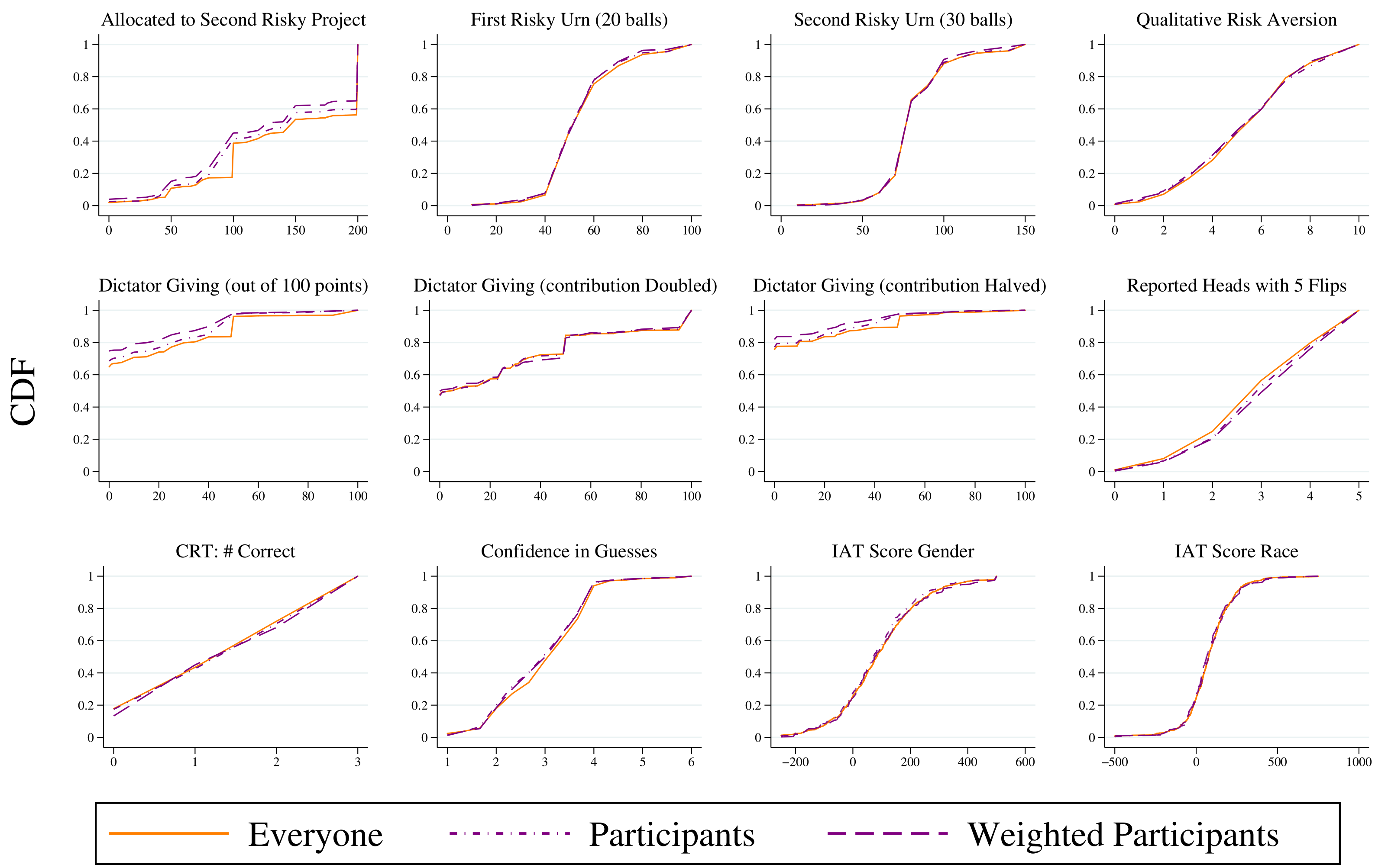


Figure A.5: Correlations across everyone, participants, and weighted participants (5% level)

	Risk Aversion	Discounting (δ)	Dictator	Prisoner's Dilemma	Lying	Cognitive	Confidence	Compete	IAT Race	IAT Gender	Male
Risk Aversion		−	+00	−00	−00	−00	−	−	0	0	−
Discounting (δ)	−		0	0	0	+	0	0	0	0	0
Dictator	+00	0		−	−	0−−	0−−	−0−	00+	0	−00
Prisoner's Dilemma	−00	0	−		+	0	+	0	0	0	+
Lying	−00	0	−	+		+00	+00	+	0	0++	+00
Cognitive	−00	+	0−−	0	+00		+0+	+0+	−00	0	+
Confidence	−	0	0−−	+	+00	+0+		+	0	+00	+
Compete	−	0	−0−	0	+	+0+	+		0	+00	++0
IAT Race	0	0	00+	0	0	−00	0	0		+	0
IAT Gender	0	0	0	0	0++	0	+00	+00	+		+
Male	−	0	−00	+	+00	+	+	++0	0	+	

Notes: indicates complete agreement, complete disagreement, and two out of three samples agreeing.

