# Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study*

**Ben Gillen**
California Institute of Technology
bgillen@caltech.edu
hss.caltech.edu/~bgillen/

**Erik Snowberg**
California Institute of Technology and NBER
snowberg@caltech.edu
hss.caltech.edu/~snowberg/

**Leeat Yariv**
California Institute of Technology
lyariv@hss.caltech.edu
hss.caltech.edu/~lyariv/

August 23, 2015

**Abstract**

Measurement error is ubiquitous in experimental work. It leads to imperfect statistical controls, attenuated estimated effects of elicited behaviors, and biased correlations between characteristics. We develop simple statistical techniques for dealing with experimental measurement error. These techniques are applied to data from the Caltech Cohort Study, which conducts repeated incentivized surveys of the Caltech student body. We illustrate the impact of measurement error by replicating three classic experiments, and showing that results change substantially when measurement error is taken into account. Collectively, these results show that failing to properly account for measurement error may cause a field-wide bias: it may lead scholars to identify "new" effects and phenomena that are actually similar to those previously documented.

JEL Classifications: C81, C9, D8, J71
Keywords: Measurement Error, Experiments, ORIV, Competition, Risk, Ambiguity

*Snowberg gratefully acknowledges the support of NSF grants SES-1156154 and SMA-1329195. Yariv gratefully acknowledges the support of NSF grant SES-0963583 and the Gordon and Betty Moore Foundation grant 1158. We thank Marco Castillo, Yoram Halevy, Muriel Niederle, and Lise Vesterlund for comments and suggestions, as well as seminar audiences at Caltech, SITE, and UBC.

# 1 Introduction

Measurement error is ubiquitous in experimental work. Lab elicitations of attitudes are subject to random variation in participants' attention and focus, as well as rounding due to finite choice menus. Moreover, there is an imperfect link between elicited proxies and the attitudes they intend to capture. Despite the ubiquity of measurement error, the experimental literature offers only a small set of tools for dealing with it—improved elicitation techniques and multiple rounds. Instead, we focus on developing statistical techniques.

At the heart of our approach is the combination of duplicate elicitations (usually two) of behavioral proxies and methods from the econometrics literature. While multiple elicitations would be impossible for a researcher using, say, the Current Population Survey, in experimental economics they are very easy to obtain.

The statistical tools we develop deal with three types of inference breakdowns that arise from different uses of experimental proxies measured with error: as controls, as causal variables, or to estimate correlations between latent preference characteristics. We demonstrate the potential perils of measurement error, and the effectiveness of our techniques, using a unique new data set tracking behavioral proxies of the entire Caltech undergraduate student body, the Caltech Cohort Study. In particular, we replicate several classic results: on gender differences in competition, on risk elicitation techniques, and on the relationship between ambiguity aversion and attitudes toward compound lotteries. Correcting for measurement error alters previously accepted conclusions in all three of the experiments we have examined. We show that gender differences in competition are driven by differences in risk aversion and overconfidence, that several risk-aversion measures are fairly closely related, and that there is little difference between reactions to compound lotteries and to ambiguous lotteries.

Collectively, these results show that failing to properly account for measurement error may cause a field-wide bias. In particular, it may lead scholars to identify "new" effects and phenomena that are actually similar to those previously documented.

## 1.1 Simulated Examples

To understand the inferential dangers posed by measurement error, we present two simulated examples. In the first, a researcher is interested in estimating the effects of a variable $D$—say, gambling—on some outcome variable $Y$—say, participation in dangerous sports—using an experimentally measured variable $X$—say, elicited risk attitudes—as a control. The model that we use to simulate data is

$$Y^* = 0 \times D + X^* \quad \text{with} \quad D = 0.5 \times X^* + \eta \quad \text{and} \quad X = X^* + \nu \tag{1}$$

where $\varepsilon \sim \mathcal{N}[0, 1]$, $\eta \sim \mathcal{N}[0, 0.9]$, $X^* \sim \mathcal{N}[0, 1]$, and $\nu \sim \mathcal{N}[0, \sigma_\nu^2]$. That is, risk attitudes drive both gambling and participation in dangerous sports, but that attitude is measured through a lab-based elicitation technique that contains error. We assume the researcher only has access to $Y = Y^* - \varepsilon$, a noisy measure of $Y^*$.

A diligent researcher would fit a regression model of the form

$$Y = \alpha D + \beta X + \varepsilon \tag{2}$$

hoping to control for the role of risk attitudes in the effect of gambling on participation in dangerous sports. Table 1 shows, from simulations, how the estimates, $\hat{\alpha}$ and $\hat{\beta}$, depend on how much measurement error there is in the variance of $X$, that is $\frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_{X^*}^2}$.

The estimated coefficients depend strongly on the amount of measurement error in $X$. With $N = 100$—a typical sample size for an experiment—the coefficient on gambling $D$ becomes statistically significant when measurement error reaches approximately 1/3 of the variance in $X$. To put this in perspective, we estimate from our data that measurement error accounts for 30–40% of the variance of elicited proxies for risk attitudes (see (4) and surrounding text).

Depressingly, adding more observations does nothing to reduce the bias in the estimated

Table 1: Simulated regressions of (2), with controls $X$ measured with error. True model: $\alpha = 0$, $\beta = 1$.

| Error as a percent of Var[$X$]: | 0 | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|
| Panel A: N=100 | | | | | | |
| $\hat{\alpha}$ | 0.00 | 0.06 | 0.11 | 0.16 | 0.21* | 0.26*** |
| | (0.11) | (0.11) | (0.12) | (0.12) | (0.12) | (0.12) |
| $\hat{\beta}$ | 1.00*** | 0.87*** | 0.75*** | 0.64*** | 0.54*** | 0.44*** |
| | (0.12) | (0.11) | (0.11) | (0.10) | (0.10) | (0.09) |
| Panel B: N=1,000 | | | | | | |
| $\hat{\alpha}$ | 0.00 | 0.06* | 0.11*** | 0.16*** | 0.21*** | 0.26*** |
| | (0.03) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) |
| $\hat{\beta}$ | 1.00*** | 0.87*** | 0.75*** | 0.64*** | 0.54*** | 0.43*** |
| | (0.04) | (0.04) | (0.03) | (0.03) | (0.03) | (0.03) |

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses. Coefficients and standard errors are averages from 10,000 simulated regressions.

coefficients. In fact, when $N = 1,000$, the approximate size of the CCS, the coefficient on gambling $\hat{\alpha}$ appears statistically significant when measurement error accounts for only 10–15% of the variance in $X$. This illustrates that, although the techniques we will suggest rely on adding data, the way in which data is added is critical.

Note also that if the researcher was instead interested in the magnitude and significance of $\beta$, which captures the effect of $X^*$ on $Y$, she would erroneously underestimate the true effect. Indeed, for $N = 100$, when measurement error accounts for only 20% of the variance in $X$, the estimated coefficient $\hat{\beta}$ is significantly less than 1. This issue can only be solved by removing improper controls from the regression.

The problem of measurement error biasing coefficients is particularly acute when researchers estimate correlations between $X$ and $Y$ (keeping in mind that (1) implies $X^* = Y^*$). For the simulated results in Table 2, we simultaneously vary the proportion of measurement

Table 2: Correlations with $X$ and $Y$ measured with error. True model: $\text{Corr}[X^*, Y^*] = 1$.

| Error as a percent of Var[$X$] and Var[$Y$]: | 0 | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|
| Panel A: N=100 | | | | | | |
| $\widehat{\text{Corr}}[X, Y]$ | 1.00 (0.00) | 0.90*** (0.02) | 0.80*** (0.04) | 0.70*** (0.05) | 0.60*** (0.06) | 0.50*** (0.08) |
| Panel B: N=1,000 | | | | | | |
| $\widehat{\text{Corr}}[X, Y]$ | 1.00 (0.00) | 0.90*** (0.01) | 0.80*** (0.01) | 0.70*** (0.02) | 0.60*** (0.02) | 0.50*** (0.02) |

Notes: ***, **, * denote statistically significantly different from 1 at the 1%, 5%, and 10% level, with standard errors in parentheses. Coefficients and standard errors are averages from 10,000 simulated regressions.

error in both $X$ and $Y$. The results speak for themselves: even a bit of measurement error causes significant deviations from the proper correlation of one. Keeping in mind that the scale of measurement error in experimental elicitations of proxies is on the order of 30–40%, it is extremely unlikely that one would ever estimate a correlation close to one, even if that were the correct relationship.[1]

## 1.2 Techniques

When control variables are measured with error, our solution is simple—elicit more controls. Naturally, with the moderate-size data sets used in experiments, using many controls can be problematic. Thus, we further show how principal component analysis allows for the use of a small, but informative, set of controls. This idea is applied, in Section 3, to show that the gender gap in competitiveness can be explained by risk attitudes and overconfidence, although Niederle and Vesterlund (2007) concluded that it was a disjoint phenomenon. That is, in explaining competition, the coefficient on gender becomes insignificant when multiple

[1]Note that the standard errors are less in Table 2, as Var[$\varepsilon$], set to 1 in all columns of Table 1, now varies across the columns: starting at 0 and climbing to 1 in the final column.

controls for risk attitudes and overconfidence are used.

We use a different approach when estimating causal effects or correlations, drawing inspiration from instrumental variables. Our approach, which we call *Obviously Related Instrumental Variables* (ORIV) uses duplicate elicitations of $X$ and $Y$ as instruments. Specifically, we obtain duplicate measures of $X$, denoted $X^a$ and $X^b$, which are both proxies for $X^*$ that are measured with error. By regressing $X^a$ and $X^b$, we extract the information contained in $X^b$ that can explain $X^a$. If the measurement error in the two elicitations is orthogonal—as we assume—then the resulting predicted values $\hat{X}^a(X^b)$ contain only information about $X^*$. We then use a stacked regression to combine the information from both $\hat{X}^a(X^b)$ and $\hat{X}^b(X^a)$, resulting in an efficient use of the data.

This technique is easily extended to allow for multiple measures of the outcome $Y$. This is particularly useful in estimating correlations, where there is no clear distinction between outcome and explanatory variables, and measurement error in either can attenuate estimates.

ORIV produces consistent coefficients, correlations, and standard errors. This technique is applied, in Section 4, to show that various risk elicitation methods are more correlated than previously thought, and that the patterns of correlations between them are indicative of phenomena outside the lab. We further use ORIV to show, in Section 5, that ambiguity aversion and reaction to compound lotteries are very close to perfectly correlated—once we account for measurement error. This leads us to conclude, in Section 6 that failing to correct for measurement error has lead the field to "over-identify" new phenomena.

## 1.3 Related Literature

Mis-measurement of data has been an important concern for statisticians and econometricians since the late $19^{\underline{th}}$ century (Adcock, 1878). Indeed, estimating the relationship between two variables when both are measured with error is a foundational problem in the statistics literature (Frisch, 1934; Koopmans, 1939; Wald, 1940). The use of instrumental variables

to address the classical errors-in-variables problem was proposed by Reiersøl (1941, 1945, 1950), with notable developments by Durbin (1954) and Sargan (1958) (see Hausman, 2001, for a review). These techniques were first applied to economic problems by Friedman (1957), who estimated consumption functions, and noted that annual income is a noisy measure of permanent income, which could attenuate estimates of the marginal propensity to consume from permanent income.[2] Since then, instrumental variables have been used to control for measurement error in an array of empirical applications including medicine (Carroll and Stefanski, 1994), psychology (Fiske et al., 2010), and epidemiology (Greenland, 2000).

There are very few instances in which measurement error played an explicit role in the analysis of experimental data. An early example, Battalio et al. (1973), shows that even small reporting errors can lead to a rejection of the generalized axiom of revealed preferences. Following this work, some scholars argued for a "theory of errors" under which observed violations of expected utility are an artifact of human error (Hey, 1991).

Recent experimental papers have taken a renewed interest in the problems caused by measurement error. In an innovative approach, Castillo et al. (2015), posit a structural model of measurement error, following Harless and Camerer (1994). They use several risk elicitation methods to study the effects of risk attitudes, accounting for measurement error, on disciplinary referrals of children.[3] Coffman and Niehaus (2015) adjust for measurement error in self-interest and other-regard by projecting both on a common set of explanatory variables. Blattman et al. (2015), in a field setting, focuses on gaining the trust of respondents in order to quantify the amount of measurement error in responses to sensitive questions. More often than not, however, experimental scholars limit their efforts at reducing the effects of measurement error through repetition of tasks across multiple rounds of an experiment.[4]

[2]For a review of the history and applications of instrumental variables more generally, see Angrist and Krueger (2001).

[3]Similarly, *Quantal Response Equilibrium* posits a structural model in which agents make mistakes that are inversely related to the payoff losses they generate (McKelvey and Palfrey, 1995, 1998).

[4]There is extensive literature dealing with measurement error in survey data, see, for example, Bertrand and Mullainathan (2001), Bound et al. (2001), and references therein. This literature acknowledges some of the issues we discuss, but offers limited techniques for overcoming them.

Our paper is, to our knowledge, the first to offer simple, yet general, experimental techniques for mitigating the effects of measurement error.

## 2 The Caltech Cohort Study

Caltech is an independent, privately supported university located in Pasadena, California. It has approximately 900 undergraduate students, of which $\sim 40\%$ are women.

In the Fall of 2013, 2014, and Spring of 2015, we administered an incentivized, online survey to the entire undergraduate student body. We used incentivized tasks to elicit an array of attributes, including: risk aversion, ambiguity aversion, competitiveness, cognitive sophistication, implicit attitudes toward gender and race, generosity, honesty, overconfidence, overprecision, and optimism. Students were also asked a large set of questions addressing their lifestyle and social habits: sleep patterns, study routines, social networks, study networks, physical attributes, and so on.[5]

The data used in this paper comes from the Fall 2014 and Spring 2015 installments. In the Fall of 2014, 92% of the entire student body (893/972) responded to the survey. Of those, 39% were female (349/893), and the average payment was \$24.34. In the Spring of 2015, 91% of the entire student body (819/899) responded to the survey. Of those, 39% were female (322/819), and the average payment was \$29.08. The difference in average payments across years was due to the inclusion of several additional incentivized items in 2015.[6] Of those who had taken the survey in 2015, 96% (786/819) also took the survey in 2014. As Section 4 requires data from both surveys, we use this subsample throughout for consistency.[7]

Unlike in most experimental settings, there is little concern about self-selection into our

[5]For screenshots of the 2015 survey, go to: people.hss.caltech.edu/~lyariv/ScreenshotsSpring2015.pdf.

[6]The number of overall students was substantially less in the Spring of 2015, as about 50 students departed the institute due to hardship or early graduation. Further, we did not approach students who had spent more than four years at Caltech, accounting for approximately 25 students.

[7]In the Fall of 2013 88% of the student body (806/916) responded to the survey, of which 38.5% (310/806) were female. The average payment was \$20.58. Of those who took the survey in 2013 and did not graduate, 89% (546/615) also took the survey in the Fall of 2014.

experiments from the subject population, due to our 90%+ response rates (Cleave et al., 2013; Falk et al., 2013; Harrison et al., 2009). However, Caltech is highly selective, which may cause one to worry that the overall population is different from the pool used in most lab experiments. Nonetheless, three points should mitigate such concerns. First, the raw results of the replications performed in this paper yield virtually identical results to those reported in the original papers.[8] Second, responses from our survey to several standard elicitations—of risk, altruism in the dictator game, etc.—are similar to those reported from several other pools (see Appendix D for details). Third, 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 our subject pool is more "special" than that used in many other lab experiments. As the results reported in this paper are replications of other studies, these points suggest that our results are likely due to our more sophisticated treatment of measurement error, rather than an artifact of the subject population.

## 2.1 Measures Used

Our results deal with a subset of the measured attributes, which we detail here. Question wordings can be found in Appendix E. Throughout, 100 survey tokens were valued at $1.

### 2.1.1 Overconfidence

We break overconfidence into three categories, following Moore and Healy (2008). These measures are used in Section 3 as controls.

**Overestimation and Overplacement:** Participants complete two tasks: a five-question cognitive reflection test (CRT; see Frederick, 2005), and five Raven's matrices (Raven, 1936).

[8]This should increase confidence in the original studies that we replicate, as it implies that it is not participants' self-selection into the lab that is driving the results in those studies.

Participants are given a maximum of 20 seconds per CRT item, and 30 seconds per Raven's matrix. After each task, participants are asked how many questions they believe they answered correctly. This, minus the participants' true performance, gives a measure of overestimation. Participants are also asked where they think they are in the performance distribution of all participants. This, minus the participants' true percentile, gives a measure of overplacement. In total, this gives three co-linear measures: performance, expected performance, and overconfidence. Two of these three can be used independently to control for confidence or overconfidence.

**Overprecision:** Participants are shown a random picture of a jar of jellybeans, and asked to guess how many jellybeans the jar contains. They are then asked—on a six point qualitative scale from "Not confident at all" to "Certain"—how confident they are of their guess. This is repeated three times. Following Ortoleva and Snowberg (2015), each of these measures is interpreted as a measure of overprecision.

**Perception of Academic Performance:** A final measure of overconfidence asks participants to state where in the grade distribution of their entering cohort they believe they would fall over the next year. This is treated as a measure of confidence in placement.

### 2.1.2 Risk

Risk measures are used in Section 3 as controls, and in Section 4 as an outcome of interest. Further, the Risk MPL described below is used as an outcome of interest in Section 5.[9]

**Projects:** Following Gneezy and Potters (1997), participants are asked to allocate 200 tokens between a safe option (keeping them), and a project that returns some multiple of the tokens with probability $p$, otherwise returning nothing. In Fall 2014, two projects were used: the first returning 3 tokens per token invested where $p = 40\%$ of the time, and the

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

second returning 2.5 tokens 50% of the time. In the Spring of 2015, the first project was modified to return 3 tokens 35% of the time.

**Qualitative:** Following Dohmen et al. (2011), participants are asked to rate themselves, on a scale of 0–10, in terms of how willing they are to take risks. As this question was asked once in the Fall of 2014, the elicitation from the Spring of 2015 is used as a secondary measure in Section 4.

**Lottery Menu:** Following Eckel and Grossman (2002), participants are asked to choose between six 50/50 lotteries with different stakes.[10] The first lottery contained the same payoff in each state, and thus corresponded to a sure amount. The second through sixth lotteries contained increasing amounts in the first state, and decreasing amounts in the second state.

**Risk MPL:** Participants respond to two Multiple Price Lists (MPLs) that ask them to choose between a lottery over a draw from an urn, and sure amounts. The lottery would pay off if a ball of the color of the participant's choosing 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 are first asked to choose the color (red or black) that they want to pay off, if drawn. They are then presented with a list of choices between a certainty equivalent that increases in units of 10 tokens from 0 to 100 or the gamble on the urn.[11] Both measures were elicited on the Fall 2014 and Spring 2015 surveys.

### 2.1.3 Ambiguity and Compound Lotteries

Reactions to ambiguous and compound lotteries are considered in Section 5.

[10] The variant we use comes from Dave et al. (2010).

[11] 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. Additionally, participants needed only to make one choice and all other rows were automatically filled in to be consistent with that choice.

**Compound MPL:** This follows the same protocol as the Risk MPLs described above, except participants are told that the number of red balls would be uniformly drawn between 0 and 20 for the first urn, and between 0 and 30 for the second. As this is a measure of risk attitudes, it is also used as a control in Section 4.

**Ambiguous MPL:** This elicitation emulates the standard (Ellsberg, 1961) urn. It follows the same protocol as the two other MPLs. Participants were informed that the composition of the urn was chosen by a Caltech administrator, Dean John Dabiri.

To reduce instructions, both of the MPLs for a given attitude (Risk, Compound, Ambiguity) are run sequentially, in random order. These three blocks are spread across the survey, and it is randomly determined which block is given first, second, or third. As no order effects were observed, we aggregate results across the different possible orderings.

# 3 Controls Measured with Error

To make the claim that an estimated effect is independent of other factors, many studies attempt to control for those other factors. If those other factors are measured with error, a small number of controls is insufficient to make this claim, as illustrated in the simulations of Section 1.1. In this section, we show that additional controls can ameliorate this issue. This is illustrated by replicating the competitiveness and gender study of Niederle and Vesterlund (2007)—henceforth NV—within the Caltech Cohort Study. Like that study, we find a robust difference in the rates at which men and women compete. However, NV conclude that:

> Finally, controlling for gender differences in general factors such as overconfidence, risk, and feedback aversion, we estimate the size of the residual gender difference in the tournament-entry decision. Including these controls, gender differences are still significant and large. Hence, we conclude that, in addition to gender differences in overconfidence, a sizeable part of the gender difference in

> tournament entry is explained by men and women having different preferences for performing in a competitive environment.

In contrast, we show that the gender gap is well explained by risk aversion and overconfidence alone, using multiple measures of each.

As shown in Section 1.1, measurement error in $X$—in this case controls for risk aversion and overconfidence—can result in a biased estimate of the coefficient on $D$—gender—on competition, $Y$. To understand this intuitively, consider the model in (1) where $Y^* = X^*$, $D$ and $X^*$ are correlated, and $X = X^* + \nu$ is a noisy measure of $X^*$. Now, consider the limit where the variance of $\nu$ is very large, so that $X$ is almost entirely noise. Then, it would appear that $X$ has no effect on $Y$, but the variation in $Y$ and $D$ would be correlated, leading to a (possibly significant) estimate of the effect of $D$ on $Y$—even though $D$ has no direct effect on $Y$. As $Y$ may also be measured with error, we use linear probability models throughout so that this will not bias the results.[12]

To put this in terms of our substantive example, it is well known that overconfidence is correlated with gender (see, for example, Moore and Healy, 2007, 2008), and, depending on the elicitation method, risk aversion may be correlated with gender as well (see Holt and Laury, 2014, for a survey, and our discussion in Section 4.6). Thus, if competitiveness is driven by overconfidence or risk aversion, mis-measurement of these traits will lead to an overestimate of the effect of gender.

What can be done to mitigate this issue? The simplest approach is to include multiple measures for each of the possible controls $X$.[13] This is the approach we take when examining the replication of NV.

[12] It is well known that measurement error in left-hand-side variables may bias estimated coefficients in discrete choice models, see Hausman (2001).

[13] Another approach is to directly reduce measurement error in $X$ by averaging a large number of measures. This will work whenever the law of large numbers (LLN) holds. However, as the LLN holds only asymptotically, a very large number of controls may be necessary. In our data, where we have 19 total measures of these two attitudes, in total, entering these directly in the regression is a more promising strategy.

## 3.1 Measuring Competitiveness in the Caltech Cohort Study

Part of the Spring 2015 survey mimicked the essential elements of NV's original design. Participants first had three minutes to complete as many sums of five two-digit numbers as they could.[14] The participants were informed that they would be randomly grouped with three others at the end of the survey. If they completed the most sums in that group of four, they would receive 40 experimental tokens (or \$0.40) for each sum correctly solved, and would otherwise receive no payment for the task. Ties are broken randomly. As in NV, at the end of this task, participants were asked to guess their rank, from 1 to 4, within the group of four participants. They were paid 50 tokens (or \$0.50) if their guess was correct.[15]

In the second task, participants were told they would have an additional three minutes to complete sums, as in the first task. However, before doing so, they chose whether to be paid according to a piece-rate scheme or a tournament. The piece-rate scheme paid 10 tokens for each sum solved correctly. The tournament had a similar payment scheme to the first three-minute task. The difference was that the participant's performance in the tournament would be compared to the performance of three randomly chosen participants in the *first* task.[16] This ensured that the participant would not need to be concerned about the motivation or other characteristics that might drive someone to compete in the second task. Otherwise, the payment structure was identical to that in the first task.

There are a few dimensions on which our implementation differs from NV's. In each of our tasks participants are allowed three, instead of five, minutes to complete the sequence of sums. Per-sum payments are scaled down by a factor of four. Additionally, in NV, participants were assigned to groups of two women and two men. This created a small imbalance between the gender distribution of partners observed by women—one other woman and two men—and

[14] The five numbers were randomly determined, as in NV.

[15] As in NV, upon ties, we interpreted the guess in the way that was most favorable to participants. For example, if a participant correctly solved 14 sums, while others in the group solved 14, 12, and 11 sums, the participant would get the 50 token reward with a guessed rank of either 1 or 2.

[16] Different from NV, these three participants were chosen independently of the group in the first task.

men—one other man and two women. We use random groups, and therefore both women and men face identical distributions of potential partners. The modal outcome was one woman and two men, which occurred 43% of the time.[17]

In keeping with the nature of payments throughout the CCS, participants were paid based on the outcomes of both tasks, rather than on one randomly chosen task, as in NV. This could potentially lead to participants hedging by choosing the piece-rate payment scheme for the second task. If the gender gap in competitiveness is not driven by risk aversion, this hedging motivation should affect both genders equally. However, as we will show, differences in risk aversion are important for driving the gender gap in competitiveness.

NV include two additional parts: a preliminary task allowing participants to try out the piece-rate scheme, and a final choice that allows participants to select an additional piece-rate scheme or tournament scheme based on their performance in that preliminary task. Their goal was to provide a control for risk aversion and overconfidence. As the CCS has multiple other controls for both risk aversion and overconfidence (see Section 2.1), we omit these two parts to reduce the complexity and time taken to elicit competitiveness.[18]

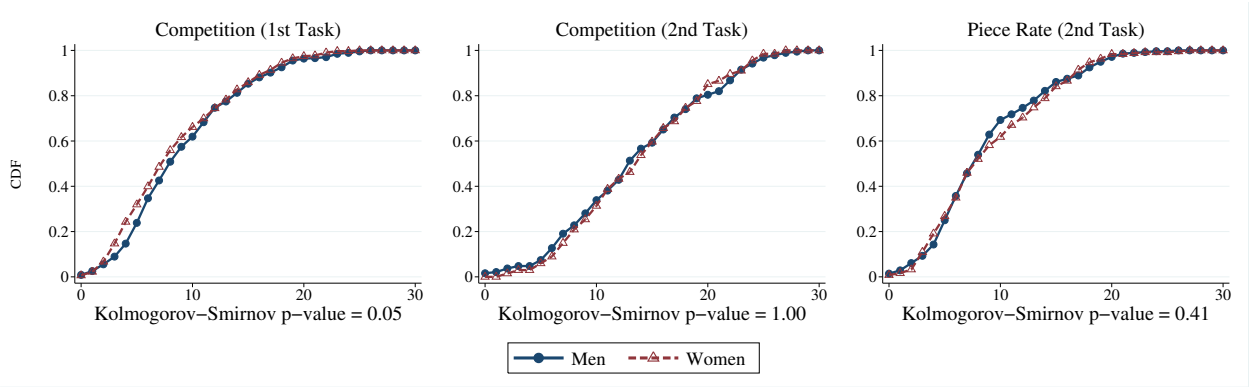
## 3.2 Performance of Participants

Figure 1 illustrates the performance of men and women in both tasks, splitting participants in the second task into those who chose to compete and those who chose the piece-rate

[17]Participants were not aware of the genders of those in their group, as groups were not formed until after a participant completed the survey. Because our experiment was conducted on a survey rather than in the lab, there were additional differences. Notably, we could not enforce our prohibition on the use of calculators. To mitigate this, sums were presented as images so it would be impossible to cut and paste into an online calculator. Participants could not use the back button during the task, and logging out and returning caused them to skip the task altogether. Participants were asked to sign an acknowledgement that they would abide by the Caltech Honor Code. Despite all this, we cannot eliminate the possibility that some students used calculators. However, in order for this to affect our results, it would have to be the case that women believed men used calculators more often, and that this belief affected not only their decision to compete, but also their choices in unrelated risk aversion and overconfidence elicitations.

[18]In Appendix B, we also show that when the final choice from NV is measured with error, as it would be in a random utility model like a Logit or Probit, this provides an imperfect control, and thus biased results. A proper specification using their data results in an insignificant coefficient on gender.

Figure 1: Performance in different tasks by gender




scheme. There is no significant difference between the performance of men and women in the second task, regardless of the compensation scheme.

However, in the first task there is a difference in performance. This difference occurs among those that complete fewer than 10 sums. In that group, women solve slightly fewer sums, on average, then men. This may appear to differ from the findings of NV, who observe statistically identical performance. However, we have roughly 10 times as many participants as they do. If we draw a group of 40 men and 40 women—the number of participants in NV—10,000 times randomly from our data, we observe a p-value less than 10% only 8.3% of the time. Thus, performance data seems consistent with NV.[19]

## 3.3 Gender, Competition, and Controls

This subsection analyzes the extent to which risk aversion and overconfidence drive the gender gap in competitiveness. Table 3 summarizes specifications meant to illustrate different points. These are linear probability models, and hence, the coefficient on gender is directly interpretable as the percentage-point gap between men and women in choosing to compete.[20]

[19]This task was chosen by NV because it was found to elicit similar performance between men and women. This may not hold for all tasks: see Gneezy et al. (2003).

[20]As noted in Footnote 12, discrete choice models may produce biased estimates of coefficients when the left-hand-side variable is measured with error. Nonetheless, in our data, Probit and Logit specifications

The first column shows the baseline difference in competition: women compete 19.0% less of the time than men. women choose competition incentives 21.4% of the time, while men choose them 40.4% of the time. While these numbers are somewhat less than those reported in NV, their relative sizes are quite similar. This difference is highly statistically significant. The second column controls for participants' estimates of their own rank, as well as their performance, linearly, as in NV's main specification. Similar to their results, the inclusion of these controls reduces the coefficient on gender by approximately 1/3.[21]

However, there is a non-linear relationship between expected rank and perceived probability of victory in a competition.[22] Therefore, the third column enters participants' subjective ranks non-parametrically, by including a dummy variable for each possible response (3 categories). This estimation confirms that the effect of perceived rank on competition is, indeed, non-linear, although the coefficient on gender remains unchanged.[23] The third column also enters performance non-linearly (29 and 26 categories, respectively), as there is a non-linear relationship between performance and competition. The coefficient on gender in the third column is lower than in the second. This is entirely driven by entering performance in the first task non-parametrically, as expected from Figure 1.

The fourth column begins introducing additional controls for risk aversion and overconfidence. In this column, two (non-randomly) selected controls are entered, one for each

produce almost identical levels of statistical significance as in Table 3.

[21]While we view this as the main specification in NV, they consider an additional specification that uses participants' choices in their final task to control for risk and feedback aversion. In Appendix B, we show that when there is noise in this measure it is an imperfect control, and yields biased results. Using their data, and a specification that accounts for measurement error, results in an insignificant coefficient on gender.

[22]If an individual believes a random participant is inferior to her with probability $p$, then her probability of winning is $p^3$. Furthermore, her expected rank is given by $\sum_{i=0}^{3}(i+1)\binom{3}{i}(1-p)^i p^{3-i} = 3(1-p)+1$. We can therefore back out the probability $p$ from any reported rank $r$ (ignoring rounding). With a reported rank of $r$, the probability of winning the competition is $\left(\frac{4-r}{3}\right)^3$.

[23]Rates of competition are 65.6% for participants who predicted they would come in first (in a random group of 4), and 31.4%, 15.3%, and 5.0% for participants predicting they would come in second, third, and fourth (last), respectively. These rankings represented 195, 293, 215, and 80 participants, respectively. This distribution differs from that reported in NV: in their data 30 out of the 40 men and 17 out of the 40 women guessed a rank of 1 and only one man and two women guessed a rank of 4. Our participants were better calibrated, and this likely resulted in the lower observed rates of competition.

Table 3: Gender, competition, and controls

| Dependent Variable | Chose to Compete ($N = 783$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Male | 0.19*** | 0.13*** | 0.11*** | 0.11*** | 0.048 | 0.039 | 0.063* | 0.050 |
| | (.034) | (.031) | (.031) | (.031) | (.031) | (.033) | (.034) | (.034) |
| Guessed Competition Rank | | -0.15*** | $F = 29$ | $F = 28$ | $F = 23$ | $F = 22$ | $F = 25$ | $F = 21$ |
| | | (.017) | $p = 0.00$ | $p = 0.00$ | $p = 0.00$ | $p = 0.00$ | $p = 0.00$ | $p = 0.00$ |
| Tournament Performance | | 0.086*** | $F = 1.6$ | $F = 1.6$ | $F = 1.6$ | $F = 1.6$ | $F = 1.6$ | $F = 1.5$ |
| | | (.020) | $p = 0.02$ | $p = 0.02$ | $p = 0.02$ | $p = 0.03$ | $p = 0.03$ | $p = 0.05$ |
| Performance Difference | | -0.021 | $F = 1.4$ | $F = 1.5$ | $F = 1.4$ | $F = 1.4$ | $F = 1.4$ | $F = 1.4$ |
| | | (.017) | $p = 0.09$ | $p = 0.07$ | $p = 0.11$ | $p = 0.10$ | $p = 0.07$ | $p = 0.11$ |
| Risk Aversion: MPL #1 | | | | 0.042*** | | | | |
| | | | | (.015) | | | | |
| Overplacement: CRT | | | | 0.026* | | | | |
| | | | | (.015) | | | | |
| Risk Aversion: Project #2 | | | | | 0.067*** | 0.065*** | | |
| | | | | | (.016) | (.019) | | |
| Perceived Performance (pctile.): CRT | | | | | -0.042*** | -0.038** | | |
| | | | | | (.016) | (.016) | | |
| Risk Aversion: Project #1 | | | | | | 0.0073 | | |
| | | | | | | (.018) | | |
| Overprecision: Guess #2 | | | | | | 0.025* | | |
| | | | | | | (.015) | | |
| All Risk Aversion Controls | | | | | | | $F = 3.7$ | $F = 3.6$ |
| | | | | | | | $p = 0.01$ | $p = 0.02$ |
| All Overconfidence Controls | | | | | | | | $F = 1.7$ |
| | | | | | | | | $p = 0.05$ |
| Adjusted $R^2$ | 0.038 | 0.23 | 0.26 | 0.27 | 0.28 | 0.29 | 0.28 | 0.29 |

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses. Coefficients and standard errors on all non-dichotomous measures are standardized. F-statistics and p-values are presented when variables are entered categorically rather than linearly. There are 3 categories for Guessed Competition Rank, 29 categories for Tournament Performance, 26 categories for Performance Difference, 6 variables for Risk Aversion Controls, and 12 variables for overconfidence controls.

attribute. As can be seen, this does not affect the coefficient on gender, despite the fact that both controls have statistically significant coefficients. The fifth column contains a different two (non-randomly) selected controls, which cuts the coefficient on gender by more than half, and renders it statistically insignificant. Taken together, these columns show that the statistical significance of controls is not a good indicator of whether or not a trait is fully controlled for. Moreover, it suggests that measurement error in the controls themselves allows for a perceived gender gap to flourish.

The seventh column adds two more controls for overconfidence and risk, which potentially mitigates issues due to measurement error. In this specification, the coefficient on gender drops even further. The final two columns enter all available controls, first for risk (6 controls), and then for risk and overconfidence (an additional 12 controls). Comparing the two columns suggests that much of the decrease in the coefficient in the final, preferred, specification is due to controls for risk rather than overconfidence. We return to a discussion of risk attitudes and gender in Section 4.6.

It is worth noting that these conclusions are not driven by our unusually large sample size. If anything, the size of our dataset helps reduce standard errors and identify weak effects that, with a smaller dataset, would appear insignificant. To see this, we draw a random sample of 40 women and 40 men (the size and gender composition of NV's experiment) from our data $10,000$ times and replicate the specification in Column 6 of Table 3 (without the performance controls). In the simulations, we use only five controls: two overconfidence controls, two risk controls, and perceived rank in the first competition task. The coefficient on gender is significant at the 1% level 2.2% of the time, at the 5% level 7.6% of the time, and at the 10% level 13% of the time.

Table 4: Principal components can be used to save degrees of freedom.

| Dependent Variable: | Chose to Compete ($N = 783$) | | | | | |
|---|---|---|---|---|---|---|
| Male | 0.19*** | 0.10*** | 0.10*** | 0.054* | 0.054* | 0.41 |
| | (.034) | (.034) | (.034) | (.033) | (.033) | (.033) |
| First Principal Component | | 0.14*** | 0.14*** | 0.15*** | 0.15*** | 0.15*** |
| | | (.017) | (.017) | (.016) | (.016) | (.016) |
| Second Principal Component | | | 0.020 | 0.020 | 0.019 | 0.019 |
| | | | (.016) | (.015) | (.015) | (.015) |
| Third Principal Component | | | | -0.14*** | -0.13*** | -0.14*** |
| | | | | (.015) | (.015) | (.015) |
| Fourth Principal Component | | | | | -0.0094 | -0.0093 |
| | | | | | (.015) | (.015) |
| Fifth Principal Component | | | | | | 0.034** |
| | | | | | | (.015) |
| Adjusted $R^2$ | 0.038 | 0.12 | 0.12 | 0.20 | 0.20 | 0.20 |

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses. Coefficients and standard errors on all non-dichotomous measures are standardized.

## 3.4 Preserving Degrees of Freedom

Our preferred specification, the final column of Table 3, contains 76 control variables (including the non-parametric controls for perceived rank and performance). With a dataset the size of ours, this does not present a challenge, as can be seen from the fact that standard errors of the coefficient on gender are stable across specifications. However, this many controls would be infeasible in a sample on the order of the original NV experiment. Moreover, a potential concern with the final column of Table 3 is that adding too many controls measured with error may bias the coefficient on gender downwards. Although this is unlikely to be the case here due to the size of our dataset and the fact that we get similar results with only five controls, it may be a concern in smaller datasets. This raises the question of how to use a relatively exhaustive control strategy with available degrees of freedom.

A growing literature in statistics and econometrics considers inference in the presence of a high-dimensional set of sparse controls.[24] One way to cope with this issue is to perform model selection after rotating the controls into their principal components, following the strategy proposed by Belloni et al. (2013). This transformation concentrates the information from the controls into relatively few factors, effectively controlling for a rich characterization of risk preferences and overconfidence without giving up too many degrees of freedom.

Table 4 illustrates this approach. We first conduct a principal components analysis of all 76 controls used in the last column of Table 3. We then enter these principal components sequentially in the columns of Table 4. As can be seen, the first principal component causes the coefficient on gender to fall by approximately half, the third by a further half, and the fifth by an additional 25%. The second and fourth have no impact on the coefficient on gender. The adjusted-$R^2$ begins to drop when the 12th component is entered, at which point the coefficient on male is 0.0431 (s.e. 0.0330)—that is, the coefficient on male does not change meaningfully as components 6 through 11 are added. Importantly, this suggests one can control for all relevant variation using only five controls. Moreover, this strategy allows the use of non-parametric or semi-parametric versions of the controls we enter linearly.

As can be seen throughout the table, the second and fourth principal components are statistically insignificant, indicating the potential for using LASSO, or similar variable selection techniques, when using principal components. The Belloni et al. (2013) approach would imply a two-stage model selection strategy, using first a LASSO regression to select the control principal components that correlate with tournament participation and then using a second LASSO regression to select the components that correlate with gender. We refer interested readers to their paper, as detailing their algorithm is beyond the scope of this work, and our

[24]This literature's roots lie in machine learning techniques for automating model selection, including the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and the Dantzig Selector (Candes and Tao, 2007). Performing inference after model selection presents a non-trivial statistical problem (Leeb and Pötscher, 2005) with recent innovations from Belloni et al. (2011), Fan and Liao (2014), Belloni et al. (2013), and Van de Geer et al. (2014) establishing new techniques for robust inference after model selection.

results are robust to selection after including the first five principal components.

## 3.5 Substantive Interpretation

Our analysis suggests that gender differences in competitiveness can be explained through differences in risk attitudes and overconfidence. This is counter to one of the main messages of NV: that the gender gap in tournament entry is influenced by men and women differing in their *preference* for performing in a competitive environment—beyond differences in risk aversion, overconfidence, and feedback aversion. It is important to note that using multiple controls for risk aversion and overconfidence, or using principal components, does not allow us to say how important these two factors are to competition, only that together they explain the gender gap in competition.

Our results do not, by any means, imply that it is better to elicit risk attitudes and overconfidence instead of competitiveness. There is a tradeoff: competition is potentially more directly relevant for an array of economically important decisions, and is definitely a more parsimonious measure. Indeed, competition has been been shown to explain several interesting behaviors, such as choice of college major (see, for example, Buser et al., 2014). However, risk aversion and overconfidence feature in many theories, and are therefore of potential use in bringing theory to bear on a particular situation.

There are also practical considerations. NV report that their experiment had an average runtime of approximately 45 minutes. By using two tasks (rather than four) and allowing participants to solve sums for three minutes (rather than five), we reduced the average time participants spent on the competition task to around 8 minutes. Naturally, eliciting multiple measures of risk and overconfidence may be time consuming as well. Nevertheless, our entire survey had an average runtime of less than 30 minutes, including the competition task.

# 4 Measurement Error Left and Right

We now shift to a situation where the variables of interest, rather than controls, are measured with error. This leads to attenuation bias in estimating the relationship between different variables. We introduce a simple method, *Obviously Related Instrumental Variables* (ORIV), to correct for this. We apply this technique to estimating the correlation between different risk measures in this section, and between risk and ambiguity aversion in the next.

It is well known that measurement error in outcome, or dependent, variables does not bias estimated relationships, although it increases standard errors. Measurement error in explanatory, or dependent, variables is a much more serious problem, biasing estimated coefficients towards zero and distorting standard errors. This leads to an improper understanding of the relationship between explanatory variables and outcomes. These problems are compounded when estimating a correlation: the distinction between outcome and explanatory variables is blurred, and measurement error in either biases estimates towards zero.

The following subsections develop the ORIV approach gradually. The first subsection introduces the substantive question of investigating correlations between different risk measures. The next gives a simple treatment of the standard application of instrumental variables to correct for measurement error in explanatory variables. The following three subsections show, theoretically and empirically, how to combine information from multiple instruments, and how to consistently estimate correlation coefficients when both explanatory and outcome variables are measured with error. The discussion in this section focuses on implementation, with the formal properties of the estimators developed in Appendix A.

## 4.1 Risk Elicitation Techniques

There is a substantial experimental literature assessing the validity of common experimental techniques for eliciting attitudes towards risk and uncertainty (see the literature review in

Holt and Laury, 2014). These studies often elicit risk attitudes in the same set of participants using different techniques. By using a within-participant design, researchers attempt to understand technique-driven differences in elicited proxies for risk aversion. These works find small correlations between different techniques, making it difficult to study the individual correlates of risk preferences. To mention a few examples, Dave et al. (2010) compares the Lottery Menu task with the Holt and Laury (2002) task—in which participants choose one lottery in each of a sequence of lottery pairs in which the means and variances change from pair to pair. Participants appear to be more risk averse in the Holt and Laury task. Deck et al. (2008) compares behavior in the Holt and Laury task to that in a task that was a variation on the game show "Deal or No Deal", and find the correlation between risk attitudes from the two tasks is only 0.008, with a p-value of 0.94. Deck et al. (2010) compares the same two tasks, adding two others (including the Lottery Menu task used here), as well as survey questions touching upon risk in six different domains. The highest pairwise correlations they find is less than 0.3. Similarly, Anderson and Mellor (2009) compares responses to the Holt and Laury task to survey questions about hypothetical gambles. They find small correlations, and provide a survey of the literature with similar results.

Ultimately, the literature concludes that risk elicitation is a "risky business" (the pun is not ours, see Friedman et al., 2014, for a survey). Indeed, these authors conclude that:

> Estimated parameters exhibit remarkably little stability outside the context in which they are fitted. Their power to predict out-of-sample is in the poor-to-nonexistent range, and we have seen no convincing victories over naive alternatives.

However, none of the studies on which this conclusion is based account for measurement error when estimating correlations between elicitation techniques. In what follows, we inspect several commonly used risk-attitude elicitation techniques, and estimate their within-participant correlations using an IV strategy to control for measurement error. This generates much higher within-subject correlations than previously reported. Moreover, the corrected corre-

lations suggest that elicitation techniques fall into one of two sets: those that elicit certainty equivalents for lotteries, and those that elicit allocations of assets between safe and risky options. The latter category exhibits greater corrected correlations with other measures, and more stability over time. Further, elicitations based on allocation decisions display substantial gender effects—which is consistent with investment behavior in the field—while certainty equivalent elicitations do not.