Figure A.6: Correlations across everyone, participants, and weighted participants (1% level)

	Risk Aversion	Discounting (δ)	Dictator	Prisoner's Dilemma	Lying	Cognitive	Confidence	Compete	IAT Race	IAT Gender	Male
Risk Aversion		−−0	0	0	−00	0	−00	−	0	0	−
Discounting (δ)	−−0		0	0	0	++0	0	0	0	0	0
Dictator	0	0		−	−	00−	0−0	−0−	0	0	0
Prisoner's Dilemma	0	0	−		+	0	0++	0	0	0	+
Lying	−00	0	−	+		0	+00	+	0	0	+00
Cognitive	0	++0	00−	0	0		0	+00	0	0	+
Confidence	−00	0	0−0	0++	+00	0		+	0	+00	+
Compete	−	0	−0−	0	+	+00	+		0	+00	++0
IAT Race	0	0	0	0	0	0	0	0		+	0
IAT Gender	0	0	0	0	0	0	+00	+00	+		+
Male	−	0	0	+	+00	+	+	++0	0	+	

Notes: [light grey] indicates complete agreement, [dark grey] complete disagreement, and [medium grey] two out of three samples agreeing.

Figure A.7: Correlations on the CCS and in the Lab (5% level)

	Risk Aversion	Discounting (δ)	Dictator	Prisoner's Dilemma	Lying	Cognitive	Confidence	Compete	IAT Race	IAT Gender	Male
Risk Aversion		0−	0	0	0	0	0	−0	0	0	0−
Discounting (δ)	0−		0	−0	0	0	0	0	0	0	0
Dictator	0	0		−	−	0	0	0	0	0	0
Prisoner's Dilemma	0	−0	−		+	0	0	0	0	0	0+
Lying	0	0	−	+		0	0+	0	0	0	0
Cognitive	0	0	0	0	0		0	0	0	+0	+
Confidence	0	0	0	0	0+	0		0+	0	0	0+
Compete	−0	0	0	0	0	0	0+		0	0	0
IAT Race	0	0	0	0	0	0	0	0		+0	0
IAT Gender	0	0	0	0	0	+0	0	0	+0		+
Male	0−	0	0	0+	0	+	0+	0	0	+	

Notes: [light grey] indicates complete agreement, [dark grey] complete disagreement, and [medium grey] two out of three samples agreeing.

Table A.3: Response time to CCS solicitation is not indicative of measured behaviors.

	Samples				Differences	
	Everyone (E)	One Email	One Week (W)	More Than One Week (M)	E−W	E−M
First Risky Project	59	59	59	61	0.74	−1.4
(out of 100)	(1.2)	(1.8)	(1.5)	(2.1)	(1.9)	(2.4)
Second Risky Project	143	141	142	145	1.2	−2.2
(out of 200)	(2.1)	(3.0)	(2.6)	(3.5)	(3.3)	(4.1)
First Risky Urn	59	59	59	60	0.30	−0.56
(20 balls)	(.52)	(.73)	(.64)	(.88)	(.82)	(1.0)
Second Risky Urn	86	86	86	86	−0.01	0.02
(30 balls)	(.73)	(1.0)	(.89)	(1.3)	(1.2)	(1.5)
Qualitative Risk Aversion	5.8	5.7	5.7	6.0	0.10	−0.18
	(.08)	(.12)	(.10)	(.13)	(.12)	(.15)
Discounting (δ)	0.77	0.77	0.76	0.78	0.00	−0.01
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
First Dictator Game	14	14	14	15	0.35	−0.65
(given out of 100)	(.84)	(1.3)	(1.1)	(1.4)	(1.4)	(1.6)
Second Dictator Game	38	37	38	38	0.03	−0.05
(given out of 300)	(2.4)	(3.5)	(3.0)	(3.9)	(3.8)	(4.6)
Dictator, Tokens Given	26	27	27	25	−0.75	1.4
are Doubled	(1.2)	(1.8)	(1.5)	(1.9)	(1.9)	(2.3)
Dictator, Tokens Given	9.0	8.8	8.8	9.3	0.20	−0.36
are Halved	(.68)	(1.0)	(.85)	(1.2)	(1.1)	(1.3)
Prisoner's Dilemma	68	69	69	66	−0.76	1.39
(% dominant strat.)	(1.5)	(2.2)	(1.9)	(2.5)	(2.4)	(2.9)
Reported Heads	3.3	3.3	3.3	3.3	0.01	−0.01
(out of 5)	(.04)	(.06)	(.05)	(.07)	(.07)	(.08)
Reported Switches	5.5	5.5	5.5	5.4	−0.03	0.05
(out of 9)	(.07)	(.10)	(.09)	(.12)	(.11)	(.14)
Raven's Matrices	1.8	1.9	1.8	1.8	−0.01	0.02
(out of 5)	(.04)	(.07)	(.05)	(.08)	(.07)	(.09)
CRT	1.7	1.8	1.7	1.6	−0.05	0.09
(out of 3)	(.04)	(.06)	(.05)	(.06)	(.06)	(.07)
Confidence in Guesses	3.1	3.1	3.1	3.2	0.02	−0.04
	(.03)	(.05)	(.04)	(.05)	(.05)	(.06)
Competition	33	29	31	37	2.1	−3.91
(% competing)	(1.7)	(2.4)	(2.0)	(2.9)	(2.6)	(3.3)
IAT Race	81	83	82	80	−0.91	1.7
	(5.6)	(8.4)	(6.8)	(9.9)	(8.8)	(11)
IAT Gender	95	81	84	115	11	−20
	(5.9)	(8.6)	(6.9)	(10.8)	(9.1)	(12)
Percent Male	62	60	60	65	2.0	−3.7
	(1.7)	(2.5)	(2.1)	(2.8)	(2.7)	(3.3)
N	819	374	530	289		

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses.

Table A.4: Those that wait more than a week to participate are less likely to go to the lab.

	Samples				Differences	
	Everyone (E)	One Email	One Week (W)	More Than One Week (M)	E−W	E−M
Percent Lab Participant	43	47	47	35	−4.4	8.1**
	(1.7)	(2.6)	(2.2)	(2.8)	(2.8)	(3.3)
Avg. Lab Sessions	1.3	1.5	1.5	0.85	−0.24	0.44***
	(.09)	(.15)	(.12)	(.11)	(.15)	(.14)
N	819	374	530	289		

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses.

Figure A.8: Distribution of responses in the Spring 2015 survey vs. the Lab ($N = 96$)