This section uses four measures of risk as described in Section 2.1: Qualitative, Risk MPL, Project, and Lottery Menu. Before we proceed, we note a few details about how we handle the data from those elicitations to standardize estimated quantities for easy comparison. First, when using two measures from the same form of elicitation, we put these on a common scale. In particular, the certainty equivalents from the 30-ball urn Risk MPL (which go up to 150) are divided by 1.5 to be on the same scale as the certainty equivalents from the 20-ball urn Risk MPL (which go up to 100).[25] Second, when comparing objects like estimated CRRA coefficients or derived certainty equivalents, these are also put on the same scale. For example, when examining the relationship between certainty equivalents from the Risk MPLs and Projects—the former allowing for risk-loving answers and the latter not—those who gave risk-loving answers on the urns are re-coded to give a risk-neutral answer. Without this censoring, results are substantially similar.[26]

## 4.2 A First Take on Measurement Error Correction

It is well known that measurement error attenuates estimated coefficients (see, for example, Greene, 2011). Here we review that basic finding to set up a framework for our estimator.

To estimate the relationship between two variables measured with independent i.i.d. error, $Y = Y^* + \nu_Y$ and $X = X^* + \nu_X$ (with $\mathbb{E}[\nu_Y \nu_X] = 0$ and $\text{Var}[\nu_k] = \sigma^2_{\nu_k}$), the ideal regression

[25] This implicitly assumes a CRRA utility function.

[26] This censoring affects 22% of the responses in the 20-ball urn, and 32% of responses in the 30-ball urn.

model would be $Y^* = \alpha^* + \beta^* X^* + \varepsilon^*$. Instead, we can only estimate $Y = \alpha + \beta X + \varepsilon$, where $\alpha$ is a constant and $\varepsilon$ is mean-zero random noise. Annotating finite-sample estimates with hats and population moments without hats, this results in an estimated relationship of

$$\begin{aligned} \hat{\beta} &= \frac{\widehat{\text{Cov}}[Y, X]}{\widehat{\text{Var}}[X]} = \frac{\widehat{\text{Cov}}[\alpha + \beta^* X^* + \varepsilon + \nu_Y, X^* + \nu_X]}{\widehat{\text{Var}}[X^* + \nu_X]} \\ \mathbb{E}\Big[\hat{\beta}\Big] &= \operatorname*{plim}_{n\to\infty} \hat{\beta} = \beta^* \frac{\sigma^2_{X^*}}{\sigma^2_{X^*} + \sigma^2_{\nu_X}} < \beta^*. \end{aligned} \tag{3}$$

The estimated coefficient $\hat{\beta}$ is thus biased towards zero. Importantly, the bias in (3) depends on the amount of information about the true explanatory variable $X^*$ in $X$.

In a lab experiment, it is relatively easy to elicit two replicated measures of the same underlying parameter $X^*$. That is, suppose we have $X^a = X^* + \nu^a_X$ and $X^b = X^* + \nu^b_X$, with $\nu^a_X, \nu^b_X$ i.i.d. random variables, and $\mathbb{E}\big[\nu^a_X \nu^b_X\big] = 0$—that is, measurement errors are independent of each other, and thus uncorrelated. With the additional assumption that $\text{Var}[\nu^a_X] = \text{Var}[\nu^b_X] \equiv \text{Var}[\nu_X]$, we have that

$$\widehat{\text{Corr}}[X^a, X^b] \to_p \text{Corr}[X^a, X^b] = \frac{\sigma^2_{X^*}}{\sigma^2_{X^*} + \sigma^2_{\nu_X}} \tag{4}$$

which allows us to ballpark the degree of bias in estimated coefficients. The modal correlation between two elicitations of the same risk measure is approximately 0.6, suggesting that the variance of measurement error is of the order of 2/3 of the variance of $X^*$.[27]

Using instrumental variables (IV), the second noisy measure of $X^*$ can be used to recover a consistent estimate of the true coefficient $\beta^*$. Recalling from the derivation of (3) that

[27] This correlation also provides a way to derive a correction factor for the attenuation bias in the regression estimates from (3) dating back to Spearman (1904). Defining the "disattenuated" estimator of $\beta$ as $\tilde{\beta} = \frac{\hat{\beta}}{\widehat{\text{Corr}}[X^a, X^b]}$ and invoking the continuous mapping theorem, it is clear that $\tilde{\beta}$ provides a consistent estimator for $\beta$. This approach, while consistent, may be inefficient in the presence of multiple replicates. As illustrated in Appendix A, our ORIV approach provides a simple formulation for consolidating the information from multiple replicates of both $X$ and $Y$.

$\widehat{\text{Cov}}[X^a, X^b] = \widehat{\text{Var}}[X^*]$, we use two-stage-least-squares (2SLS) to instrument $X^a$ with $X^b$

$$X^a = \pi_0 + \pi_1 X^b + \varepsilon_X \quad \Rightarrow \quad \hat{\pi}_1 = \frac{\widehat{\text{Cov}}[X^a, X^b]}{\widehat{\text{Var}}[X^b]} = \frac{\widehat{\text{Var}}[X^*]}{\widehat{\text{Var}}[X^b]}, \tag{5}$$

and then condition on this instrumented relationship to estimate $Y = \alpha + \beta(\hat{\pi}_0 + \hat{\pi}_1 X^b) + \varepsilon_Y$. This second stage regression provides

$$\hat{\beta} = \frac{\widehat{\text{Cov}}[\alpha^* + \beta^* X^* + \varepsilon^* + \nu_Y, \hat{\pi}_0 + \hat{\pi}_1 X^b]}{\widehat{\text{Var}}[\hat{\pi}_0 + \hat{\pi}_1 X^b]} = \frac{\beta^* \hat{\pi}_1 \widehat{\text{Var}}[X^*]}{\hat{\pi}_1^2 \widehat{\text{Var}}[X^b]} \to_p \beta^*,$$

a consistent estimate of $\beta^*$, the true relationship between $Y^*$ and $X^*$.

## 4.3 Two Instrumentation Strategies

Multiple measures for $X^*$ admit multiple instrumentation strategies that, in finite samples, can produce very different results. The ORIV estimator consolidates the information from these two formulations of the problem. In our working example, we have two equally valid elicitations and two possible instrumentation strategies: one may instrument $X^a$ with $X^b$, or $X^b$ with $X^a$. In this subsection, we illustrate the divergent results these two strategies may produce. The next subsections show how to deal with this issue by combining these sources of information into a single estimated relationship.

Table 5 illustrates the estimated relationships between different elicitation techniques. These relationships are first estimated using a standard regression, and then the two different IV strategies discussed above. The coefficients are from regressions where both the left and right-hand-side variables are standardized, and thus are effectively correlations, which removes scale effects and provides for easy comparison.

The table contains three insights. First, in line with previous work, raw correlations are low to moderate, ranging from 0.15 to 0.39. Second, corrected measures are substantially

Table 5: Correlation between different risk measures is understated due to measurement error.

| Dependent Variable | Qualitative Assessment | | | | Lottery Menu | | | |
|---|---|---|---|---|---|---|---|---|
| Project #1 | 0.31*** | | | | 0.24*** | | | |
| | (.034) | | | | (.034) | | | |
| Project #2 | | 0.29*** | | | | 0.29*** | | |
| | | (.034) | | | | (.034) | | |
| Project #1 (Instrumented) | | | 0.55*** | | | | 0.59*** | |
| | | | (.069) | | | | (.073) | |
| Project #2 (Instrumented) | | | | 0.58*** | | | | 0.50*** |
| | | | | (.067) | | | | (.070) |
| Risk MPL #1 | 0.15*** | | | | 0.19*** | | | |
| | (.036) | | | | (.035) | | | |
| Risk MPL #2 | | 0.17*** | | | | 0.23*** | | |
| | | (.035) | | | | (.035) | | |
| Risk MPL #1 (Instrumented) | | | 0.22*** | | | | 0.44*** | |
| | | | (.048) | | | | (.067) | |
| Risk MPL #2 (Instrumented) | | | | 0.21*** | | | | 0.37*** |
| | | | | (.047) | | | | (.067) |

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses. Coefficients are from regressions where both the right and left-hand-side variables are standardized, and thus are correlations. $N = 776$ for all regressions.

higher, achieving a level of up to 0.71 between the Project measure and the Lottery Menu measure. Last, our Project measures exhibit substantially higher correlations, both raw and corrected, with both the Quantitative and the Lottery Menu measures.

Although different instrumentation strategies may produce similar results—as in the third and fourth columns of Table 5—they can also produce different results—as in the seventh and eighth columns. Moreover, given that estimated standard deviations—which are inflated by measurement error—are used to standardize the variables in Table 5, neither strategy is likely to produce an accurate result. The next subsection deals with both of these issues.

## 4.4 Obviously Related Instrumental Variables

We construct ORIV estimates and corrected correlations in three steps. First, we consider the case where only explanatory variables are measured with error. We then extend the analysis to the case where both the outcome and explanatory variables are measured with error. Finally, we show how to derive consistent correlations from the consistent and asymptotically efficient ORIV estimates of the regression coefficient $\beta$. Throughout, we focus on designs in which there are at most two replications for each measure. This is done for simplicity, and because it fits precisely the implementation carried out using the Caltech Cohort Study.[28]

### 4.4.1 Errors in Explanatory Variables

We continue with the model stated in Section 4.2, noting that unlike the analysis in Section 4.3, these measures are not standardized. The ORIV regression estimates a stacked model to consolidate the information from the two available instrumentation strategies:

$$\begin{pmatrix} Y \\ Y \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \beta \begin{pmatrix} X^a \\ X^b \end{pmatrix} + \varepsilon \quad \text{instrumenting the LHS with} \quad W = \begin{pmatrix} X^b & 0_N \\ 0_N & X^a \end{pmatrix}, \tag{6}$$

[28] Appendix A extends the ORIV estimator to settings where more than one replicate is available.

where $N$ is the number of participants, and $0_N$ is an $N \times N$ matrix of zeroes. To implement this, one need only to create a stacked dataset and run a 2SLS regression. This can be thought of as estimating a first stage, as in (5), for both instrumentation strategies, and then estimating

$$\begin{pmatrix} Y \\ Y \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \beta \begin{pmatrix} \hat{X}^a \\ \hat{X}^b \end{pmatrix} + \varepsilon, \tag{7}$$

where $\hat{X}^a$ and $\hat{X}^b$ are the predicted values derived from the two first-stage regressions. Sample Stata code illustrating this estimation procedure appears in Appendix C.2.[29]

With a single replication, the stacked regression will produce an estimate of $\beta^*$ that is the average of the two alternative instrumentation approaches in the previous subsections. Intuitively, this occurs as, with no theoretical reason to favor one or the other, it is equally likely that the smaller is too small as it is that the larger is too large. The estimator splits the difference, leading to a consistent estimate of $\beta^*$, the true relationship between $Y^*$ and $X^*$. We denote this estimate using $\hat{\beta}^*$.

**Proposition 1.** *ORIV produces consistent estimates of $\beta^*$.*

This technique uses each individual twice, which results in standard errors that are too small, as the regression appears to have twice as much data as it really does. Many practitioners understand intuitively the idea that one should use clustered standard errors to treat multiple observations as having the same source.[30]

[29] Note that if one estimates 2SLS in stages, the estimated standard errors from the second stage, in (7), would be incorrect, as they do not take into account the fact that the $\hat{X}^k$s are estimated. Therefore it is preferable to estimate (6) directly, using a statistical package's 2SLS command, as this will give correct asymptotic standard errors.

[30] Mathematically, clustering is needed as $\text{Cov}[\varepsilon_i, \varepsilon_{N+i}] = \text{Var}[\varepsilon_i^*]$ for $i \in \{1, 2, 3, \ldots, N\}$. This implies that the variance-covariance matrix of residuals is given by

$$\begin{pmatrix} (\text{Var}[\varepsilon^*] + \beta^2 \text{Var}[\nu^a]) I_N & \text{Var}[\varepsilon^*] I_N \\ \text{Var}[\varepsilon^*] I_N & (\text{Var}[\varepsilon^*] + \beta^2 \text{Var}[\nu^b]) I_N \end{pmatrix}, \text{ where } I_N \text{ is an } N \times N \text{ identitymatrix.}$$

Clustering takes care of the common $\varepsilon_i^*$ for participant $i$ on- and off-diagonal.

As the diagonal terms differ as to whether $\text{Var}[\nu^a]$ or $\text{Var}[\nu^b]$ remains, this suggests a different weighting of $X^a$ and $X^b$ is optimal. This could be implemented using Feasible-Generalized Least Squares (FGLS)

**Proposition 2.** *The ORIV estimator satisfies asymptotic normality under standard conditions. The estimated standard errors, when clustered by participant, are consistent to the asymptotic standard errors.*

Simulations, and experience with data, suggest that block-bootstrapped standard errors are, if anything, slightly smaller, implying that asymptotic standard errors are slightly conservative. As such, practitioners should not hesitate to use asymptotic standard errors.

### 4.4.2 Errors in Outcome and Explanatory Variables

When estimating the relationship between two elicited variables there is no reason to believe that one is measured with error ($X$), but the other is not ($Y$). The existence of measurement error in $Y$ does not change Propositions 1 and 2, although estimated standard errors will, of course, increase, reflecting the degree of uncertainty in the estimated coefficient. Still, if one has access to two estimates of $Y$ there is no reason not to use them, as they will increase efficiency. Moreover, when estimating correlations, where neither variable can be said to be the explanatory or outcome variable, measurement error in either variable will bias the estimated correlation towards zero.

To incorporate two measures of $Y^*$ with measurement error ($Y^a = Y^* + \nu_Y^a$, $Y^b = Y^* + \nu_Y^b$, $\mathbb{E}[\nu_Y^a] = \mathbb{E}\left[\nu_Y^b\right] = 0$) in the ORIV estimation procedure, one would simply estimate

$$\begin{pmatrix} Y^a \\ Y^a \\ Y^b \\ Y^b \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} + \beta \begin{pmatrix} X^a \\ X^b \\ X^a \\ X^b \end{pmatrix} + \varepsilon \quad \text{with instruments} \quad W = \begin{pmatrix} X^b & 0_N & 0_N & 0_N \\ 0_N & X^a & 0_N & 0_N \\ 0_N & 0_N & X^b & 0_N \\ 0_N & 0_N & 0_N & X^a \end{pmatrix}.$$

for our ORIV estimators. However, FGLS tends to have poor small sample properties, and would likely produce worse estimates in small to moderate-sized experimental datasets. In our dataset, which is an order of magnitude larger than most, FGLS has no effect. For more detail, see Appendix A.

### 4.4.3 Estimating Correlations from Consistent Coefficients

ORIV produces $\hat{\beta}^*$, a consistent estimate of $\beta^*$. Notice that

$$\hat{\beta} = \frac{\widehat{\text{Cov}}[X,Y]}{\widehat{\text{Var}}[X]} \text{ implying } \hat{\rho}_{XY} = \hat{\beta}\sqrt{\frac{\widehat{\text{Var}}[X]}{\widehat{\text{Var}}[Y]}},$$

where $\rho_{XY}$ is the correlation. Thus, we need consistent estimates of $\text{Var}[X^*]$ and $\text{Var}[Y^*]$ to recover $\hat{\rho}^*_{XY}$. The problem is that $\text{Var}[X] = \text{Var}[X^*] + \text{Var}[\nu_X]$. As $\text{Var}[Y]$ is biased as well, it is not clear if transforming the regression coefficient into a correlation will bias the resulting correlation up or down (although overall, the correlation will be biased towards zero by measurement error). Nonetheless, we have

$$\text{Cov}[X^a, X^b] = \text{Cov}[X^* + \nu^a_X, Y + \nu^b_Y] = \text{Var}[X^*] \text{ so } \hat{\rho}^*_{XY} = \hat{\beta}^*\sqrt{\frac{\widehat{\text{Cov}}[X^a, X^b]}{\widehat{\text{Cov}}[Y^a, Y^b]}}.$$

To determine the proper asymptotic standard errors, simply multiply those estimated from the ORIV procedure by $\sqrt{\widehat{\text{Cov}}[X^a, X^b]/\widehat{\text{Cov}}[Y^a, Y^b]}$. An example of how to estimate correlations using ORIV in STATA can be found in Appendix C.3.

**Proposition 3.** *$\hat{\rho}^*_{XY}$ is consistent. Standard errors estimated from ORIV, multiplied by $\sqrt{\widehat{Cov}[X^a, X^b]/\widehat{Cov}[Y^a, Y^b]}$, are consistent.*

## 4.5 Corrected Correlations between Risk Elicitation Techniques

We now use our ORIV estimators to examine the correlations between different risk measures. Table 6 contains both the raw and corrected correlations. Both the Risk MPL task and the Project task were elicited twice. The qualitative risk measure was elicited only once in the

Fall of 2014, but again in the Spring of 2015, which serves as its second measure.[31]

Previous work comparing different risk-elicitation techniques often transforms them into a common scale (see Deck et al., 2010, for an example). For comparability, we do the same in Table 6. This is not theoretically advisable as it introduces a non-linear change in the structure of measurement error, which may lead to inconsistent estimates. However, for the question at hand, this makes little difference in the results.

In the top panel, we consider correlations between the unaltered measures. That is, we use units given by the elicitation techniques. The second panel translates these various measures into CRRA coefficients, except for the qualitative assessment, which does not lend itself to transformation. The third panel uses the imputed CRRA coefficients to calculate the implied certainty equivalent with a 50% probability of 100 tokens and a 50% probability of 0 tokens. Note that in the case of the Risk MPLs, this is the same as the questions' natural units, as these are elicitations of certainty equivalents over 50/50 lotteries.

All three panels suggest similar conclusions. First, the corrected correlations are substantially higher. While the raw correlations are arguably low, never exceeding 0.5 (and uniformly below 0.27 when considering imputed CRRA coefficients), corrected correlations are dramatically higher, reaching levels as high as 0.73. Whether this correlation is "high" or "low" is largely a judgement call. However, the literature seems to consistently suggest that correlations above 0.7 are very high (see, for example, Cohen, 1988; Evans, 1996). Moreover, many perceived strong links correspond to correlations that are 0.7 or below. For example, the correlation between parents' and their childrens' heights hovers around 0.5 (Wright and Cheetham, 1999), the correlation between height and foot length for individuals over the age of 30 is about 0.6 for females and 0.7 for males (Pawar and Dadhich, 2012), the correlation

[31]For correlations involving the Lottery Menu task, we multiply by $\sqrt{\widehat{\text{Var}}[((X^a)' \ (X^b)')']/\widehat{\text{Var}}[Y]}$. This is valid so long as the measurement error in the measures of $X$ and $Y$ are equal—that is, so long as $\text{Corr}[X^a, X^b] = \text{Corr}[Y^a, Y^b]$, as shown in (4). However, having only one measure of $Y$, we cannot test this. Nevertheless, the correlation between measures for the three that we do have are 0.67 (s.e. 0.026, projects), 0.62 (s.e. 0.028, qualitative), and 0.58 (s.e. 0.29, Risk MPLs), so this seems reasonable.

Table 6: Correlation matrices before and after controlling for measurement error

In Units Given by the Questions

| | Raw Correlations | | | Corrected for Measurement Error | | |
|---|---|---|---|---|---|---|
| | Project | Qualitative | Lottery Menu | Project | Qualitative | Lottery Menu |
| Qualitative | 0.26*** | | | 0.41*** | | |
| | (.030) | | | (.046) | | |
| Lottery Menu | 0.47*** | 0.25*** | | 0.71*** | 0.40*** | |
| | (.030) | (.033) | | (.047) | (.054) | |
| Risk MPL | 0.19*** | 0.13*** | 0.22*** | 0.30*** | 0.19*** | 0.38*** |
| | (.031) | (.033) | (.029) | (.053) | (.048) | (.060) |

Measured in CRRA coefficients

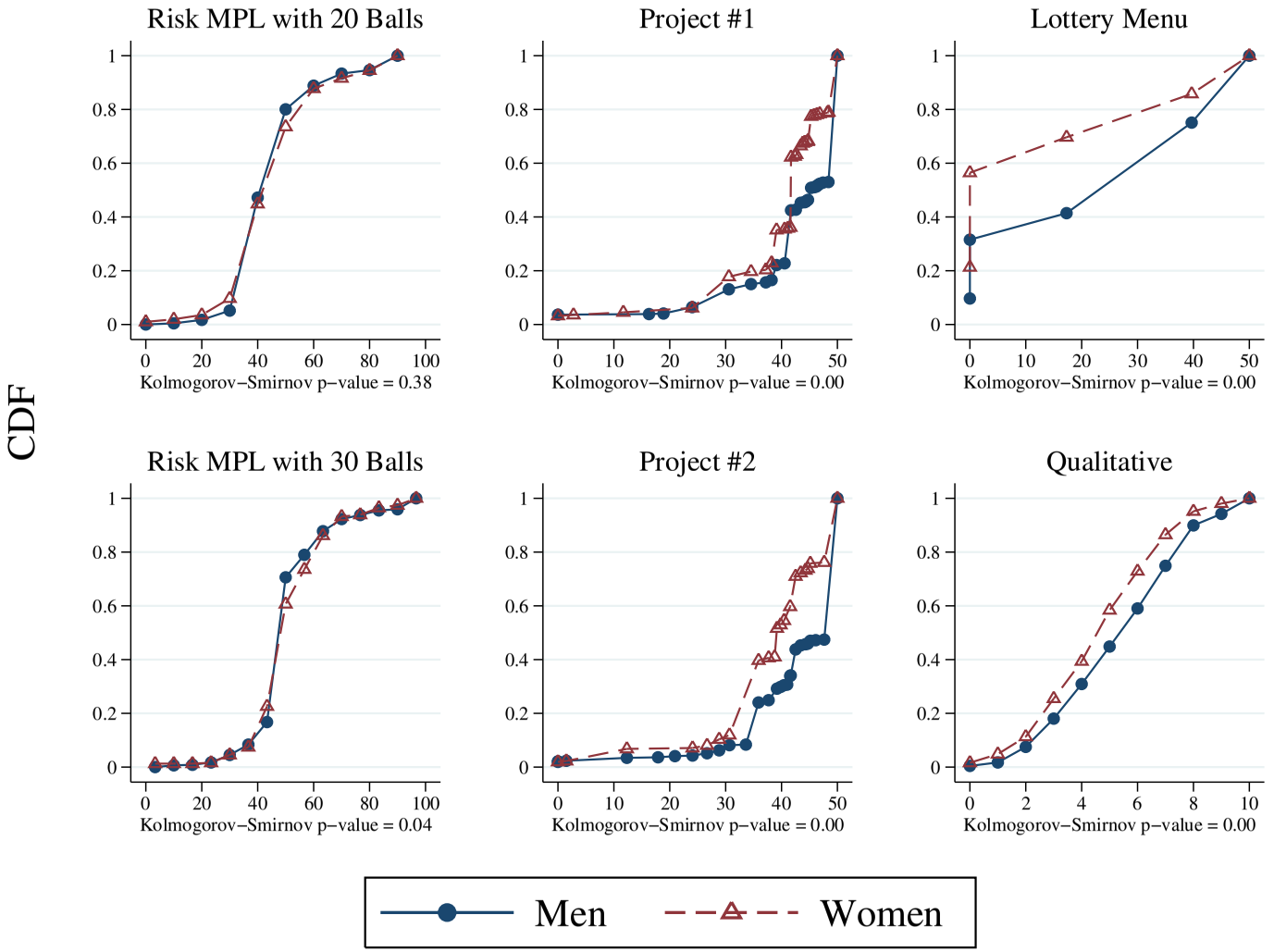
| | Raw Correlations | | | Corrected for Measurement Error | | |
|---|---|---|---|---|---|---|
| | Project | Qualitative | Lottery Menu | Project | Qualitative | Lottery Menu |
| Qualitative | 0.21*** | | | 0.36*** | | |
| | (.043) | | | (.052) | | |
| Lottery Menu | 0.27*** | 0.24*** | | 0.55*** | 0.38*** | |
| | (.057) | (.035) | | (.071) | (.057) | |
| Risk MPL | 0.18*** | 0.069** | 0.22*** | 0.37*** | 0.10** | 0.42*** |
| | (.041) | (.033) | (.038) | (.084) | (.048) | (.076) |

Measured in certainty equivalent of a 50/50 lottery over 0/100 tokens

| | Raw Correlations | | | Corrected for Measurement Error | | |
|---|---|---|---|---|---|---|
| | Project | Qualitative | Lottery Menu | Project | Qualitative | Lottery Menu |
| Qualitative | 0.25*** | | | 0.44*** | | |
| | (.029) | | | (.053) | | |
| Lottery Menu | 0.38*** | 0.23*** | | 0.73*** | 0.36*** | |
| | (.032) | (.031) | | (.086) | (.051) | |
| Risk MPL | 0.24*** | 0.13*** | 0.20*** | 0.43*** | 0.19*** | 0.34*** |
| | (.044) | (.033) | (.025) | (.080) | (.048) | (.052) |

Notes: ***, **, * denote statistical significance at the 1%, 5% and 10% level, with standard errors in parentheses. $N = 776$.

Figure 2: Risk aversion differs by gender, except when elicited using Risk MPLs.




Notes: All panels except for the qualitative assessment are put on the same scale: a certainty equivalent of a 50/50 lottery paying 0/100 tokens, via a CRRA utility function.

between average parents' education and their children's education ranges from around 0.30 in Denmark and 0.54 in Italy, with most western countries falling somewhere in-between (Hertz et al., 2007), and the correlation between Body Mass Index (BMI) and insulin resistance—which underlies the link between obesity and type-2 diabetes—is around 0.46 (Abbasi et al., 2002). Second, some measures are noticeably more correlated. Namely, the Project measure appears to be most correlated with the other elicitation techniques. It is most highly correlated with the Lottery Menu measure, with corrected correlations of 0.55–0.73, depending on the units of measurement. The Lottery Menu also exhibits relatively high correlations with the other measures. The lowest correlations are observed between the Risk MPL and the Qualitative measure.

## 4.6 Substantive Implications

There are good reasons to suspect that the Risk MPL measure captures risk attitudes over a different domain than the other measures: its smaller correlation with other measures, and the fact that, unlike other risk measures, it is uncorrelated with gender, as shown in Figure 2. In that figure, the average risk attitudes of men and women are statistically indistinguishable when considering the Risk MPLs, but all other measures show that women are substantially more risk averse than men. These differences account for the explanatory power of risk controls on the gender gap in competitiveness in Section 3.3.

The fact that different measures yield different conclusions about the relationship between gender and risk attitudes is reflected in the behavioral literature, which reaches mixed conclusions about the general relationship between gender and risk (see, for example, Byrnes et al., 1999, for a review of the relevant experimental work in psychology, and Croson and Gneezy, 2009; Eckel and Grossman, 2008b; and Niederle, 2015, for reviews of related experimental work in economics).[32] On the other hand, the finance literature has found a

[32]Our findings are consistent with some observations in Holt and Laury (2014).

more consistent difference between men and women when considering risky financial investments (see, for example, Barber and Odean, 2013; Embrey and Fox, 1997; Farrell, 2011). The Project measure intentionally mimics a stock / bond portfolio choice (or risky / safe assets), and the gender-based behavior in this task is similar to that seen in real financial investments: men invest more aggressively then women.[33]

There is also variation in the consistency of responses to the different risk elicitation techniques across time. The Project, Risk MPL, and Qualitative assessment were all elicited in both the Fall 2014 and Spring 2015 installments of the survey. The risk attitudes elicited by the Project task exhibit more stability than the Risk MPLs—a correlation of 0.65 (0.044) for the Project measure(s), compared with 0.42 (0.063) for the Risk MPL (both corrected for measurement error). The Qualitative elicitation was performed only once per survey, and the uncorrected correlation between these elicitations was 0.62.

Taken together, these different measures may be representative of risk attitudes in different settings. This would not be surprising, as psychologists have found that risk attitudes do differ across contexts (see, for example, Kruger et al., 2007; Weber et al., 2002, and references therein). While there are many criteria on which one might evaluate such measures, the Project-based measure seems particularly attractive due to its stability, correlation with other popular measures, and the fact that the literal interpretation of the measure is consistent with evidence from the field, in finance.[34]

The next section uses the ORIV technique to examine risk in a particular domain—compound lotteries—and how this relates to ambiguous (uncertain) lotteries.

[33]The lotteries in the Lottery Menu task can be viewed as corresponding to different investment allocations between a safe option and a risky one that pays three times the amount invested with 50% probability, see Eckel and Grossman (2002). Eckel and Grossman (2008a) observed that results are not sensitive to whether or not lotteries are described as risky investments to subjects, which is consistent with the Lottery Menu task exhibiting similar patterns to the Project measure.

[34]Recent evidence suggests the presence of a global risk component that explains individuals' choices across investment domains. This component exhibits some of the features the Project measure (Einav et al., 2012).

# 5 New Traits

Measurement error may lead researchers to believe that an observed behavior is not well explained by current theory. We have already shown one example of this in Section 3: using controls that are measured with error may lead researchers to believe a behavior is independent of other behaviors. It is natural to think that bias in correlations, explored in the last section, may similarly cause researchers to underestimate the relationship between two variables, and thus declare them distinct when they are, in fact, not. In this section we provide a potential example of this phenomenon by examining attitudes towards ambiguity.

Ambiguity aversion refers to a preference for known risks over unknown risks. First introduced by Ellsberg (1961), this preference implies that an ambiguity averse individual would prefer a lottery with known probability distribution of rewards over a similar lottery in which the probability distribution of rewards is unknown. This behavior is expressed in the *Ellsberg Paradox*, where participants prefer a bet on the draw of a black ball from an urn with, say, 10 red and 10 black balls than on one with 20 total balls, but with unknown composition. Ambiguity aversion is widely studied, and used to explain incomplete contracts, volatility in stock markets, selective abstention in elections, and so on (Mukerji, 2000).

Segal (1987, 1990) suggests that choices under ambiguity may come from improperly compounding a sequence of lotteries. For instance, in the Ellsberg Paradox scenario above, a participant might view a draw from the ambiguous urn as having two stages: First, the number of red balls is randomly determined, according to some subjective probability; second, a ball is drawn from the urn. If an individual fails to properly reduce these two lotteries into one, a bias will result. Halevy (2007) experimentally tests this proposition. In his study, participants face both an Ellsberg urn, and an urn where the number of red balls is uniformly determined. In his results, Halevy reports correlations of around 0.5 between behaviors in both treatments. Nonetheless, his results suggest that half the variation in the responses to

ambiguous and compound lotteries is independent. This suggests a strong, but imperfect, link between ambiguity aversion and (negative) reactions to compound lotteries.

In this section, we replicate Halevy's exercise, adding duplicate measures of certainty equivalents of both ambiguous and compound lotteries. This allows us to correct for measurement error using ORIV. As a result, ambiguity aversion and reaction to compound lotteries appear almost identical.

## 5.1 Ambiguity Aversion and Reaction to Compound Lotteries

As described in Section 3.1, the Risk MPL, Compound MPL, and Ambiguous MPL are all implemented in a very similar way. All ask for a participant's certainty equivalent value of a draw from an urn if a certain color ball is drawn. All allow the participant to select the color of the ball she would like to pay off. All have the same number of balls, and the same payoff. The only difference is that one task specifies exactly the distribution of balls in the urn: half black and half red for Risk; or unknown, but selected by the Dean of Undergraduate Students for Ambiguous; or drawn from a uniform distribution for Compound. Each measure is replicated twice: once with 20 balls and a payoff of 100 tokens if the correct color ball is drawn, and once with 30 balls and a 150-token payoff.

Our data shows evidence of ambiguity aversion, as well as a negative reaction to compound lotteries. In particular, the certainty equivalents of the ambiguous urns are 2.5 (0.48) and 2.5 (0.46) percentage points less than those of the risky urns for the 20 and 30 ball urns, respectively; and the certainty equivalents of the compound lotteries are 2.9 (0.51) and 2.8 (0.51) percentage points less than those of the risky urns. Note that these differences are statistically significant, but they are not statistically significantly different from *each other*. On average, ambiguity aversion and reaction to compound lotteries are identical.[35]

[35]On the individual level, 30% of our respondents prefer the uncertain lottery over the ambiguous one, 20% prefer the ambiguous lottery, and 50% are neutral. Halevy (2007) used the BDM method to elicit certainty equivalents, whereas we use an MPL. This choice was made because MPLs are easier to explain on a survey,

Table 7: The correlation between certainty equivalents is substantial.

| | Raw Correlations | | | Corrected for Measurement Error | | |
|---|---|---|---|---|---|---|
| | Risk CE | Compound CE | Compound Reaction | Risk CE | Compound CE | Compound Reaction |
| Compound CE | 0.55*** (.041) | | | 0.74*** (.057) | | |
| Ambiguous CE | 0.60*** (.038) | 0.65*** (.029) | | 0.78*** (.048) | 0.85*** (.035) | |
| Ambiguity Aversion | | | 0.44*** (.042) | | | 0.85*** (.086) |

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses. $N = 786$.

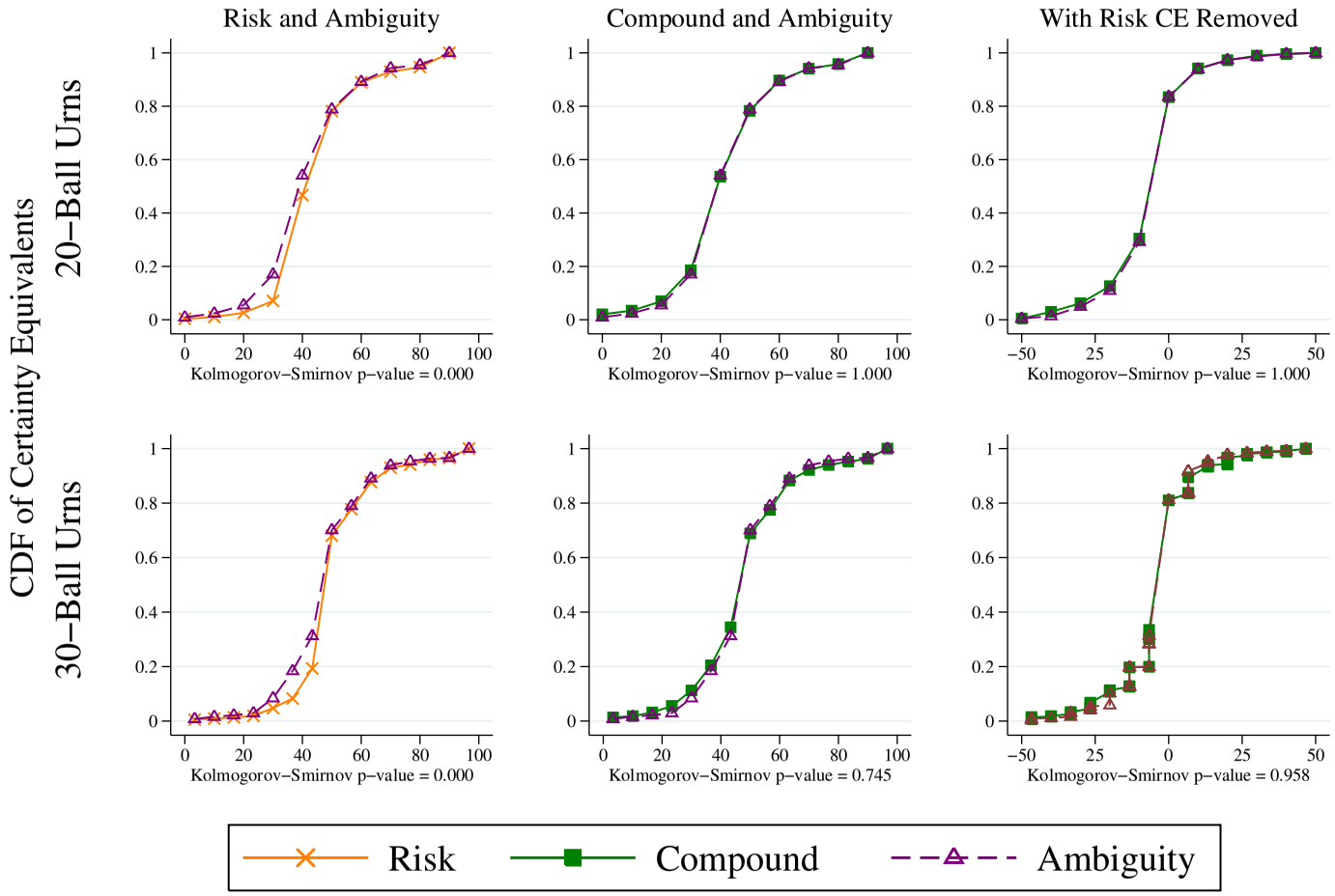
Table 7 reports the raw and corrected correlations between the three measures. The raw correlation between ambiguous and compound certainty equivalents are 0.65. This is in line with Halevy (2007), who reports a correlation of 0.45 in the first round of his experiment, and a correlation of 0.71 in his robustness round. However, once measurement error is corrected for, the correlation is much higher: 0.85.

Corrected correlations between certainty equivalents of risky and compound or ambiguous urns are substantial as well: estimated at 0.74 and 0.78, respectively. This brings up an important point: perhaps the correlation between ambiguity aversion and reaction to compound lotteries is as high as it is because the certainty equivalents of both reflect risk attitudes as well. Thus, we subtract the risk certainty equivalents from each of the compound and ambiguous certainty equivalents, leaving a measure of ambiguity aversion, and (negative) reaction to compound lotteries.[36] This results in a smaller raw correlation, but

when participants are not able to ask the experimenter for help. Once we take into account the fact that participants in our experiment express certainty equivalents in units 20 times as large as Halevy (2007), his data, which he graciously provided to us, is reasonably close to ours.

[36]Halevy's data has a correlation 0.47 in the main experiment, and 0.75 in the robustness round. Ortoleva and Dean (2015) report a correlation of 0.73 between these quantities.

Figure 3: Population responses to risky and ambiguous lotteries differ, but are identical for ambiguous and compound lotteries.




Notes: The first two columns are certainty equivalents for a lottery paying 0/100 points. This is the natural scale for the 20-ball urns; for the 30-ball urns the expressed certainty equivalent is divided by 1.5, implying a CRRA utility function. The final column subtracts the certainty equivalent of the risky lottery from those of the compound and ambiguous lotteries.

the same correlation of 0.85 once measurement error is taken into account. Moreover, the 90% confidence interval for this value is (0.71, 0.995). Thus, while we can reject the null hypothesis that the correlation is 1, we cannot reject the null that the correlation is very close to 1.[37]

Here, unlike in Section 3, our large sample size is likely the reason we find any difference at all. Drawing a random sample of 104 observations (the size of Halevy's experiment) 10,000 times, the correlation between ambiguity aversion and reaction to compound lotteries differs from 1 only 1.2% of the time at the 1% level, 4.8% of the time at the 5% level, and 9.0% of the time at the 10% level. It should be noted that these are the results from standard confidence intervals that are well-calibrated. While correlations have an upper bound of 1, and thus suffer from the Andrews's (2001) problem, our estimator can take on values greater than 1, and is normally distributed around 1 when that is the true correlation. Thus, any time a correlation greater than 1 is calculated using ORIV, this should be interpreted as strong evidence that the correlation is actually 1, rather than as a statistical issue.

As a final way of showing how closely ambiguity aversion and reaction to compound lotteries are related, we plot the CDFs of the various measures in Figure 3. The left-most panels show certainty equivalents for risky urns and ambiguous urns—the fact that these diverge below 50 is evidence of ambiguity aversion. The center panels show the certainty equivalents for ambiguous urns and compound urns: the distributions are visually and statistically identical. The final panels show the distributions of ambiguity aversion and reaction to compound lotteries. Once again, these are visually and statistically identical.

## 5.2 Substantive Implications

Ambiguity aversion and reaction to compound lotteries appear remarkably similar once measurement error is accounted for. A plausible conclusion is that sufficiently complex lotteries,

[37]The 95% confidence interval is (0.68, 1.02) and the 99% confidence interval is (0.63, 1.08).

involving the compounding of probabilities, are cognitively equivalent to ambiguous ones—individuals (even Caltech students) have a hard time envisioning the precise resulting probabilities of each outcome. Such an interpretation is in line with the original description of Knightian Uncertainty, "[T]he essential fact is that 'risk' means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomena depending on which of the two is really present and operating." (Knight, 1921). Regardless of the philosophical interpretation of these results, it is certainly clear that any successful model of ambiguity aversion should also be useful for predicting behavior in complex, but fully specified, risky environments.

# 6 Conclusion

If measurement error is such a ubiquitous, but easily correctable, issue, why has the experimental literature paid so little attention to it? The answer likely lies in the fact that attenuation bias driven by measurement error is a "conservative bias". That is, it biases the researcher against Type I errors. So when asked about measurement error, a researcher can confidently answer that it would "go against finding anything."

However, measurement error creates another, field-wide, issue that has been little appreciated. It leads to the over-identification of "new" phenomena. Indeed, our paper shows two examples of this: previously, competitiveness was thought to have a component unconnected to risk aversion and overconfidence, and ambiguity aversion and reaction to compound lotteries (the latter a form of risk aversion) were thought to be very distinct. Our results show that both are unlikely to be true. Moreover, our finding that elicitations of risk attitudes are more highly correlated than previously appreciated suggests that fewer elicitation methods are needed, and that differences in risk attitudes across domains are not as large as pre-

viously thought. Given that these are the only three results we have examined in trying to understand the influence of measurement error in experiments, it seems likely that the over-identification of new phenomena is a substantial problem.

That measurement error may lead to the identification of new phenomena where none exist may feed into the recent mushrooming of methodological work suggesting the high rates of non-replicability of research discoveries (see Ioannidis, 2005; Simonsohn, 2015, and references therein). Using the techniques developed here to account for measurement error may help researchers discover, in a more robust fashion, the deep connections between different attitudes and effects.

# References


**Abbasi, Fahim, Byron William Brown, Cindy Lamendola, Tracey McLaughlin, and Gerald M Reaven**, "Relationship between Obesity, Insulin Resistance, and Coronary Heart Disease Risk," *Journal of the American College of Cardiology*, 2002, *40* (5), 937–943.

**Adcock, Robert James**, "A Problem in Least Squares," *The Analyst*, 1878, *5* (2), 53–54.

**Agranov, Marina and Leeat Yariv**, "Collusion through Communication in Auctions," 2015. California Institute of Technology, *mimeo.*

**Anderson, Lisa R. and Jennifer M. Mellor**, "Are Risk Preferences Stable? Comparing an Experimental Measure with a Validated Survey-based Measure," *Journal of Risk and Uncertainty*, 2009, *39* (2), 137–160.

**Andrews, Donald W.K.**, "Testing when a Parameter is on the Boundary of the Maintained Hypothesis," *Econometrica*, 2001, *68* (2), 683–734.

**Angrist, Joshua and Alan B. Krueger**, "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 2001, *15* (4), 69–85.

**Barber, Brad M. and Terrance Odean**, "The Behavior of Individual Investors," in George M. Constantinides, Milton Harris, and Rene M. Stulz, eds., *Handbook of Economics of Finance*, Vol. 2B: Asset Pricing, Oxford, UK: North-Holland, 2013, pp. 1533–1570.

**Battalio, Raymond C., John H. Kagel, Robin C. Winkler, Edwin B. Fisher, Robert L. Basmann, and Leonard Krasner**, "A Test of Consumer Demand Theory using Observations of Individual Consumer Purchases," *Western Economic Journal*, 1973, *11* (4), 411–428.

**Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen**, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 2013, *81* (2), 608–650.

**_ , _ , and Lie Wang**, "Square-Root LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 2011, *98* (4), 791–806.

**Bertrand, Marianne and Sendhil Mullainathan**, "Do People Mean what they Say? Implications for Subjective Survey Data," *American Economic Review (Papers & Proceedings)*, 2001, *91* (2), 67–72.

**Blattman, Christopher, Julian C. Jamison, Tricia Koroknay-Palicz, Katherine Rodrigues, and Margaret Sheridan**, "Measuring the Measurement Error: A Method to Qualitatively Validate Survey Data," 2015. NBER Working Paper Series # 21447.

**Bound, John, Charles Brown, and Nancy Mathiowetz**, "Measurement Error in Survey Data," in James J. Heckman, ed., *Handbook of Econometrics*, Vol. 5, Amsterdam, The Netherlands: Elsevier, 2001, chapter 59, pp. 3705–3843.

**Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek**, "Gender, Competitiveness, and Career Choices," *Quarterly Journal of Economics*, 2014, *129* (3), 1409–1447.

**Byrnes, James P, David C Miller, and William D Schafer**, "Gender Differences in Risk Taking: A Meta-analysis.," *Psychological Bulletin*, 1999, *125* (3), 367–383.

**Candes, Emmanuel and Terence Tao**, "The Dantzig Selector: Statistical Estimation when $p$ is much Larger than $n$," *Annals of Statistics*, 12 2007, *35* (6), 2313–2351.

**Carroll, Raymond J. and Leonard A. Stefanski**, "Measurement Error, Instrumental Variables and Corrections for Attenuation with Applications to Meta-analyses," *Statistics in Medicine*, 1994, *13* (12), 1265–1282.

**Castillo, Marco, Jeffrey L. Jordan, and Ragan Petrie**, "Children's Rationality, Risk Attitudes, and Misbehavior," 2015. George Mason University, *mimeo.*

**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.

**Coffman, Lucas and Paul Niehaus**, "Pathways to Persuasion," 2015. Ohio State University, *mimeo.*

**Cohen, Jacob**, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition ed., Mahwah, New Jersey: Lawrence Earlbaum Associates, 1988.

**Condon, David M and William Revelle**, "The International Cognitive Ability Resource: Development and Initial Validation of a Public-domain Measure," *Intelligence*, 2014, *43* (2), 52–64.

**Croson, Rachel and Uri Gneezy**, "Gender Differences in Preferences," *Journal of Economic Literature*, 2009, *47* (2), 448–474.

**Dave, Chetan, Catherine C. Eckel, Cathleen A. Johnson, and Christian Rojas**, "Eliciting risk preferences: When is simple better?," *Journal of Risk and Uncertainty*, 2010, *41* (3), 219–243.

**Deck, Cary A., Jungmin Lee, Javier A. Reyes, and Chris Rosen**, "Measuring Risk Attitudes Controlling for Personality Traits," 2008. SSRN working paper #1148521.

**Deck, Cary, Jungmin Lee, Javier Reyes, and Chris Rosen**, "Measuring Risk Aversion on Multiple Tasks: Can Domain Specific Risk Attitudes Explain Apparently Inconsistent Behavior?," 2010. University of Arkansas, *mimeo.*

**Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, 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.

**Durbin, James**, "Errors in Variables," *Revue de l'Institut International de Statistique*, 1954, *22* (1/3), 23–32.

**Eckel, Catherine C. and Philip J. Grossman**, "Sex Differences and Statistical Stereotyping in Attitudes toward Financial Risk," *Evolution and Human Behavior*, 2002, *23* (4), 281–295.

_ **and** _ , "Forecasting Risk Attitudes: An Experimental Study Using Actual and Forecast Gamble Choices," *Journal of Economic Behavior & Organization*, 2008, *68* (1), 1–17.

**Eckel, Catherine C and Philip J Grossman**, "Men, Women and Risk Aversion: Experimental Evidence," in Charles R. Plott and Vernon L. Smith, eds., *Handbook of Experimental Economics Results*, Vol. 1, North-Holland, 2008, pp. 1061–1073.

**Einav, Liran, Amy Finkelstein, Iuliana Pascu, and Mark Cullen**, "How General Are Risk Preference? Choices under Uncertainty in Different Domains," *American Economic Review*, 2012, *101* (2), 2606–2638.

**Ellsberg, Daniel**, "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics*, 1961, *75* (4), 643–669.

**Embrey, Lori L. and Jonathan J. Fox**, "Gender Differences in the Investment Decision-making Process," *Financial Counseling and Planning*, 1997, *8* (2), 33–40.

**Engel, Christoph**, "Dictator Games: A Meta Study," *Experimental Economics*, 2011, *14* (4), 583–610.

**Evans, James D.**, *Straightforward Statistics for the Behavioral Sciences*, Brooks/Cole Publishig Company, 1996.

**Falk, Armin, 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.

**Fan, Jianqing and Runze Li**, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 2001, *96* (456), 1348–1360.

_ **and Yuan Liao**, "Endogeneity in High Dimensions," *Annals of Statistics*, 2014, *42* (3), 872–917.

**Farrell, James**, "Demographics of Risky Investing," *Research in Business and Economics Journal*, 2011, *Special Edition.*

**Fiske, Susan T., Daniel T. Gilbert, and Gardner Lindzey**, *Handbook of Social Psychology*, 5th ed., Vol. 1, Hoboken, NJ: John Wiley & Sons, 2010.

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

**Friedman, Daniel, R Mark Isaac, Duncan James, and Shyam Sunder**, *Risky Curves: On the Empirical Failure of Expected Utility*, Routledge, 2014.

**Friedman, Milton**, *A Theory of the Consumtion Function*, Princeton, New Jersey: Princeton University Press, 1957.

**Frisch, Ragnar**, *Statistical Confluence analysis by Means of Complete Regression Systems*, Universitetets Økonomiske Instituut, 1934.

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

_ **, Muriel Niederle, Aldo Rustichini et al.**, "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, 2003, *118* (3), 1049–1074.

**Greene, William H.**, *Econometric Analysis*, 7th ed., Prentice Hall, 2011.

**Greenland, Sander**, "An Introduction to Instrumental Variables for Epidemiologists," *International Journal of Epidemiology*, 2000, *29* (4), 722–729.

**Halevy, Yoram**, "Ellsberg Revisited: An Experimental Study," *Econometrica*, 2007, *75* (2), 503–536.

**Harless, David W. and Colin F. Camerer**, "The Predictive Utility of Ceneralized Expected Utility Theories," *Econometrica*, 1994, *62* (6), 1251–1289.

**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.

**Hausman, Jerry A.**, "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 2001, *15* (4), 57–67.

**Hertz, Tom, Tamara Jayasundera, Patrizio Piraino, Sibel Selcuk, Nicole Smith, and Alina Verashchagina**, "The Inheritance of Educational Inequality: International Comparisons and Fifty-year Trends," *The BE Journal of Economic Analysis & Policy*, 2007, *7* (2), Article 10.

**Hey, John D.**, *Experiments in Economics*, Blackwell Publishing, 1991.

**Holt, Charles A. and Susan K. Laury**, "Risk Aversion and Incentive Effects," *American Economic Review*, 2002, *92* (5), 1644–1655.

_ **and** _ , "Assessment and Estimation of Risk Preferences," in Mark J. Machina and W. Kip Viscusi, eds., *Handbook of Economics of Risk and Uncertainty*, Vol. 1, North-Holland, 2014, pp. 135–202.

**Ioannidis, John P.A.**, "Why most Published Research Findings are False," *Chance*, 2005, *18* (4), 40–47.

**Knight, Frank H**, *Risk, Uncertainty and Profit*, New York: Hart, Schaffner and Marx, 1921.

**Koopmans, Tjalling Charles**, *Tanker Freight Rates and Tankship Building: An Analysis of Cyclical Fluctuations, by Dr. T. Koopmans*, De erven F. Bohn nv, 1939.

**Kruger, Daniel J., Xiao-Tian Wang, and Andreas Wilke**, "Towards the Development of an Evolutionarily Valid Domain-specific Risk-taking Scale," *Evolutionary Psychology*, 2007, *5* (3), 555–568.

**Leeb, Hannes and Benedikt M Pötscher**, "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 2005, *21* (1), 21–59.

**McKelvey, Richard D. and Thomas R. Palfrey**, "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, 1995, *10* (1), 6–38.

_ **and** _ , "Quantal Response Equilibria for Extensive Form Games," *Experimental Economics*, 1998, *1* (1), 9–41.

**Moore, Don A. and Paul J. Healy**, "The Trouble with Overconfidence," 2007. Carnegie Mellon University, *mimeo.*

_ **and** _ , "The Trouble with Overconfidence," *Psychological Review*, 2008, *115* (2), 502–517.

**Mukerji, Sujoy**, "A Survey of Some Applications of the Idea of Ambiguity Aversion in Economics," *International Journal of Approximate Reasoning*, 2000, *24* (2–3), 221–234.

**Nagel, Rosemarie**, "Unraveling in Guessing Games: An Experimental Study," *The American Economic Review*, 1995, *85* (5), 1313–1326.

**Niederle, Muriel**, "Gender," in John H. Kagel and Alvin E. Roth, eds., *Handbook of Experimental Economics, Volume 2*, Elsevier, 2015.

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

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

_ **and Mark Dean**, "Is it All Connected? A Testing Ground for Unified Theories of Behavioral Economics Phenomena," 2015. Columbia University, *mimeo*.

**Pawar, Pradeep K and Abhilasha Dadhich**, "Study of Correlation between Human Height and Foot Length in Residents of Mumbai," *International Journal of Biological and Medical Research*, 2012, *3* (3), 2232–2235.

**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.

**Reiersøl, Olav**, "Confluence Analysis by means of Lag Moments and Other Methods of Confluence Analysis," *Econometrica*, 1941, *9* (1), 1–24.

_ , *Confluence Analysis by means of Instrumental Sets of Variables*, Stockholm, Sweden: Almqvist & Wiksell, 1945.

_ , "Identifiability of a Linear Relation between Variables which are Subject to Error," *Econometrica*, 1950, *18* (4), 375–389.

**Sargan, John D**, "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 1958, pp. 393–415.

**Segal, Uzi**, "The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach," *International Economic Review*, 1987, *28* (1), 175–202.

_ , "Two-Stage Lotteries without the Reduction Axiom," *Econometrica*, March 1990, *58* (2), 349–377.

**Simonsohn, Uri**, "Small Telescopes: Detectability and the Evaluation of Replication Results," *Psychological Science*, 2015, *26* (5), 559–569.

**Spearman, Charles**, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, 1904, *15* (1), 72–101.

**Tibshirani, Robert**, "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, *58* (1), 267–288.

**Van de Geer, Sara, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure**, "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 2014, *42* (3), 1166–1202.

**Wald, Abraham**, "The Fitting of Straight Lines if both Variables are Subject to Error," *The Annals of Mathematical Statistics*, 1940, *11* (3), 284–300.

**Weber, Elke U., Ann-Renee Blais, and Nancy E. Betz**, "A Domain-specific Risk-attitude Scale: Measuring Risk Perceptions and Risk Behaviors," *Journal of Behavioral Decision Making*, 2002, *15* (4), 263–290.

**Wright, Charlotte M. and Tim D. Cheetham**, "The Strengths and Limitations of Parental Heights as a Predictor of Attained Height," *Archives of Disease in Childhood*, 1999, *81* (3), 257–260.

# Online Appendix—Not Intended for Publication

# A Sampling Properties of the ORIV Estimator

There are several textbooks and survey articles that present methods for inference in linear and non-linear models with measurement error, a topic that is also addressed in almost any statistics and econometrics textbook. A common approach to this problem exploits repeated measurements (or replicates) to characterize the severity of measurement error. After estimating these distributional features of the measurement error, researchers can then compute correction factors that dis-attenuate sample estimates. Our ORIV method provides a simple unifying method for consistent estimation that integrates all information available for relating two variables to one another. To our knowledge, it is the first to do so. In this technical appendix, we present a standard suite of results establishing consistency and asymptotic normality, along with consistent standard errors, for the ORIV estimator. These properties are proved using textbook-standard arguments presented primarily for completeness.

## A.1 Assumptions for the ORIV Model

To establish the Sampling Properties of ORIV, we first detail the assumptions underlying the linear regression model in latent variables (Assumption 1) and then present the classical measurement error model for our contaminated observations (Assumption 2). The assumptions here are standard, with the usual exogeneity and bounds on moments to admit asymptotic analysis.

**Assumption 1** (Linear Regression Model in Latent Variables)**.** *The relationship between the latent variables $Y^*$ and $X^*$ satisfy the usual assumptions for linear regression:*

1. *Linear model:* $Y^* = \alpha + X^*\beta^* + \varepsilon^*$

2. *Exogeneity:* $\mathbb{E}[\varepsilon^*|X^*] = 0$

3. *Variability in treatment:* $0 < Var(X^*)$, *and,* $\mathbb{E}\left[(X^*)^4\right] < \infty$

4. *Variability in residuals:* $Var[\varepsilon^*|X^*] = \sigma^2_{\varepsilon^*}$, *and,* $\mathbb{E}\left[(\varepsilon^*)^4 |X^*\right] < \infty$

5. *Independent sampling of individuals:* $(Y^*_t, X^*_t, \varepsilon^*_t) \perp (Y^*_s, X^*_s, \varepsilon^*_s)\ \forall t \neq s$.

**Assumption 2** (Replicated Classical Measurement Error)**.** *We observe multiple replicates for noisy measurements of* $Y^*$ *and* $X^*$ *that are tainted by classical measurement error. We assume there exists* $\Delta > 0$ *such that:*

1. $X^k = X^* + \nu^k_X,\ k = 1, \dots K_X$

   (a) $\mathbb{E}\left[\nu^k_X|X^*, Y^*\right] = 0$

   (b) $\mathbb{E}\left[(\nu^k_X)^2|X^*, Y^*\right] = \sigma^2_{\nu^k_X}$

   (c) $\mathbb{E}\left[(\nu^k_X)^4|X^*, Y^*\right] < \infty$

   (d) $\nu^k_X \perp \nu^l_X,\ \forall k \neq l$

2. $Y^k = Y^* + \nu^k_Y\ k = 1, \dots K_Y$

   (a) $\mathbb{E}\left[\nu^k_Y|X^*, Y^*\right] = 0$

   (b) $\mathbb{E}\left[(\nu^k_Y)^2|X^*, Y^*\right] = \sigma^2_{\nu^k_Y}$

   (c) $\mathbb{E}\left[(\nu^k_Y)^4|X^*, Y^*\right] < \infty$

   (d) $\nu^k_Y \perp \nu^l_Y,\ \forall k \neq l$

3. *Independence across measures:* $\nu^k_Y \perp \nu^l_X,\ \forall k, l$

4. *Independent sampling of individuals:* $\forall s, t \in \{1, \dots, N\}$, *s.t.* $t \neq s$*:*

$$\nu^k_{X,t} \perp \nu^l_{X,s}, \forall k, l \in \{1, \dots, K_X\}\text{, and, } \nu^k_{Y,t} \perp \nu^l_{Y,s}, \forall k, l \in \{1, \dots, K_Y\}$$

Our objective is to perform consistent and (potentially) efficient inference on $\beta$, for which we propose the ORIV regression. Defining $1_N$ as a $(N \times 1)$ vector of 1's and $0_N$ as a $(N \times 1)$

vector of 0's, the independent and dependent variables are:

$$\tilde{Y} \equiv \left[Y^{1\prime}, \ldots, Y^{K_Y\prime}\right]' \qquad \tilde{X}^k \equiv 1_{K_Y} \otimes X^k$$
$$Y_{OR} \equiv 1_{K_X} \otimes Y \qquad X_{OR} \equiv \left[\tilde{X}^{1\prime}, \ldots, \tilde{X}^{K_X\prime}\right]'$$

Now construct the instruments for each model:

$$W^k \equiv \left[X^1, \ldots, X^{k-1}, 0_N, X^{k+1}, \ldots, X^{K_X}\right] \qquad W_{OR} \equiv \begin{bmatrix} \tilde{W}^1 & 0 & \cdots & 0 \\ 0 & \tilde{W}^2 & \ddots & 0 \\ 0 & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & \tilde{W}^{K_X} \end{bmatrix}$$
$$\tilde{W}^k \equiv 1_{K_Y} \otimes W^k$$

Then, as we soon show, the ORIV estimator fits the linear model with the exogeneity restriction:

$$Y_{OR} = X_{OR}\beta^* + \varepsilon_{OR},\ \mathbb{E}[\varepsilon_{OR}|W_{OR}] = 0.$$

Letting $P_{W_{OR}} \equiv W_{OR}\left(W_{OR}'W_{OR}\right)^{-1}W_{OR}'$ denote the projection matrix onto the column space of $W_{OR}$, we can write the ORIV estimator for $\beta$ using the usual formula:

$$\hat{\beta}^* = \left(X_{OR}'P_{W_{OR}}X_{OR}\right)^{-1}X_{OR}'P_{W_{OR}}Y_{OR}. \tag{8}$$

## A.2 ORIV Estimator Consistency

We now present the standard arguments for IV estimator consistency in the ORIV setting, establishing the result stated in the text's Proposition 1:

**Proposition 1.** *ORIV produces consistent estimates of $\beta^*$.*

*Proof.* We begin by verifying the exogeneity condition $\mathbb{E}[\varepsilon_{OR,n}|W_{OR,n}] = 0$. This requires breaking $\varepsilon_{OR}$ down equation-by-equation to verify the condition for each IV specification

included in the model.

$$\begin{aligned}\varepsilon^{i,j} &= Y^i - X^j\beta^* \\ &= Y^* + \nu_Y^i - X^*\beta^* - \nu_X^j\beta^* \\ &= X^*\beta^* + \varepsilon^* + \nu_Y^i - X^*\beta^* - \nu_X^j\beta^* \\ &= \varepsilon^* + \nu_Y^i - \nu_X^j\beta^*.\end{aligned}$$

Note the ORIV estimator interacts $\varepsilon^{(i,j)}$ solely with $W^{(j)}$, so that:

$$\begin{aligned}\mathbb{E}\left[\varepsilon^{i,j}|W^j\right] &= \mathbb{E}\left[\varepsilon^* + \nu_Y^i - \nu_X^j\beta^*|W^j\right] \\ &= \cancel{\mathbb{E}\left[\varepsilon^*|W^j\right]} + \cancel{\mathbb{E}\left[\nu_Y^i|W^j\right]} + \cancel{\mathbb{E}\left[\nu_X^j|W^j\right]}\beta_{OR} \\ &= 0.\end{aligned}$$

The first term above cancels by the exogeneity condition in assumption 1.2, the second by the classical measurement error assumption 2.3, and the last by assumptions 2.1(a) and 2.1(d) because the $k^{th}$ column of $W^k$ is zeroed out.

We now recall the formula for $\hat{\beta}^*$ from equation 8:

$$\begin{aligned}\hat{\beta}^* &= \left(X_{OR}'P_{W_{OR}}X_{OR}\right)^{-1}X_{OR}'P_{W_{OR}}Y_{OR} \\ &= \left(X_{OR}'P_{W_{OR}}X_{OR}\right)^{-1}X_{OR}'P_{W_{OR}}X_{OR}\beta^* + \left(X_{OR}'P_{W_{OR}}X_{OR}\right)^{-1}X_{OR}'P_{W_{OR}}\varepsilon_{OR} \\ &= \beta^* + \left(X_{OR}'P_{W_{OR}}X_{OR}\right)^{-1}X_{OR}'P_{W_{OR}}\varepsilon_{OR}.\end{aligned}$$

To establish consistency of the estimator, notice that:

$$\begin{aligned}\hat{\beta}^* - \beta^* &= \left(X'_{OR} P_{W_{OR}} X_{OR}\right)^{-1} X'_{OR} P_{W_{OR}} \varepsilon_{OR} \\ &= \left(\frac{1}{N} X'_{OR} W_{OR} \left(\frac{1}{N} W'_{OR} W_{OR}\right)^{-1} \frac{1}{N} W'_{OR} X_{OR}\right)^{-1} \\ &\quad * \ \frac{1}{N} X'_{OR} W_{OR} \left(\frac{1}{N} W'_{OR} W_{OR}\right)^{-1} \frac{1}{N} W'_{OR} \varepsilon_{OR}.\end{aligned}$$

The bounded fourth moments in assumptions 1.4, 2.1(c), and 2.2(c) allow us to apply a strong law of large numbers for each of the averages in the above formula. Further, since $\mathbb{E}[\varepsilon_{OR,n}|W_{OR,n}] = 0 \Rightarrow \mathbb{E}[W_{OR,n}\varepsilon_{OR,n}] = 0$, the last term $\frac{1}{N} W'_{OR}\varepsilon_{OR} \to_p 0$, yielding our desired consistency result: $\left(\hat{\beta}^* - \beta^*\right) \to_p 0$. □

## A.3 Asymptotic Normality and Clustered Variances

We now establish asymptotic normality and present a consistent estimator for the variance of $\hat{\beta}^*$. This establishes the proof for Proposition 2:

**Proposition 2.** *The ORIV estimator satisfies asymptotic normality under Assumptions (1) and (2). The estimated standard errors, when clustered by participant, are consistent for the asymptotic standard errors.*

*Proof.* First we establish asymptotic normality. Let $\Sigma_{\varepsilon_{OR}} \equiv E\left[\varepsilon_{OR}\varepsilon'_{OR}\right]$. By laws of large numbers,

$$S_{X'W} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{N} X'_{OR} W_{OR} \ , \ \ S_{W'W} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{N} W'_{OR} W_{OR} \text{ , and } \Omega_{OR} \equiv \operatorname*{plim}_{n\to\infty} \frac{1}{N} W'_{OR} \Sigma_{\varepsilon_{OR}} W_{OR}.$$

From the Central Limit Theorem, $N^{-1/2} W'_{OR}\varepsilon_{OR} \to_d N\left(0, \Omega_{OR}\right)$

Then asymptotic normality of $\hat{\beta}_{OR}$ now follows since:

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}^* - \beta^*\right) &= \left(\frac{1}{N}X'_{OR}W_{OR}\left(\frac{1}{N}W'_{OR}W_{OR}\right)^{-1}\frac{1}{N}W'_{OR}X_{OR}\right)^{-1} \\ &\quad * \ \frac{1}{N}X'_{OR}W_{OR}\left(\frac{1}{N}W'_{OR}W_{OR}\right)^{-1}N^{-1/2}W'_{OR}\varepsilon_{OR} \\ &\to_p N\left(0, \Sigma_{\hat{\beta}^*}\right), \text{ where,} \\ \Sigma_{\hat{\beta}^*} &= \left(S_{X'W}S_{W'W}^{-1}S'_{X'W}\right)^{-1}S_{X'W}S_{W'W}^{-1}\Omega S_{W'W}^{-1}S'_{X'W}\left(S_{X'W}S_{W'W}^{-1}S'_{X'W}\right)^{-1}.\end{aligned}$$

A feasible estimator of the asymptotic variance requires an estimate for $\Omega$, which we will show is available using the usual clustered variance-covariance matrix estimator. To achieve this result, we characterize the structure of $\Sigma_{\varepsilon_{OR}}$. Consider

$$\begin{aligned}\mathbb{E}\left[\varepsilon_s^{i,j}\varepsilon_t^{k,l}\right] &= \mathbb{E}\left[\left(\varepsilon_s^* + \nu_{Y,s}^i - \nu_{X,s}^j\beta^*\right)\left(\varepsilon_t^* + \nu_{Y,t}^k - \nu_{X,t}^l\beta^*\right)\right] \\ &= \mathbb{E}\big[\varepsilon_s^*\varepsilon_t^* + \cancel{\varepsilon_s^*\nu_{Y,t}^k} - \cancel{\varepsilon_s^*\nu_{X,t}^l}\beta^* \\ &\qquad + \cancel{\nu_{Y,s}^i\varepsilon_t^*} + \nu_{Y,s}^i\nu_{Y,t}^k - \cancel{\nu_{Y,s}^i\nu_{X,t}^l}\beta^* \\ &\qquad - \cancel{\nu_{X,s}^j\varepsilon_t^*} - \cancel{\nu_{X,s}^j\nu_{Y,t}^k} + \nu_{X,s}^j\nu_{X,t}^l(\beta^*)^2\big] \\ &= \mathbb{E}\left[\varepsilon_s^*\varepsilon_t^* + \nu_{Y,s}^i\nu_{Y,t}^k + \nu_{X,s}^j\nu_{X,t}^l(\beta^*)^2\right].\end{aligned}$$

Therefore, we have:

$$\mathbb{E}\left[\varepsilon_s^{i,j}\varepsilon_t^{k,l}\right] = \begin{cases} 0, & \text{if } s \neq t, \forall i,j,k,l \\ s_{00} \equiv \sigma_{\varepsilon^*}^2, & \text{if } s = t, i \neq k, j \neq l \\ s_{i0} \equiv \sigma_{\varepsilon^*}^2 + \sigma_{\nu_Y^i i}^2, & \text{if } s = t, i = k, j \neq l \\ s_{0j} \equiv \sigma_{\varepsilon^*}^2 + \sigma_{\nu_X^j}^2(\beta^*)^2, & \text{if } s = t, i \neq k, j = l \\ s_{ij} \equiv \sigma_{\varepsilon^*}^2 + \sigma_{\nu_Y^i}^2 + \sigma_{\nu_X^j}^2(\beta^*)^2, & \text{if } s = t, i = k, j = l \end{cases}. \tag{9}$$

This formula allows us to populate the entries of $\Sigma_{\varepsilon_{OR}}$. Denote by $S_Y$ the diagonal matrix with entries $\left[\sigma^2_{\nu_Y^1}, \ldots, \sigma^2_{\nu_Y^{K_Y}}\right]$ and, similarly, by $S_X$ the diagonal matrix with entries $\left[\sigma^2_{\nu_X^1}, \ldots, \sigma^2_{\nu_X^{K_X}}\right]$. Letting $I_N$ denote the $N \times N$ Identity matrix and working with Kronecker products, we can write:

$$\begin{aligned}\Sigma_{\varepsilon_{OR}} =& \sigma^2_{\varepsilon^*}\left[\left(1_{(K_X K_Y)} 1'_{(K_X K_Y)}\right) \otimes I_N\right] \\ &+ \left[S_Y \otimes \left(\left(1_{K_X} 1'_{K_X}\right) \otimes I_N\right)\right] \\ &+ \beta^{*2}\left[\left(\left(1_{K_Y} 1'_{K_Y}\right) \otimes S_X\right) \otimes I_N\right].\end{aligned} \tag{10}$$

Importantly, the result indicates that all of the non-zero entries in $\Sigma_{\varepsilon_{OR}}$ correspond to instances where the individual representing the unit of observation is the same in two different regression matrices. That is, the covariance matrix for residuals is striped, with heteroskedastic clusters on each individual as a unit of observation.

Define $\hat{\varepsilon}_{OR,(i)} \equiv Y_{OR,i} - X_{OR,i}\hat{\beta}^*$ and consider the estimator for $\hat{\Sigma}_{\varepsilon_{OR}}$ that set its $(i,j)$ entry equal to $\hat{\varepsilon}_{OR,(i)}\hat{\varepsilon}_{OR,(i)}$ if the $(i,j)$ entry in $\Sigma_{\varepsilon_{OR}}$ is non-zero. Then, by the Law of Large Numbers:

$$\begin{aligned}\hat{\Omega} &\equiv \frac{1}{N} W'_{OR}\hat{\Sigma}_{\varepsilon_{OR}} W_{OR} \to_p \Omega \\ N\hat{\Sigma}_{\hat{\beta}^*} &\equiv N\frac{X'_{OR}W_{OR}\left(W'_{OR}W_{OR}\right)^{-1}\hat{\Omega}\left(W'_{OR}W_{OR}\right)^{-1}W'_{OR}X_{OR}}{\left(X'_{OR}P_{W_{OR}}X_{OR}\right)^2} \to_p \Sigma_{\hat{\beta}_{OR}}.\end{aligned}$$

□

## A.4 Correlation Consistency and Asymptotic Normality

We now establish consistency and asymptotic normality of the correlation estimator as stated in Proposition 3:

**Proposition 3.** *$\hat{\rho}^*_{XY}$ is consistent. Standard errors estimated from ORIV, multiplied by $\sqrt{\widehat{Cov}[X^a, X^b]/\widehat{Cov}[Y^a, Y^b]}$, are consistent.*

*Proof.* This result follows from a straightforward application of the Continuous Mapping Theorem. In particular, note that:

$$\begin{array}{l}\widehat{\text{Cov}}[X^a, X^b] \to_p \text{Cov}[X^a, X^b] = \sigma^2_{X^*} \\ \widehat{\text{Cov}}[Y^a, Y^b] \to_p \text{Cov}[Y^a, Y^b] = \sigma^2_{Y^*}\end{array} \Rightarrow \sqrt{\frac{\widehat{\text{Cov}}[X^a, X^b]}{\widehat{\text{Cov}}[Y^a, Y^b]}} \to_p \sqrt{\frac{\text{Cov}[X^a, X^b]}{\text{Cov}[Y^a, Y^b]}} = \frac{\sigma_{X^*}}{\sigma_{Y^*}}.$$

Given the asymptotic normality of $\hat{\beta}^*$ and its consistency for $\beta^* = \frac{\text{Cov}(X^*, Y^*)}{\sigma^2_{X^*}}$, we have:

$$\hat{\rho}^*_{XY} = \sqrt{\frac{\widehat{\text{Cov}}[X^a, X^b]}{\widehat{\text{Cov}}[Y^a, Y^b]}}\hat{\beta}^* \to_p \frac{\text{Cov}\left(X^*, Y^*\right)}{\sigma_{X^*}\sigma_{Y^*}} = \rho^*_{XY}.$$

Now, using Slutsky's Theorem:

$$\sqrt{N}\left(\hat{\rho}^*_{XY} - \rho^*_{XY}\right) = \sqrt{\frac{\widehat{\text{Cov}}[X^a, X^b]}{\widehat{\text{Cov}}[Y^a, Y^b]}}\sqrt{N}\left(\hat{\beta}^* - \beta^*\right) \to_d N\left(0, \frac{\sigma_{X^*}}{\sigma_{Y^*}}\Sigma_{\hat{\beta}^*}\right),$$

and a consistent asymptotic variance estimate is available:

$$\sqrt{\frac{\widehat{\text{Cov}}[X^a, X^b]}{\widehat{\text{Cov}}[Y^a, Y^b]}}\hat{\Sigma}_{\hat{\beta}^*} \to_p \frac{\sigma_{X^*}}{\sigma_{Y^*}}\Sigma_{\hat{\beta}^*}$$

□

In the asymptotic approximation, the sampling error in $\text{Cov}[X^a, X^b]$ and $\text{Cov}[Y^a, Y^b]$ makes a negligible contribution to the variance of the estimated correlation. Therefore, using a bootstrap to compute standard errors might be warranted in smaller samples. Our simulation exercises, however, have indicated the asymptotic approximation performs reasonably well.

## A.5 Equivalent Estimators

The ORIV estimator provides a convenient and intuitive representation for consolidating all the information available to the experimenter. This convenience brings with it some redundancy and so we highlight here a couple of special cases yielding numerically equivalent estimators to ORIV. These equivalences also motivate an interpretation of ORIV as a model combination estimator that will be useful in characterizing the estimator's efficiency.

### A.5.1 Averaging Left-Hand-Side Variables: $K_Y \geq 2$

Our presentation in the main body of the text proposed stacking additional elicitations $Y^1, \ldots, Y^{K_Y}$ for $Y^*$ to consolidate the information available from these different elicitations. This representation is inherently overspecified and, in fact, the general variance-covariance matrix for residuals will be singular. The singularity arises because the variance-covariance matrix of residuals accounts for $N \times K_X \times K_Y$ effective observations but we only observe $N \times (K_Y + K_X)$ data points, only $N \times (K_Y + K_X - 1)$ of which are informative for exogenous variation in the noisily measured treatment variable. This singularity can be easily resolved in a numerically equivalent representation of the model that takes the simple average of the elicitations, $\bar{Y} = \frac{1}{K_Y} \sum_{k=1}^{K_Y} Y^k$, as the only measure for $Y^*$.

**Proposition 4.** *Suppose $K_Y \geq 2$, let $\bar{Y} = \frac{1}{K_Y} \sum_{k=1}^{K_Y} Y^k$, and define:*

$$\begin{aligned} \bar{Y}_{OR} &= 1_{K_X} \otimes \bar{Y} \\ \bar{X}_{OR} &= \left[X^{1\prime}, \ldots, X^{K_X\prime}\right]' \end{aligned} \qquad \bar{W}_{OR} = \begin{bmatrix} W^1 & 0 & \cdots & 0 \\ 0 & W^2 & \ddots & 0 \\ 0 & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & W^{K_X} \end{bmatrix}.$$

*Then* $\bar{\beta}_{OR} = \left(\bar{X}_{OR}' P_{\bar{W}_{OR}} \bar{X}_{OR}\right)^{-1} \bar{X}_{OR}' P_{\bar{W}_{OR}} \bar{Y}_{OR} = \hat{\beta}^*$.

*Proof.* The first step in the proof shows that $X'_{OR}P_{W_{OR}}X_{OR} = K_Y\bar{X}'_{OR}P_{\bar{W}_{OR}}\bar{X}_{OR}$, which comes immediately from the SUR structure ORIV imposes on the first stage regressions of $X^k$ on $W^k$. Due to the blocked structure of $\bar{W}_{OR}$,

$$\bar{X}'_{OR}P_{\bar{W}_{OR}}\bar{X}_{OR} = \sum_{k=1}^{K_X} X^{k\prime}P_{W^k}X^k.$$

Due to the stacked structure of $\tilde{X}^k$ and $\tilde{W}^k$, $\tilde{X}^{k\prime}P_{\tilde{W}^k}\tilde{X}^k = K_Y X^{k\prime}P_{W^k}X^k$, which when blocked across the measurements of $X^*$, gives:

$$X'_{OR}P_{W_{OR}}X_{OR} = \sum_{k=1}^{K_X} \tilde{X}^{k\prime}P_{\tilde{W}^k}\tilde{X}^k = \sum_{k=1}^{K_X} K_Y X^{k\prime}P_{W^k}X^k = K_Y\bar{X}'_{OR}P_{\bar{W}_{OR}}\bar{X}_{OR}.$$

To complete the proof, we apply a parallel argument to show that $X'_{OR}P_{W_{OR}}Y_{OR} = K_Y\bar{X}'_{OR}P_{\bar{W}_{OR}}\bar{Y}_{OR}$.

□

### A.5.2 Averaging Estimates of Individual IV Models: $K_Y = 1$

In the special case where $K_Y = 1$, either because there is only one measurement for $Y^*$ or because multiple measurements have been concentrated into $\bar{Y}$, the ORIV estimator is numerically equivalent to a weighted average of estimates from the individual IV specifications.

**Proposition 5.** *Let* $\hat{\beta}_k = \left(X^{k\prime}P_{W^k}X^k\right)^{-1}X^{k\prime}P_{W^k}Y$ *and* $\omega_k = \left(X^{k\prime}P_{W^k}X^k\right)$*, then:*

$$\tilde{\beta}^* = \frac{1}{\sum_{k=1}^{K_X}\omega_k}\sum_{k=1}^{K_X}\omega_k\hat{\beta}_k = \hat{\beta}^*.$$

*Proof.* We begin with the observation that $\tilde{W}^k = W^k$ and $\tilde{X}^k = X^k$ when $K_Y = 1$. From

the previous subsection, we can write:

$$\hat{\beta}^* = \left(\sum_{k=1}^{K_X} X^{k\prime} P_{W^k} X^k\right)^{-1} \sum_{k=1}^{K_X} X^{k\prime} P_{W^k} Y = \left(\sum_{k=1}^{K_X} \omega_k\right)^{-1} \sum_{k=1}^{K_X} X^{k\prime} P_{W^k} X^k \hat{\beta}_k = \frac{\sum_{k=1}^{K_X} \omega_k \hat{\beta}_k}{\sum_{k=1}^{K_X} \omega_k} = \tilde{\beta}^*.$$

□

A further specialized case of Proposition 5 applies when $K_X = 2$. In this setting,

$$\omega_1 = X^{1\prime} P_{W^1} X^1 = X^{1\prime} P_{X^2} X^1 = X^{2\prime} P_{X^1} X^2 = X^{2\prime} P_{W^2} X^2 = \omega_2.$$

Consequently, in the setting where $K_Y = 1$ and $K_X = 2$, the ORIV estimator can be computed by taking the simple average of the two IV estimators $\hat{\beta}_1$ and $\hat{\beta}_2$.

## A.6 Estimator Efficiency and GLS Estimation

We close our discussion of the ORIV estimator with a brief comment on the estimator's efficiency. In general, when different measurement errors have different magnitudes, a GLS implementation of ORIV can improve efficiency. However, in the special case where measurement errors are homoskedastic, then the ORIV estimator is asymptotically efficient.

### A.6.1 Efficiency in Homoskedastic Setting: $K_Y = 1$

In this subsection, we consider efficiency when $K_Y = 1$ while maintaining the following homogeneity assumption for variances in the model.

**Assumption 3** (Identical Variances Across Replicates)**.** *Suppose measurement error has the same variance for all replicates, so that:*

1. $\sigma_{\nu,X,k} = \sigma_{\nu,X,j} = \sigma_{\nu,X},\ \forall j,k$
2. $\sigma_{\nu,Y,k} = \sigma_{\nu,Y,j} = \sigma_{\nu,Y},\ \forall j,k.$

Intuitively, efficiency obtains under Assumption 3 because each observation is equally informative and, as such, should be weighted equally. To verify efficiency, we analyze the variance-covariance matrix for estimates from individual models. Let $\hat{B} = \left[\hat{\beta}_1, \ldots, \hat{\beta}_{K_X}\right]'$. The asymptotic variance-covariance matrix for this vector is defined by the limit of the scaled covariances:

$$\begin{aligned} N\text{Cov}\left[\hat{\beta}_i, \hat{\beta}_j\right] &= N\frac{X^{i\prime}P_{W^i}P_{W^j}X^j}{\left(X^{i\prime}P_{W^i}X^i\right)\left(X^{j\prime}P_{\tilde{W}^j}X^{j)}\right)}s_{1(1\{i=j\})} \\ &= \frac{\frac{1}{N}X^{i\prime}P_{W^i}P_{W^j}X^j}{\left(\frac{1}{N}X^{i\prime}P_{W^i}X^i\right)\left(\frac{1}{N}X^{j\prime}P_{\tilde{W}^j}X^{j)}\right)}s_{1(1\{i=j\})} \to_p \zeta_{i,j}. \end{aligned}$$

We will show below that the homoskedastic measurement error guaranteed by Assumption 3 implies $\zeta_{i,j} = \zeta_{k,l}, \forall i \neq j, k \neq l$, and $\zeta_{i,i} = \zeta_{j,j}, \forall i, j$, allowing us to express the asymptotic variance-covariance matrix for $\hat{B}$ as:

$$\Sigma_{\hat{B}}^{\infty} = 1_{K_X}1_{K_X}'\zeta_{1,2} + I_{K_X}\zeta_{1,1}.$$

If we want to form an efficient linear combination of the estimators in $\hat{B}$ while maintaining consistency, we would be looking for a vector of weights, $w = [w_1, w_2, \ldots, w_{K_X}]'$, that sum to one and minimize:

$$\begin{aligned} w^* &= \arg\min_w w'\Sigma_{\hat{B}}^{\infty}w, \text{ such that, } w'1_{K_X} = 1 \\ &\Rightarrow w^* = K_X^{-1}1_{K_X}. \end{aligned}$$

That is, in the homoskedastic setting, equally weighting the estimates of the individually valid IV estimators is asymptotically efficient.

**Proposition 6** (Two-Parameter Covariance Matrix for Individual IV Estimators)**.** *Under Assumptions (1)–(3), the variance-covariance matrix for estimates from the individual IV*

*models for $\beta^*$ features constant correlations and homoskedastic variances. That is,*

$$\Sigma_{\hat{B}}^{\infty} = 1_{K_X} 1'_{K_X} \zeta_{1,2} + I_{K_X} \zeta_{1,1}.$$

*Proof.* By virtue of the homoskedastic measurement error, in the limit:

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} \frac{1}{N} X^{i\prime} P_{W^i} X^i &= \operatorname*{plim}_{N\to\infty} \frac{1}{N} X^{j\prime} P_{W^j} X^j \equiv \omega_{11}, \text{ and} \\
\operatorname*{plim}_{N\to\infty} \frac{1}{N} X^{i\prime} P_{W^i} P_{W^j} X^j &= \operatorname*{plim}_{N\to\infty} \frac{1}{N} X^{k\prime} P_{W^k} P_{W^l} X^l \equiv \omega_{12}, \text{ and},, \forall i \neq j, k \neq l
\end{aligned}$$

The exact formulas for the constants $\omega$ involve some algebraic manipulation. The relatively simple constant for the norm of $X^i$ projected onto its instruments $W^i$ is just the expected $R^2$ of the first stage regression:

$$\omega_{11} = \frac{(K_X - 1)\,\mathbb{E}[(X^*)^2]^2}{(K_X - 1)\,\mathbb{E}[(X^*)^2] + \sigma^2_{\nu_X}}.$$

With respect to $\omega_{12}$, notice that

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} \frac{1}{N} X^{i\prime} W^i &= \mathbb{E}\big[X^i_n W^{i\prime}_n\big] = \mathbb{E}\big[(X^*)^2\big]\, 1'_{K_X-1} \\
\operatorname*{plim}_{N\to\infty} \frac{1}{N} W^{i\prime} W^i &= \mathbb{E}\big[W^i_n W^{i\prime}_n\big] = \mathbb{E}\big[(X^*)^2\big]\, 1_{K_X-1} 1'_{K_X-1} + \sigma^2_{\nu_X} I_{K_X-1} \\
\operatorname*{plim}_{N\to\infty} \frac{1}{N} W^{i\prime} W^j &= \mathbb{E}\big[W^i_n W^{j\prime}_n\big] = \mathbb{E}\big[(X^*)^2\big]\, 1_{K_X-1} 1'_{K_X-1} + \sigma^2_{\nu_X} E_i' E_j
\end{aligned}$$

Here $E_i$ is a $K_X \times (K_X - 1)$ matrix constructed by removing the $i^{th}$ column from the $K_X \times K_X$ identity matrix. Establishing the probability limit result for $\operatorname*{plim}_{N\to\infty} \frac{1}{N} X^{i\prime} P_{W^i} P_{W^j} X^j$ requires chaining together and inverting a series of these expressions:

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} X^{i\prime} P_{W^i} P_{W^j} X^j = \operatorname*{plim}_{N\to\infty} \frac{1}{N} X^{i\prime} W^i \left(\frac{1}{N} W^{i\prime} W^i\right)^{-1} \frac{1}{N} W^{i\prime} W^j \left(\frac{1}{N} W^{j\prime} W^j\right)^{-1} \frac{1}{N} W^{j\prime} X^j$$

Importantly, the dependence on which measurements are involved, $i$ and $j$ is entirely wrapped up in inner products of the form $1'_{K_X-1}E_i'E_j1_{K_X-1} = K_X - 2$ □

### A.6.2 Efficiency in Heteroskedastic Settings: $K_Y = 1$

We now consider efficiency in the heteroskedastic setting, relaxing assumption 3 to allow for the case where $\sigma_{\nu_X^k} \neq \sigma_{\nu_X^j}$ while maintaining the assumption that $K_Y = 1$. In this setting, the heteroskedastic errors will admit a more efficient GLS estimator. Here, $S_X$ is the diagional matrix with $\sigma^2_{\nu_X^k}$ in the $(k,k)^{th}$ entry, so that:

$$\Sigma_{\varepsilon_{OR}}^{K_Y=1} = \left(\sigma^2_{\varepsilon^*} + \sigma^2_{\nu_Y}\right)\left[\left(1_{K_X}1'_{K_X}\right) \otimes I_N\right] + \beta^{*2}\left(S_X \otimes I_N\right)$$

Importantly, even though the off-diagonal terms retain a homoskedastic structure, now the diagonal terms reflect the heteroskedasticity in the measurement error for $X$. It is this heteroskedasticity that allows for enhanced efficiency. For efficient estimation in the presence of heteroskedasticity, the usual formulas for GLS estimation can be used for ORIV.

To characterize the efficient weights associated with each of the individual IV models, we can again consider the model combination exercise pertaining to the combination of estimates in $\hat{B}$ to minimize variance. $\Sigma_{\hat{B}}$ is no longer going to have constant variances and covariances. Without a homogeneity assumption, we cannot further simplify its representation beyond the results above:

$$\begin{aligned} w^* &= \arg\min_w w'\Sigma^{\infty}_{\hat{B}}w, \text{ such that, } w'1_{K_X} = 1 \\ &\Rightarrow w^* = \frac{\Sigma^{\infty -1}_{\hat{B}}1_{K_X}}{1'_{K_X}\Sigma^{\infty -1}_{\hat{B}}1_{K_X}}. \end{aligned}$$

Without reweighting observations, the ORIV estimator will assign weights of $w_{i,OR} = \frac{X^{i\prime}P_{W^i}X_i}{\sum_{k=1}^{K}\left(X^{k\prime}P_{W^k}X_k\right)}$. Therefore, consider reweighting observations for the $i^{th}$ model to achieve

the optimal weights. Specifically, defining $\lambda_i \equiv \sqrt{\frac{w_i^*}{w_{i,OR}}}$ and $\lambda \equiv [\lambda_1, \ldots, \lambda_{K_X}]'$, let

$$\begin{aligned} \tilde{Y}_{OR,\lambda} &= \lambda \otimes Y^1 \\ \tilde{X}_{OR,\lambda} &= \left[\lambda_1 X^{1\prime}, \ldots, \lambda_{K_X} X^{K_X\prime}\right]' \end{aligned} \qquad \tilde{W}_{OR,\lambda} = \begin{bmatrix} \lambda_1 W^1 & 0 & \cdots & 0 \\ 0 & \lambda_2 W^2 & \ddots & 0 \\ 0 & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_{K_X} W^{K_X} \end{bmatrix}.$$

Then $\hat{\beta}^*_{GLS} = \left(\tilde{X}'_{OR,\lambda} P_{\tilde{W}_{OR,\lambda}} \tilde{X}_{OR,\lambda}\right)^{-1} \tilde{X}'_{OR,\lambda} P_{\tilde{W}_{OR,\lambda}} \tilde{Y}_{OR,\lambda}$ is the efficient, asymptotically unbiased estimator for $\beta^*$.

### A.6.3 Efficiency in Heteroskedastic Settings: $K_Y \geq 2$

The efficiency arguments in the homoskedastic setting when $K_Y = 1$ extend immediately to models with $K_Y > 1$ by the equivalence result in Proposition 4. The heteroskedastic setting is slightly complicated because the simple average is no longer the most efficient way to combine the information in the different measurements of $Y$. Instead, we replace the simple average with the efficient average, $\bar{Y}_{GLS} = \frac{1}{\sum_{k=1}^{K_Y} \sigma^{-2}_{\nu^k_Y}} \sum_{k=1}^{K_Y} \sigma^{-2}_{\nu^k_Y} Y^k$.

Having minimized measurement error in the left-hand-side variable, $Y^*$, we can construct the weighted GLS estimator from the previous subsection. Define $\tilde{\bar{Y}}_{OR,\lambda} = \lambda \otimes \bar{Y}_{GLS}$. Then $\hat{\beta}^*_{GLS} = \left(\tilde{X}'_{OR,\lambda} P_{\tilde{W}_{OR,\lambda}} \tilde{X}_{OR,\lambda}\right)^{-1} \tilde{X}'_{OR,\lambda} P_{\tilde{W}_{OR,\lambda}} \tilde{\bar{Y}}_{OR,\lambda}$ is the efficient, asymptotically unbiased estimator for $\beta^*$.

### A.6.4 Estimating Measurement Error Variance and FGLS Efficiency: $K_Y = 1$

Feasibly implementing the GLS adjustments proposed in the previous two subsections requires estimating the weighting parameters $\lambda$. Clearly, the weights for $w_{i,OR}$ can be readily estimated from sample data. As such, we focus on estimating $w^*$, an exercise that turns to

estimating $\Sigma_B^\infty$. In the heteroskedastic setting,

$$\text{Cov}\left[\hat{\beta}_i, \hat{\beta}_j\right] = \begin{cases} \frac{X^{i\prime} P_{W^i} P_{W^j} X^j}{\left(X^{i\prime} P_{W^i} X^i\right)\left(X^{j\prime} P_{\tilde{W}^j} X^{j)}\right)} s_{10}, & \text{if } i \neq j \\ \frac{1}{X^{i\prime} P_{W^i} X^i} s_{1i}, & \text{if } i = j \end{cases}.$$

Feasible implementation then requires estimating $s_{10}$ and $s_{1i}$. Recalling their definition from equation 9, a feasible estimator for $s_{10}$ would average over the cross-products between residuals corresponding to two different elicitations of $X^*$ for the same individual. The feasible estimator for $s_{1i}$ then just sums those squared residuals corresponding to the $i^{th}$ model's specification.

$$\hat{s}_{10} = \frac{1}{N}\sum_{n=1}^{N} \frac{2}{(K_X - 1)(K_X - 2)} \sum_{i=1}^{K_X - 1} \sum_{j=i+1}^{K_X} \hat{\varepsilon}_n^i \hat{\varepsilon}_n^j \text{ , and, } \hat{s}_{1i} = \frac{1}{N}\sum_{n=1}^{N} \hat{\varepsilon}_n^{i2}.$$

Applications of the Law of Large Numbers then ensure that $\hat{s}_{1i} \to_p s_{1i}, i = 0, \ldots, K_X$. Substituting these values into the formula for $\Sigma_B^\infty$ defines a consistent estimator $\hat{\Sigma}_B \to_p \Sigma_B^\infty$, which can be used to consistently estimate the efficient weights, $\hat{w}^* = \frac{\hat{\Sigma}_B^{-1} 1_{K_X}}{1_{K_X}' \hat{\Sigma}_B^{-1} 1_{K_X}}$.

### A.6.5 Estimating Measurement Error Variance and FGLS Efficiency: $K_Y \geq 2$

Implementing feasible efficient estimators with multiple measurements of $Y*$, each of which has different variances, requires estimating the variance-covariance matrix for the different measurements. This setting is straightforward, as the objective of inference can be directly estimated. Let $\hat{\Sigma}_Y \to \Sigma_Y$ denote the sample and population covariance matrix for the different measurements of $Y$. Recalling that $S_Y$ is the diagonal matrix of measurement error variances for $Y^*$, it is immediately apparent that:

$$\Sigma_Y = \sigma_{\varepsilon^*}^2 1_{K_Y} 1_{K_Y}' + S_Y.$$

Consequently, a consistent estimate of $\sigma^2_{\nu^k_Y}$ could estimate $\sigma^2_{\varepsilon^*}$ by averaging the off-diagonal entries of $\Sigma_Y$, and subtracting this average from the $k^{th}$ diagonal entry.

### A.6.6 A Note of Caution on FGLS Implementations

When implementing FGLS for ORIV, it is best to exploit the structure of the model to its fullest extent. Simply applying FGLS to the full ORIV specification will perform very poorly. Since the population covariance matrix for residuals is singular, in finite samples, its inverse is very poorly scaled. Even attempting to simply combine elicitations for $Y$ efficiently can result in extreme weights without imposing homogeneity on the off-diagonal entries in $\Sigma_Y$. While we do not present detailed results to this effect, our simulation analysis suggests that FGLS estimators perform poorly relative to the simple ORIV estimator. Unless there is good reason to model substantial heterogeneity in measurement error, our informal recommendation is to avoid FGLS corrections.

# B Measurement Error in Binary Choices

As mentioned in the text, Niederle and Vesterlund (2007) attempt to control for preferences with another competition choice. Namely, in the last stage of the experiment (their Task 4), subjects are given a second opportunity to be paid for their performance in the piece-rate task (Task 1). They may choose whether to be paid as a piece rate (if Task 4 was randomly chosen for payment by the experimenters), or to enter their Task 1 performance into a tournament with the other three participants in the group. This choice has the same payoff features as the competitive choice given in their main task (Task 3). The idea is that the choice in Task 4 would control for risk aversion, overconfidence, and feedback aversion, so that different choices in Task 3 and Task 4 would be explained by a difference in preferences for competition per se. However, in the presence of measurement error, this control is subject to

the issues highlighted in Section 3. Here, we explore the effect of measurement error in binary controls theoretically. We then show that an analysis of NV's data that properly accounts for measurement error leads to the same conclusion we arrive at, that gender differences in competition are driven by gender differences in fundamental attitudes such as risk aversion and overconfidence.

Although this section is motivated by the approach in NV, it is developed in generality as it applies to any case where a binary measure, measured with error, is used as a control. The next subsection carries out this general development, and the following subsection gives a numerical example that is closely related to NV, as well as a re-analysis of NV's data.

## B.1 Stochastic Choice with Binary Preference Control

In the general formulation, let individual $i$ have latent characteristics $X_i$ (risk and overconfidence), treatment $D_i$ (gender), and face two binary decision tasks in which they report $Y_i^a \in \{0,1\}$ (competition) as well as its replicate $Y_i^b \in \{0,1\}$ (entering piece-rate performance in a competition). Each individual answers 1 to both the decision task and the replicate with probability $p_i(X_i)$ that depends on $X_i$ but is independent of treatment, $D_i$, so that

$$\mathbb{E}[Y_i^a|X_i, D_i] = \mathbb{E}[Y_i^a|X_i] = p_i(X_i) = \mathbb{E}\big[Y_i^b|X_i\big] = \mathbb{E}\big[Y_i^b|X_i, D_i\big]\,.$$

Though $Y_i^a|X_i \perp D_i \perp Y_i^b|X_i$, the unconditional statement is not generally true because of potential dependence between $D_i$ and $X_i$. Consequently, $p_i$ is correlated with $D_i$ only because both are correlated with $X_i$.

By a standard application of the Frisch-Waugh-Lovell Theorem, we can estimate the effect of $D$ on $Y^a$ conditional on a set of controls in two stages. That is, we can get rid of the dependence on $X_i$, through $p_i(X_i)$, to directly understand how measurement error would produce a biased estimate of the effect of $D$ on $Y^a$, even controlling for $Y^b$. In the first stage,

we regress $Y^a$ and $D$ on the controls and recover the residuals from both regressions:

$$\begin{aligned} Y^a &= \pi_{Y^a,0} + \pi_{Y^a,1} p_i(X_i) + u_{Y^a} \\ D &= \pi_{D,0} + \pi_{D,1} p_i(X_i) + u_D. \end{aligned}$$

In the second stage, we regress the residual variation in outcomes, $u_{Y^a}$, on the residual variation in treatment, $u_D$.

$$u_{Y^a} = \beta u_D + \varepsilon.$$

Without measurement error, the estimate of $\beta$ would be zero.

However, when $Y_i^b$ as a proxy for $p_i(X_i)$ this introduces measurement error. In this case, the first-stage estimates for $\pi_{Y^a,\cdot}$ and $\pi_{D,\cdot}$ will be biased towards zero by measurement error. Consequently, in the second stage regression of $u_{Y^a}$ on $u_D$, both residuals will be tainted by persistent variation in the controls due to measurement error. Further, the contamination in both residuals will be correlated. This correlated contamination is what drives distorted inference in the second stage regression.

Replacing each individual's choice probability $p_i$ with the replicate $Y_i^b$ provides just this form of measurement error. To show this formally, denote this error as $\nu_i = Y_i^b - p_i$, which takes the value $-p_i$ with probability $1 - p_i$ and the value $1 - p_i$ with the complementary probability $p_i$. Then we have

$$\begin{aligned} Y^a &= \tilde{\pi}_{Y^a,0} + \tilde{\pi}_{Y^a,1} Y^b + \tilde{u}_{Y^a} = \tilde{\pi}_{Y^a,0} + \tilde{\pi}_{Y^a,1}(p + \nu) + \tilde{u}_{Y^a} \\ D &= \tilde{\pi}_{D,0} + \tilde{\pi}_{D,1} Y^b + \tilde{u}_D = \tilde{\pi}_{D,0} + \tilde{\pi}_{D,1}(p + \nu) + \tilde{u}_D. \end{aligned}$$

The residuals for this regression clearly differ from those in the regression without measurement error. Let $\gamma_{Y^a} = \frac{\text{Var}[Y^a]}{\text{Var}[Y^a]+\text{Var}[\nu]}$ and $\gamma_D = \frac{\text{Var}[D]}{\text{Var}[D]+\text{Var}[\nu]}$, so that $\tilde{\pi}_{\cdot,1} = \gamma_{\cdot}\pi_{\cdot,1}$. Note that

$\pi_{1,\cdot} - \tilde{\pi}_{1,\cdot} = (1 - \gamma_\cdot)\pi_{1,\cdot}$ and

$$\begin{aligned} \pi_{0,\cdot} &= \mathbb{E}[Y^a] - \pi_{1,\cdot}\mathbb{E}[p] \\ \tilde{\pi}_{0,\cdot} &= \mathbb{E}[Y^a] - \tilde{\pi}_{1,\cdot}\mathbb{E}\big[Y^b\big] \end{aligned} \quad \Rightarrow \quad \begin{aligned} \tilde{\pi}_{0,\cdot} &= \pi_{0,\cdot} + (\pi_{1,\cdot} - \tilde{\pi}_{1,\cdot})\mathbb{E}[p] \\ &= \pi_{0,\cdot} + (1 - \gamma_\cdot)\pi_{1,\cdot}\mathbb{E}[p]\,. \end{aligned}$$

We can now relate the residuals $\tilde{u}_\cdot$ to their uncontaminated counterpart $u_\cdot$:

$$\begin{aligned} u_\cdot - \tilde{u}_\cdot &= \tilde{\pi}_{0,\cdot} - \pi_{0,\cdot} + (\tilde{\pi}_{1,\cdot} - \pi_{1,\cdot})p + \tilde{\pi}_{1,\cdot}\nu \\ &= (1 - \gamma_\cdot)\pi_{1,\cdot}\{\mathbb{E}[p] + p\} + \gamma_\cdot\pi_{1,\cdot}\nu. \end{aligned}$$

Regressing the contaminated residual in outcomes on the contaminated residual in treatment then yields a spurious correlation, as

$$\begin{aligned} \mathrm{Cov}[\tilde{u}_{Y^a}, \tilde{u}_D] &= \mathrm{Cov}[u_{Y^a} + (1 - \gamma_{Y^a})\pi_{1,Y^a}p + \gamma_{Y^a}\pi_{1,Y^a}\nu, u_D + (1 - \gamma_D)\pi_{1,D}p + \gamma_D\pi_{1,D}\nu] \\ &= \mathrm{Cov}[u_{Y^a}, u_D] + (1 - \gamma_{Y^a})\pi_{1,Y^a}(1 - \gamma_D)\pi_{1,D}\mathrm{Var}[p] + \gamma_{Y^a}\pi_{1,Y^a}\gamma_D\pi_{1,D}\mathrm{Var}[\nu], \end{aligned}$$

where constant expectations are dropped for compactness. We also see that:

$$\mathrm{Var}[\tilde{u}_D] = \mathrm{Var}[u_D] + (1 - \gamma_D)^2\pi_{1,D}^2\mathrm{Var}[p] + \gamma_D^2\pi_{1,D}^2\mathrm{Var}[\nu].$$

Consequently, even though $\mathrm{Cov}[u_{Y^a}, u_D] = 0$, when we test the second stage regression:

$$\tilde{u}_{Y^a} = \tilde{\beta}\tilde{u}_D,$$

we are likely to get biased results indicating a significant treatment effect in the contaminated data because:

$$\mathbb{E}\Big[\tilde{\beta}\Big] = \frac{(1 - \gamma_Y^a)\pi_{1,Y^a}(1 - \gamma_D)\pi_{1,D}\mathrm{Var}[p] + \gamma_{Y^a}\pi_{1,Y^a}\gamma_D\pi_{1,D}\mathrm{Var}[\nu]}{\mathrm{Var}[u_D] + (1 - \gamma_D)^2\pi_{1,D}^2\mathrm{Var}[p] + \gamma_D^2\pi_{1,D}^2\mathrm{Var}[\nu]} \neq 0.$$

## B.2 Numerical Example

Note that choosing to compete in NV is tantamount to choosing a lottery that pays a fixed amount with some probability—the probability of winning the tournament—and zero otherwise, over a sure thing—the piece-rate payment.[1] We use this observation to present a numerical example.

In the setup of the competition task, the choice will be driven by the interaction of subjective probabilities of winning (overconfidence) with risk aversion. To simplify, we consider only variations in risk aversion for a fixed lottery and certainty equivalent. Participants are given the choice between a lottery that pays \$100 with a 25% probability and receiving a \$20 payment with certainty. Each participant has CRRA utility with risk aversion parameter $\theta_i$. Conditional on risk aversion, participants' choices are governed by logistic choice probabilities:

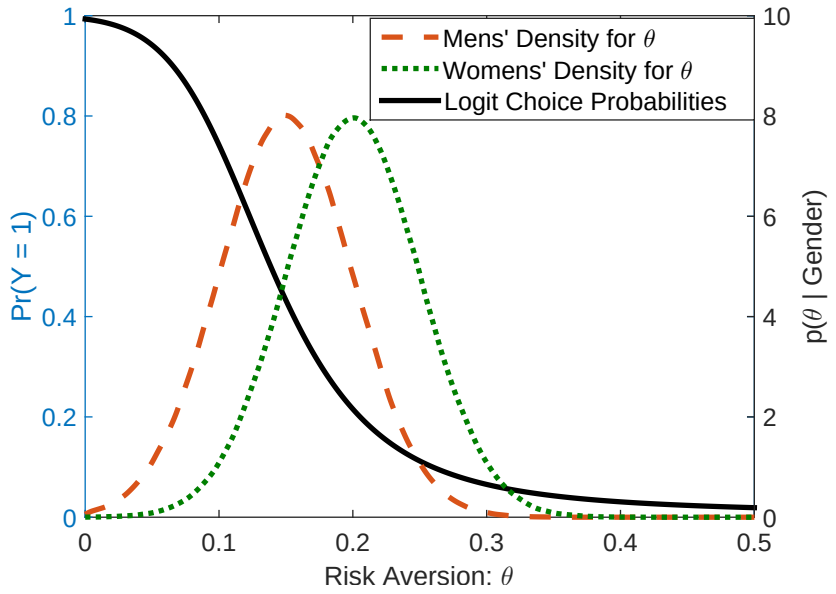
$$p_i = Pr\{\text{Choose Lottery}|\theta_i\} = \frac{\exp\{\frac{0.25}{1-\theta_i}100^{1-\theta_i}\}}{\exp\{\frac{0.25}{1-\theta_i}100^{1-\theta_i}\} + \exp\{\frac{1}{1-\theta_i}20^{1-\theta_i}\}}$$

To synthetically calibrate preferences, subjects who have $\theta_i = \frac{\log(4/5)}{\log(1/5)} \approx 0.14$ are indifferent between the lottery and the certainty equivalent, choosing each with equal probability. We assume the risk-aversion parameter is distributed normally for men and women, with each distribution having standard deviation 0.05. To calibrate choices to the 19% gap in competition we observe in our data, we assume the distribution for men has mean 0.15, and for women, 0.2. This results in men and women who will choose the lottery about 46% and 27% of the time, respectively. Figure 4 illustrates the probability of choosing the lottery conditional on risk aversion and the distributions over risk aversion for men and women.

Let $Y_i^a = 1\{\text{Player } i \text{ Chooses Lottery in Main Decision}\}$ represent the observed choice for subject $i$ in the main decision task. If we knew $\theta_i$ and could compute each individual's $p_i$,

[1] Indeed, participants were informed of the number of correct sums they solved during each task.

Figure 4: Example Illustrating Lottery Choice Probabilities and Gender-Differences in Risk Aversion




then $p_i \equiv \mathbb{E}[Y_i^a|\theta_i]$ would be the best predictor for $Y_i^a$ and the ideal explanatory variable and control. As such, following NV, we describe $p_i$ as an individual's *preferences*. A regression of $Y_i^a$ on gender and preferences $p_i$ would fail to find any relationship between gender and lottery take-up (competition) in a large sample.

If we only observe a replicate $Y_i^b = 1\{\text{Player } i \text{ Chooses Lottery in Replicate Task}\}$, that replicate will have classical, mean-zero measurement error for the proper control $p_i$. Letting $\eta_i = Y_i^b - p_i$, the variance of the realization of this measurement error, $p_i(1-p_i)$ correlates with gender because the distribution of $p_i$ depends on gender.

Numerically, if we simulate the choices $Y_i^a$, $Y_i^b$ for 800 individuals (calibrated to our sample size) and regress the choice of the lottery in $Y_i^a$ on gender 10,000 times, we observe an average coefficient of 0.16 (or 16%, s.e. 0.034, $p < 0.01$). This can be understood to occur because although the coefficient on $Y^b$ should be unity, it is, instead 0.19 (s.e. 0.036, $p < 0.01$) due to measurement error.

This suggests that by forcing the coefficient on $Y_i^b$ to be 1, the resulting regression can produce a valid test. That is, by regressing $Y_i^a - Y_i^b$ on gender, the resulting coefficient will be a conservative test of the effect of gender controlling for the second choice.[2] Following this prescription in our simulated data gives a very accurate and precise estimate of $-0.000078$ (s.e. 0.034, $p = 1.00$).

The same exercise can be conducted using the data from NV. For calibration, using a linear probability model in their main specification produces a coefficient on male of 0.27 (s.e. 0.10, $p < 0.01$). Adding their measure of $Y^b$ to the right hand side reduces this coefficient to 0.21 (s.e. 0.10, $p < 0.05$). Using a proper specification, with $Y^b$ on the left-hand side, this results in a coefficient on male of 0.075, with an average standard error of 0.12 ($p = 0.54$). Including the proxies for $X$ in their specifications (appearing in their Tables VI and VIII), the coefficient drops to 0.053 (s.e. 0.12, $p = 0.67$). Adding an unused, but reasonable, additional proxy for $X$ results in a coefficient of 0.087 (s.e. 0.11, $p = 0.43$).[3] Thus, had NV used a specification that accounts for measurement error in this control, they would have found the same thing we have: that preferences for competition based on risk aversion and overconfidence explain the gender gap in competition.

# C Example STATA

## C.1 Using Principal Components as Controls

```
tab sumsRank, gen(ss)
tab sumsCorrectCompete, gen(scc)
tab performanceDiff, gen(pdd)
```

[2]The test is conservative because although the left-hand-side will now properly correspond to $Y - \mathbb{E}[Y|X]$, the right-hand-side will consist of $D$ rather than $D - \mathbb{E}[D|X]$. This test is asymptotically efficient, but it will be inefficient in small samples.

[3]We thank Muriel Niederle and Lise Vesterlund for generously sharing their data.

```
#delimit;
pca ss* scc* pdd* *RiskyProjectAllocation riskyUrn*OMaxValue compoundUrn*
*Over* CRTCorrect CRTPercentile ravenCorrect ravenPercentile guess*Confidence gpaPercent
#delimit cr

predict p1-p76

//stardardize principal components to make output easily interpretable
foreach x of varlist p1-p5 {
sum `x'
replace `x' = (`x'-r(mean))/r(sd)
}

. reg sumsCompete male p1-p5

      Source |       SS       df       MS              Number of obs =     783
-------------+------------------------------           F(  6,   776) =   34.50
       Model |  36.3549507     6  6.05915845           Prob > F      =  0.0000
    Residual |  136.291282   776  .175633095           R-squared     =  0.2106
-------------+------------------------------           Adj R-squared =  0.2045
       Total |  172.646232   782  .220775233           Root MSE      =  .41909

------------------------------------------------------------------------------
 sumsCompete |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0411864   .0331224     1.24   0.214    -.0238336    .1062065
          p1 |   .1497314   .0158151     9.47   0.000      .118686    .1807768
          p2 |   .0193967   .0149865     1.29   0.196    -.0100222    .0488155
          p3 |  -.1382706    .015186    -9.11   0.000    -.1680811   -.1084601
          p4 |  -.0092983   .0149876    -0.62   0.535    -.0387194    .0201228
          p5 |   .0336481   .0152317     2.21   0.027     .0037477    .0635484
       _cons |   .3035024   .0248917    12.19   0.000     .2546393    .3523655
------------------------------------------------------------------------------
```

## C.2 ORIV When $Y$ is Measured without Error

This subsection and the next implement ORIV assuming that the $X$ and $Y$ variables are on the same scale. If they are not, one should first put them on the same scale. If there is an obvious way to do this, as in the case of certainty equivalents of lotteries with the same

probabilities, but a different high option, this should be done. Otherwise standardization of the variables may be attractive.

```
use tmp, clear
keep uid highLowValue firstProjectValue secondProjectValue
rename firstProjectValue mainVar
rename secondProjectValue instrument
gen control1 = 1
save tmp1, replace
use tmp, clear
keep uid highLowValue firstProjectValue secondProjectValue
rename firstProjectValue instrument
rename secondProjectValue mainVar
gen control2 = 1
append using tmp1
replace control1 = 0 if control1 == .
replace control2 = 0 if control2 == .
. ivregress 2sls highLowValue (mainVar = instrument) control*, cluster(uid) nocons

Instrumental variables (2SLS) regression              Number of obs =    1766
                                                      Wald chi2(3)  =       .
                                                      Prob > chi2   =       .
                                                      R-squared     =       .
                                                      Root MSE      =  20.673

                                  (Std. Err. adjusted for 883 clusters in uid)
------------------------------------------------------------------------------
             |               Robust
highLowValue |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mainVar |   1.238774   .1477303     8.39   0.000      .949228     1.52832
    control2 |  -28.40269   6.270678    -4.53   0.000    -40.69299   -16.11238
    control1 |  -28.56091   6.278213    -4.55   0.000    -40.86598   -16.25584
------------------------------------------------------------------------------
Instrumented:  mainVar
Instruments:   control2 control1 instrument
```

## C.3 ORIV Estimates of a Correlation

```
use tmp, clear
keep uid firstRiskyProjectAllocation riskyUrn20RN riskyUrn30RN
rename firstRiskyProjectAllocation LHS
rename riskyUrn20RN mainVar
rename riskyUrn30RN instrument
gen control1 = 1
save tmp1, replace
use tmp, clear
keep uid firstRiskyProjectAllocation riskyUrn20RN riskyUrn30RN
rename firstRiskyProjectAllocation LHS
rename riskyUrn20RN instrument
rename riskyUrn30RN mainVar
gen control2 = 1
append using tmp1
save tmp1, replace
use tmp, clear
keep uid secondRiskyProjectAllocation riskyUrn20RN riskyUrn30RN
rename secondRiskyProjectAllocation LHS
rename riskyUrn20RN mainVar
rename riskyUrn30RN instrument
gen control3 = 1
append using tmp1
save tmp1, replace
use tmp, clear
keep uid secondRiskyProjectAllocation riskyUrn20RN riskyUrn30RN
rename secondRiskyProjectAllocation LHS
rename riskyUrn20RN instrument
rename riskyUrn30RN mainVar
gen control4 = 1
append using tmp1
replace control1 = 0 if control1 == .
replace control2 = 0 if control2 == .
replace control3 = 0 if control3 == .
replace control4 = 0 if control4 == .
. ivregress 2sls LHS (mainVar = instrument) control*, cluster(uid) nocons

Instrumental variables (2SLS) regression              Number of obs =    3532
                                                      Wald chi2(5)  =       .
                                                      Prob > chi2   =       .
```

```
                                                      R-squared     =        .
                                                      Root MSE      =   56.934

                                  (Std. Err. adjusted for 883 clusters in uid)
------------------------------------------------------------------------------
             |               Robust
         LHS |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mainVar |    2.80734   .4042871     6.94   0.000     2.014952    3.599728
    control4 |   14.25088   19.37017     0.74   0.462    -23.71396    52.21572
    control3 |   23.55569   18.06123     1.30   0.192    -11.84367    58.95505
    control2 |  -3.854446    19.4315    -0.20   0.843    -41.93948    34.23059
    control1 |   5.450364   18.12677     0.30   0.764    -30.07744    40.97817
------------------------------------------------------------------------------
Instrumented:  mainVar
Instruments:   control4 control3 control2 control1 instrument

use tmp, clear
. corr firstRiskyProjectAllocation secondRiskyProjectAllocation, cov
(obs=883)

             | first~on secon~on
-------------+------------------
firstRisk~on |  3923.25
secondRis~on |  2186.45  2724.73

. corr riskyUrn20RN riskyUrn30RN, cov
(obs=883)

             | ris~20RN ris~30RN
-------------+------------------
riskyUrn20RN |  59.2741
riskyUrn30RN |   30.895  44.3019

//corrected correlation
. display 2.80734*sqrt(30.895)/sqrt(2186.45)
.04805784

//corrected standard error of correlration
. display .4042871*sqrt(30.895)/sqrt(2186.45)
.33371009
```

# D Comparison of Subject Pools

We now compare responses to standard choice tasks on the Caltech Cohort Study (CCS) to other subject pools. The participant's responses on the CCS resemble those previously reported. That is, the participants in the CCS are not dramatically different from those in other subject pools. This is in line with the replication results reported in the paper: before correcting for measurement error, our data yield virtually identical conclusions to those reported in the original experiments.

## D.1 The Beauty Contest

Each installment of the CCS contained a beauty contest game following Nagel (1995). Each participant is asked to choose a number between 0 and 100. They are told that they will be grouped with 9 randomly-chosen participants, and if their choice is closest to 2/3 of the average of the choices in that group, they will receive a $5 reward, otherwise nothing.[4]

This task is often viewed as a proxy for cognitive sophistication. Any choice above 66.66 is dominated. If everyone uses undominated strategies, any choice above 44.44 will not pay off. In this way, each iteration can be associated with a greater level of sophistication. In the limit, this iterated elimination of dominated strategies yields the unique equilibrium of the game—everyone should choose 0. Nagel (1995) reported numbers very far from 0 when the game was played for the first time by Bonn University students. In her data, the mean number chosen was 36.73 and the median was 33. Data from the first installment of the CCS in the Fall of 2013 yields similar results: In our sample of 806 participants, the mean was 37.17 with a median of 33.[5]

[4] If they tie with a subset of $k$ participants in their group, each gets $1/k$ of the $5 reward.

[5] While Nagel (1995) does not report her raw data, the distribution of selected numbers in our data visually resembles the distribution depicted in her Figure 1B.

## D.2 The Cognitive Reflection Test

Next we compare responses on the Cognitive Reflection Test (CRT), described briefly in Section 2.1. The CRT was introduced by Frederick (2005) and consists of three quantitative questions, each having a seemingly intuitive answer that is wrong. In the Spring 2015 installment of the CCS, we included the three questions proposed by Frederick (2005), but with different wording and scaling.

The number of correct answers on the CRT is viewed as a proxy for meta-cognition. It is associated with time preferences: participants who score higher are more "patient". In the CCS, the fraction of participants with 0, 1, 2, and 3 correct answers are 19%, 25%, 28%, and 27%, respectively. Frederick (2005) reports responses from 11 subject populations with aggregate proportions of 33%, 28%, 23%, and 17%. In the CCS, participants were less likely to answer no question correctly, and more likely to answer all three questions correctly, while the rates for 1 or 2 correct answers are similar.

## D.3 Dictator Giving

All installments of the CCS included the dictator game, in which participants have to divide a fixed allocation between themselves and another randomly chosen participant. We have noted that generosity declines markedly the second time a student participates in the CCS, so we focus on the initial, Fall 2013, installment. The mean amount given away is 22%, with 43% giving away nothing, 13% giving away a third, and 27% giving away half.

Engel (2011) compiles data from 83 papers, consisting of 616 treatments and 20,813 individual observations. The mean amount given away across all treatments was 28%. At the individual level, he reports that 36% give away nothing, 9% give away a third, and 17% give away half. These numbers are very close to those we observe. Participants in the CCS are slightly more extreme in the sense that they are more likely to give away nothing, but

when they do give, they are more generous.

### D.4 Risky Projects

We conclude in Section 4 that a particularly appealing risk elicitation is the "Project" measure, described in Section 2.1, and based on Gneezy and Potters (1997). Here we focus on a particular implementation of this elicitation in the Fall 2014 installment of the CCS. This implementation allowed participants to allocate 200 tokens between a safe option and a risky project, the latter returning 2.5 tokens with 50% probability per token allocated.[6] In our data, the mean allocation to the risky project is 74%, with a median of 75%.

In comparison, Agranov and Yariv (2015) elicit precisely the same measure in 8 sessions with 80 students at the University of California, Irvine (UCI). There, the mean allocation was 72% with a median of 70%. The distribution of allocations in their data looks very similar to that in the CCS. For example, in the CCS approximately 30% of participants allocated half or less of their endowment to the risky project, while at UCI approximately 35% of participants did.

## E Question Wordings

Screenshots and other design details of the Caltech Cohort Study can be found at 2. In this section we present the question wordings that were used in this paper. Throughout this section, comments in square brackets are meant to express information that is not found on the screen, but is useful in understanding the flow of the survey. A new item number generally indicates another screen (even though there may not be a particular question associated with it).

[6] We focus on this version of the task since it is the most commonly used and therefore allows for a natural comparison with existing data.

## E.1 Competition

This question, meant to elicit competitiveness, follows NV, and was used on the Spring 2015 CCS. It consists of several parts, several of which create data which are used both by LV and by us as controls.

1. This next task asks you to add together series of numbers. You will be given three minutes to complete as many sums as possible. When all surveys are submitted, we will randomly group you with 3 other people (so you will be in a group of 4). **You will be paid only if you achieve the highest number of completed sums within this group, in which case you will be paid 40 tokens per sum completed.**

   In case of a tie between those who completed the highest number of sums, we will randomly determine the participant who will be paid.

2. [3 minutes in which to do as many sums as the participant can]

3. In the randomly determined group of 4 (you and 3 others), where do you think you rank in terms of the number of sums completed (1 corresponding to the highest number of sums completed in the group, 4 corresponding to the lowest number of sums completed in the group). You will earn an additional 50 tokens if your guess is correct.

   [radio buttons next to numbers 1 through 4]

4. You will be given an additional three minutes to complete as many sums as possible. Please pick how you would like to be paid from the following two options:

   (a) 10 tokens per sum completed; or

   (b) When all surveys are submitted, we will randomly group you with 3 other people (so you will be in a group of 4). We will compare the number of sums you complete now with the number of sums the other 3 completed in the previous stage you

just concluded. **You will be paid only if you achieve the highest number of completed sums within this group, in which case you will be paid 40 tokens per sum completed.**

In case you tie with another participant(s), we will randomly determine whom of these gets first place, and you will be paid only if it is you who is declared the winner.

(c) [3 minutes in which to do as many sums as the participant can]

## E.2 Overconfidence

These questions were used as controls in the regressions on competitiveness, and elicited on the Spring 2015 CCS.

The first three questions are used as a control for overprecision, and are based on Ortoleva and Snowberg (2015).

1. We will now measure your ability to assess numbers quickly.

   We will show you three pictures of jars of jellybeans. Please give us your best guess as to the number of jellybeans in each jar.

2. [With random picture of a jar of jellybeans] Please enter the number of jellybeans you think are in this jar (between 1 and 1000).

3. How confident are you of your answer to this question?

   (a) No confidence at all

   (b) Not very confident

   (c) Somewhat unconfident

(d) Somewhat confident

(e) Very confident

(f) Certain

[Repeat three times.]

The next set of questions ask the participant to complete a set of tasks, and then asks the participant to guess their performance (as a measure of estimation / overestimation), and then their guess as to how their performance compares with their peers (placement / overplacement). Participants are given 30 seconds to answer each logical question.

4. This next task asks you to solve five logical puzzles. You will be given thirty seconds to complete each puzzle, and will be paid 20 tokens for each puzzle solved correctly.

5. [5 Raven's matrices from Condon and Revelle (2014). All participants executed the same matrices in the same order.]

6. How many of the 5 puzzles do you think you solved correctly?

7. Out of 100 other randomly picked survey participants, what percentage do you think solved more puzzles correctly than you?

The next set follows the same structure, but uses questions from the cognitive reflection test (CRT).

8. This next task asks you to answer five logical questions. You will have up to 30 second to answer each question and will be paid 20 tokens for each question answered correctly.

9. A monitor and a keyboard cost 350$intotal.Themonitorcosts$300 more than the keyboard. How much does the keyboard cost?

10. It takes 10 computers 10 minutes to run 10 simulations. How long does it take 200 computers to run 200 simulations?

11. In the pond in front of Baxter Hall, there is a patch of lily pads. The patch doubles in size every day. If it takes 36 days for the patch to cover the entire pond, how many days would it take to cover half the pond?

12. Professor Wiseman spent one-fourth of his life as a boy, one-eighth as a youth, and one-half as an active man. If Professor Wiseman spent 8 years as an old and wise man, how many years did he spend as an active man?

13. A 4 foot pole casts a shadow that is 2 feet long on the ground. If the pole was 16 feet in height, how long would the shadow be?

14. How many of the 5 questions do you think you answered correctly?

15. Out of 100 other randomly picked survey participants, what percentage do you think answered more questions correctly than you?

## E.3 Risk Elicitations

The Fall, 2014 CCS contained a large number of risk elicitations. The Spring, 2015 CCS included variants of the risky projects, the risky urns, and the qualitative risk question, which were used as controls for the competition task.

### E.3.1 Risky Projects

These risky projects are based on Gneezy and Potters (1997). Note that they were separated by several questions on the survey.

1. You are endowed with 200 tokens (or $2) that you can choose to keep or invest in a risky project. Tokens that are not invested in the risky project are yours to keep.

   The risky project has a 40% chance of success.

   If the project is successful, you will receive 3 times the amount you chose to invest.

   If the project is unsuccessful, you will lose the amount invested.

   Please choose how many tokens you want to invest in the risky project. Note that you can pick any number between 0 and 100, including 0 or 100:

2. You can invest in another risky project if you would like. You can invest up to 200 tokens, or you can choose to keep them.

   The risky project has a 50% chance of success.

   If the project is successful, you will receive 2.5 times the amount you chose to invest.

   If the project is unsuccessful, you will lose the amount invested.

   Please choose how many tokens you want to invest in the risky project. Note that you can pick any number between 0 and 200, including 0 or 200:

### E.3.2 Risky Urns

These are standard urn gambles, with certainty equivalents elicited using a multiple price list. Note that the order of the lotteries was randomized, as was the order (spaced throughout the survey) with the compound and ambiguous urn lotteries.

4. In the next task we will ask you to assess the value of several gambles. The gambles will be designed using virtual urns filled with red and black balls. We will give you some information on the composition of each urn.

Well let you pick which color ball you would like to pay off for each gamble. If you choose a gamble on an urn, we will draw one ball from it. If that ball is the color you chose, you will receive a reward. If it is the other color, you will receive nothing.

The mechanism for selecting the gamble may take some getting used to. We will provide you with a list of rewards from 0 tokens to 150 tokens in increments of 5 tokens, and for each one, we ask that you think whether you prefer that amount, or taking the urn gamble. Once you click on an option, well fill in the rest of the choices for you so that they make sense (this will save you time!).

However, you should keep clicking on options that you prefer until the choice in each line is what you would like. You should do so because at the end of the survey, we will randomly select one row from the list, and you will get whatever you chose on that line. If you specified you prefer a sure amount on that line, youll get that amount. If you specified that you preferred the gamble, then we will draw a ball from the urn, and pay you according to the description above.

5. The next choice will involve an urn containing 20 balls, 10 of which are red and 10 of which are black.

   Which color ball would you like to paid 100 tokens for (if it is drawn from the urn in the following questions)? Note that this means you will be paid 0 tokens if the other color is drawn.

   (a) red

   (b) black

6. **Urn with Equal Number of Red and Black Balls**

   The urn from which we can draw a ball is composed of 10 red balls and 10 black balls.

   The urn gamble pays 100 tokens if the ball drawn is [red].

What do you prefer? (make sure a radio button in each row is selected)

[Multiple price list with "Urn Gamble" on the left hand side, and "X tokens" on the right hand side, with X in increments of 10 from 0 to 100.]

7. The next choice will involve an urn containing 30 balls, 15 of which are red and 15 of which are black.

   Which color ball would you like to paid 150 tokens for (if it is drawn from the urn in the following questions)? Note that this means you will be paid 0 tokens if the other color is drawn.

   (a) red

   (b) black

8. **Urn with Equal Number of Red and Black Balls**

   The urn from which we can draw a ball is composed of 15 red balls and 15 black balls.

   The urn gamble pays 150 tokens if the ball drawn is [red].

   What do you prefer? (make sure a radio button in each row is selected)

   [Multiple price list with "Urn Gamble" on the left hand side, and "X tokens" on the right hand side, with X in increments of 10 from 0 to 150.]

### E.3.3 Lottery Menu

This question is based on Eckel and Grossman (2002).

9. Which of the following gambles would you prefer?

   Each of the gambles will give you a 50% chance of the Low Payoff, and a 50% chance of the High Payoff.

The gamble you chose will be run at the end of the survey, and we will tell you your payoff then.

| | Low Payoff | High Payoff |
|---|---|---|
| **Gamble 1:** | 140 | 140 |
| **Gamble 2:** | 120 | 180 |
| **Gamble 3:** | 100 | 220 |
| **Gamble 4:** | 80 | 260 |
| **Gamble 5:** | 60 | 300 |
| **Gamble 6:** | 10 | 350 |

#### E.3.4 Qualitative Risk Assessment

This qualitative assessment of risk comes from Dohmen et al. (2011). This was also elicited on the Spring 2015 CCS, and that question was used to instrument the question on the Fall 2014 CCS.

10. How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?

    Please tick a box on the scale, where the value 0 means: not at all willing to take risks and the value 10 means: very willing to take risks

    [radio buttons, presented horizontally, with numbers from 0 to 10 next to each option.]

### E.4 Compound and Ambiguous Lotteries

Compound and ambiguous lotteries follow the same format as the risky urn lotteries above. These three blocks of urn gambles were spaced out throughout the survey, and their order

was randomized. The compound lotteries were also used as a control (risk-aversion measure) in the regressions regarding competitiveness.

1. The next choice will involve an urn containing 20 red and black balls. The composition of the urn is randomly determined. That is, we will first randomly draw a number between 0 and 20 (all equally likely). That number will be the number of red balls. The remaining balls will be black.

   Which color ball would you like to paid 100 tokens for (if it is drawn from the urn in the following questions)? Note that this means you will be paid 0 tokens if the other color is drawn.

   (a) red

   (b) black

2. **Urn with Uncertain Number of Red and Black Balls**

   The composition of the urn is randomly determined. That is, we will first randomly determine a number between 0 and 20 (all equally likely). That number will be the number of red balls, the rest of the 20 balls (if any) will be black.

   The urn gamble pays 100 tokens if the ball drawn is [red].

   What do you prefer? (make sure a radio button in each row is selected)

   [Multiple price list with "Urn Gamble" on the left hand side, and "X tokens" on the right hand side, with X in increments of 10 from 0 to 100.]

3. The next choice will involve an urn containing 30 red and black balls. The composition of the urn is randomly determined. That is, we will first randomly draw a number between 0 and 30 (all equally likely). That number will be the number of red balls. The remaining balls will be black.

Which color ball would you like to paid 150 tokens for (if it is drawn from the urn in the following questions)? Note that this means you will be paid 0 tokens if the other color is drawn.

(a) red

(b) black

4. **Urn with Uncertain Number of Red and Black Balls**

   The composition of the urn is randomly determined. That is, we will first randomly determine a number between 0 and 30 (all equally likely). That number will be the number of red balls, the rest of the 30 balls (if any) will be black.

   The urn gamble pays 150 tokens if the ball drawn is [red].

   What do you prefer? (make sure a radio button in each row is selected)

   [Multiple price list with "Urn Gamble" on the left hand side, and "X tokens" on the right hand side, with X in increments of 10 from 0 to 150.]

5. The next choice will involve an urn containing ,0 balls, each of which could be red or black. Dean Dabiri has chosen the exact composition of the urn: the balls could all be red, they could all be black, or there could be any combination of red and black balls.

   Which color ball would you like to paid 100 tokens for (if it is drawn from the urn in the following questions)? Note that this means you will be paid 0 tokens if the other color is drawn.

   (a) red

   (b) black

6. **Urn with Unknown Number of Red and Black Balls**

The urn has a combination of 20 red and black balls chosen by Dean Dabiri.

The urn gamble pays 100 tokens if the ball drawn is [red].

What do you prefer? (make sure a radio button in each row is selected)

[Multiple price list with "Urn Gamble" on the left hand side, and "X tokens" on the right hand side, with X in increments of 10 from 0 to 100.]

7. The next choice will involve an urn containing 30 balls, each of which could be red or black. Dean Dabiri has chosen the exact composition of the urn: the balls could all be red, they could all be black, or there could be any combination of red and black balls.

   Which color ball would you like to paid 150 tokens for (if it is drawn from the urn in the following questions)? Note that this means you will be paid 0 tokens if the other color is drawn.

   (a) red

   (b) black

8. **Urn with Unknown Number of Red and Black Balls**

   The urn has a combination of 20 red and black balls chosen by Dean Dabiri.

   The urn gamble pays 150 tokens if the ball drawn is [red].

   What do you prefer? (make sure a radio button in each row is selected)

   [Multiple price list with "Urn Gamble" on the left hand side, and "X tokens" on the right hand side, with X in increments of 10 from 0 to 150.]