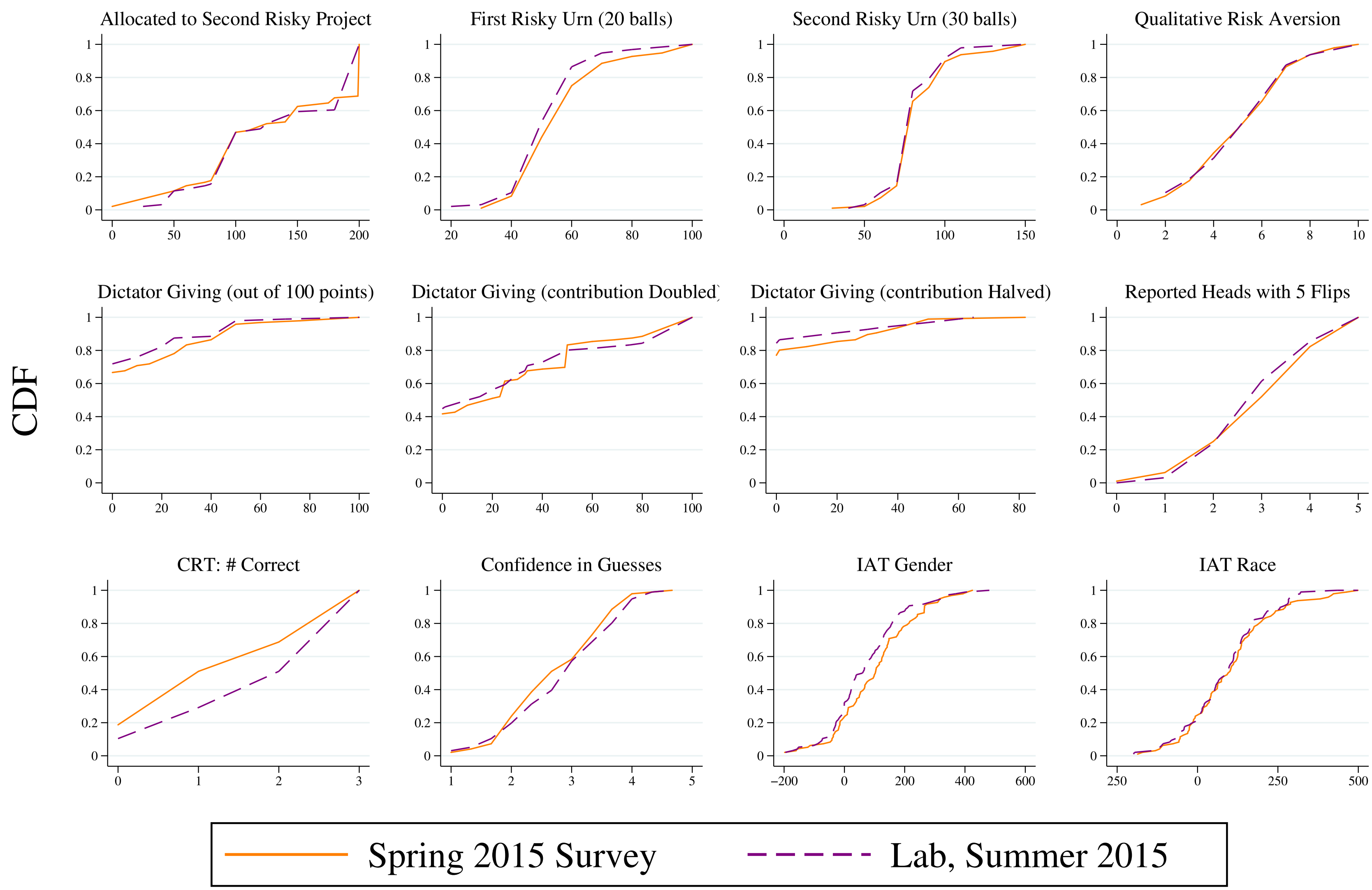


Table A.5: There are few significant differences based on the amount of lab participation.

	By Partcipation			Differences		
	All Participants (P)	Below Median (B)	Above Median (A)	P−B	P−A	B−A
First Risky Project	55	57	52	−1.9	2.4	4.4
(out of 100)	(1.8)	(2.5)	(2.7)	(3.1)	(3.2)	(3.6)
Second Risky Project	139	144	132	−5.1	6.4	11.5*
(out of 200)	(3.2)	(4.2)	(4.9)	(5.2)	(5.9)	(6.4)
First Risky Urn	58	58	58	0.11	−0.14	−0.25
(20 balls)	(.77)	(1.1)	(1.0)	(1.4)	(1.3)	(1.5)
Second Risky Urn	86	86	86	0.05	−0.06	−0.10
(30 balls)	(1.1)	(1.6)	(1.4)	(2.0)	(1.8)	(2.2)
Qualitative Risk Aversion	5.7	5.7	5.8	0.07	−0.09	−0.16
	(.12)	(.17)	(.18)	(.21)	(.22)	(.25)
Discounting (δ)	0.78	0.78	0.77	0.00	0.00	0.01
	(.01)	(.01)	(.02)	(.02)	(.02)	(.02)
First Dictator Game	12	14	9.4	−1.9	2.3	4.2*
(given out of 100)	(1.1)	(1.6)	(1.6)	(2.0)	(2.0)	(2.3)
Second Dictator Game	32	36	26	−4.4	5.5	9.8
(given out of 300)	(3.2)	(4.5)	(4.4)	(5.5)	(5.4)	(6.2)
Dictator, Tokens Given	26	27	26	−0.28	0.35	0.63
are Doubled	(1.8)	(2.5)	(2.7)	(3.1)	(3.2)	(3.6)
Dictator, Tokens Given	7.8	9.2	5.9	−1.5	1.8	3.3*
are Halved	(.94)	(1.4)	(1.2)	(1.7)	(1.5)	(1.8)
Prisoner's Dilemma	67.1	65.9	68.7	1.3	−1.6	−2.8
(% dominant strat.)	(2.3)	(3.1)	(3.5)	(3.9)	(4.2)	(4.7)
Reported Heads	3.4	3.3	3.5	0.09	−0.11	−0.20
(out of 5)	(.06)	(.08)	(.10)	(.11)	(.11)	(.13)
Reported Switches	5.5	5.4	5.6	0.10	−0.13	−0.23
(out of 9)	(.11)	(.15)	(.17)	(.18)	(.20)	(.22)
Raven's Matrices	1.8	1.8	1.8	−0.02	0.02	0.04
(out of 5)	(.07)	(.09)	(.10)	(.11)	(.12)	(.14)
CRT	1.7	1.7	1.7	0.00	0.00	0.00
(out of 3)	(.06)	(.08)	(.09)	(.10)	(.10)	(.12)
Confidence in Guesses	3.1	3.0	3.1	0.02	−0.03	−0.05
	(.05)	(.07)	(.07)	(.08)	(.08)	(.10)
Competition	34	34	33	−0.64	0.81	1.5
(% competing)	(2.5)	(3.4)	(3.8)	(4.3)	(4.6)	(5.1)
IAT Race	87	90	83	−3.2	4.0	7.1
	(8.5)	(12)	(11)	(15)	(14)	(17)
IAT Gender	85	73	100	12	−15	−27
	(8.5)	(10)	(14)	(13)	(17)	(17)
Percent Male	55	57	54	−1.5	1.9	3.4
	(2.7)	(3.6)	(4.0)	(4.4)	(4.8)	(5.4)
N	350	195	155			

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses.

Table A.6: Re-weighting the CCS to be more demographically representative does not change conclusions.

	Weightings			Differences	
		Re-Weighted			
	Unweighted (U)	Gender (G)	Race (R)	U−G	U−R
First Risky Project (out of 100)	59 (1.2)	57 (1.9)	64 (2.5)	2.3 (2.3)	−4.5 (2.8)
Second Risky Project (out of 200)	143 (2.1)	138 (3.2)	155 (4.2)	4.5 (3.8)	-12^{**} (4.7)
First Risky Urn (20 balls)	59 (.52)	59 (.84)	59 (1.1)	0.14 (.99)	0.46 (1.2)
Second Risky Urn (30 balls)	86 (.73)	86 (1.2)	85 (1.5)	0.17 (1.4)	0.87 (1.7)
Qualitative Risk Aversion	5.8 (.08)	5.7 (.12)	5.8 (.16)	0.09 (.15)	−0.01 (.18)
Discounting (δ)	0.77 (.01)	0.76 (.01)	0.81 (.01)	0.00 (.01)	-0.04^{**} (.02)
First Dictator Game (given out of 100)	14 (.84)	14 (1.4)	16 (1.7)	−0.31 (1.6)	−2.5 (1.9)
Second Dictator Game (given out of 300)	38 (2.4)	39 (3.9)	44 (4.9)	−0.80 (4.5)	−5.8 (5.4)
Dictator, Tokens Given are Doubled	26 (1.2)	26 (2.0)	33 (2.4)	0.08 (2.3)	-6.5^{**} (2.7)
Dictator, Tokens Given are Halved	9.0 (.68)	9.7 (1.1)	9.0 (1.5)	−0.73 (1.3)	−0.05 (1.6)
Prisoner's Dilemma (% dominant strat.)	68 (1.5)	67 (2.4)	66 (3.1)	0.99 (2.9)	2.3 (3.4)
Reported Heads (out of 5)	3.3 (.04)	3.3 (.07)	3.2 (.09)	0.02 (.08)	0.10 (.10)
Reported Switches (out of 9)	5.5 (.07)	5.4 (.11)	5.4 (.15)	0.06 (.13)	0.09 (.16)
Raven's Matrices (out of 5)	1.8 (.04)	1.8 (.07)	1.8 (.09)	0.02 (.08)	−0.04 (.10)
CRT (out of 3)	1.7 (.04)	1.6 (.06)	1.7 (.08)	0.07 (.07)	−0.06 (.09)
Confidence in Guesses	3.1 (.03)	3.1 (.05)	3.1 (.07)	0.05 (.06)	0.00 (.07)
Competition (% competing)	33.46 (1.7)	31 (2.6)	33 (3.4)	2.4 (3.1)	0.34 (3.8)
IAT Race	81 (5.6)	83 (9.2)	90 (12)	−1.1 (11)	−8.7 (13)
IAT Gender	95 (5.9)	83 (9.2)	93 (12)	12 (11)	1.7 (13)

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses.