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

Ben Gillen

Claremont McKenna College

Erik Snowberg

California Institute of Technology, University of British Columbia, and National Bureau of Economic Research

Leeat Yariv

Princeton University, Centre for Economic Policy Research, and National Bureau of Economic Research

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 statistical techniques for handling 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 replicate three classic experiments, demonstrating that results change substantially when measurement error is accounted for. Collectively, these results show that failing to properly account for measurement error may cause a field-wide bias leading scholars to identify "new" phenomena.

Snowberg gratefully acknowledges the support of NSF grants SES-1156154 and SMA-1329195. Yariv gratefully acknowledges the support of NSF grants SES-0963583 and SES-1629613 and Gordon and Betty Moore Foundation grant 1158. We thank Jonathan Bendor, Christopher Blattman, Colin Camerer, Marco Castillo, Gary Charness, Lucas Coffman, Guillaume Frechette, Dan Friedman, Drew Fudenberg, Yoram Halevy, Ori Heffetz, Muriel Niederle, Alex Rees-Jones, Shyam Sunder, Roel van Veldhuizen, and Lise Vesterlund, as well

Electronically published June 13, 2019
[*Journal of Political Economy*, 2019, vol. 127, no. 4]

I. 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. Despite the ubiquity of measurement error, fewer than 10 percent of experimental papers published in the last decade in leading economics journals mention measurement error as a concern (see Sec. I.B for details). Moreover, the tools for dealing with measurement error in experiments—most commonly, improved elicitation techniques and multiple rounds—are relatively crude. This paper proposes a mix of statistical tools and design recommendations to handle measurement error in experimental and survey research.

At the heart of our approach is the combination of duplicate elicitations (usually two) of behavioral proxies and methods from the econometrics literature, particularly the instrumental-variables approach to errors in variables (Reiersøl 1941). 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 discussed here 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 (CCS), described in Section II. We replicate within the CCS three classic and influential studies, and observe that 30–40 percent of variance in choices is attributable to measurement error. In all three of the studies that we examine, accounting for measurement error substantially alters conclusions and implications.

First, we consider the most influential experimental study of the last decade, Niederle and Vesterlund (2007). That paper found that men are more likely to select into competition because of a preference for competitive situations that is distinct from risk attitudes and overconfidence. We replicate, as have many others, the fact that men choose to compete more often than women. However, the gender gap in competition is well explained by risk attitudes and overconfidence once measurement error is properly accounted for. This is true in both our data and that of the original study. Second, Friedman et al. (2014), summarizing their own re-

as two anonymous reviewers and the editor, Emir Kamenica, for comments and suggestions. We also appreciate the input of seminar audiences at Caltech, Hong Kong University of Science and Technology, the Ifo Institute, Nanyang Technological University, the National University of Singapore, Stanford Institute for Theoretical Economics, the University of Bonn, the University of British Columbia, the University of Southern California, and the University of Zurich. Data are provided as supplementary material online.

search and that of many other scholars, find low correlations between different lab-based methods of measuring risk attitudes. As risk attitudes are fundamental to many economic theories, the failure to reliably measure them has troubling implications for lab experiments. In contrast, we find that many commonly used measures of risk attitudes are highly correlated once measurement error is taken into account. Third, we inspect the relationship between attitudes toward ambiguous and compound lotteries, following the setup of Halevy (2007). Ambiguity aversion is a rich area of theoretical exploration, and it is used to explain many behaviors: from equity trading to voting decisions. Despite the fact that Halevy finds a substantial correlation between attitudes toward compound risk and ambiguity, his results are often seen as consistent with these attitudes corresponding to separate phenomena (see Epstein [2010] and Ahn et al. [2014], among others). We find that once measurement error is accounted for, there is very little difference between the two attitudes.

As is well known, classical measurement error in a single variable biases estimates of effects toward zero. This attenuation bias is considered conservative, as it "goes against finding anything"; that is, it reduces the probability of false positives. However, as our results demonstrate, it may also lead to the identification of "new" effects and phenomena that are, in actuality, already reflected in existing research.

A. *Simulated Examples*

Here we present simulated examples to illustrate, for the unfamiliar reader, the problems created by measurement error, and summarize our approaches. In our first example, 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 we use to simulate data is

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

where $\eta \sim \mathcal{N}[0, 0.9]$ (so the variance of D is ≈ 1), $X^* \sim \mathcal{N}[0, 1]$, 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^*, where $\varepsilon \sim \mathcal{N}[0, 1]$.

A diligent researcher would fit a regression model of the form

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

hoping to control for the role of risk attitudes in the effect of gambling on participation in dangerous sports. Table 1 shows, from simulations,

TABLE 1
SIMULATED REGRESSIONS OF EQUATION (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%
	A. $N = 100$					
$\hat{\alpha}$	.00	.06	.11	.16	.21*	.26***
	(.11)	(.11)	(.12)	(.12)	(.12)	(.12)
$\hat{\beta}$	1.00***	.87***	.75***	.64***	.54***	.44***
	(.12)	(.11)	(.11)	(.10)	(.10)	(.09)
Percent of time $\alpha = 0$ rejected at the 5% level:						
1 noisy measure of X^*	5	8	15	25	37	50
5 noisy measures of X^*	5	6	6	7	9	11
10 noisy measures of X^*	5	5	5	5	6	7
20 noisy measures of X^*	5	5	5	5	5	6
	B. $N = 1{,}000$					
$\hat{\alpha}$	.00	.06*	.11***	.16***	.21***	.26***
	(.03)	(.04)	(.04)	(.04)	(.04)	(.04)
$\hat{\beta}$	1.00***	.87***	.75***	.64***	.54***	.43***
	(.04)	(.04)	(.03)	(.03)	(.03)	(.03)
Percent of time $\alpha = 0$ rejected at the 5% level:						
1 noisy measure of X^*	5	31	81	98	100	100
5 noisy measures of X^*	5	6	11	23	42	66
10 noisy measures of X^*	5	5	7	10	16	28
20 noisy measures of X^*	5	5	5	6	8	11

NOTE.—Coefficients and standard errors (in parentheses) are averages from 10,000 simulated regressions.
* Statistical significance at the 10 percent level.
** Statistical significance at the 5 percent level.
*** Statistical significance at the 1 percent level.

how the estimates, $\hat{\alpha}$ and $\hat{\beta}$, depend on how much measurement error there is in the variance of X, that is, $\sigma_\nu^2/(\sigma_\nu^2 + \sigma_{X^*}^2)$.

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 $\hat{\alpha}$ becomes statistically significant in the average simulation when measurement error reaches approximately one-third of the variance of X. Statistically significant results are likely in such a setting, as we estimate that measurement error accounts for 30–40 percent of the variance of elicited proxies for risk attitudes; see equation (4) and surrounding text. Intuitively, false positives occur because the measurement error in X attenuates $\hat{\beta}$, allowing $\hat{\alpha}$ to pick up the variation in D related to X^*.

Depressingly, adding more observations does nothing to reduce the bias in the estimated coefficients. In fact, when $N = 1{,}000$, the approximate size of the CCS, $\hat{\alpha}$ appears statistically significant in the average sim-

TABLE 2
SIMULATED CORRELATIONS WHEN X AND Y ARE MEASURED WITH ERROR
($N = 100$; True Model: $\text{Corr}[X^*, Y^*] = 1$)

	ERROR AS A PERCENT OF $\text{Var}[X]$ AND $\text{Var}[Y]$					
	0% (1)	10% (2)	20% (3)	30% (4)	40% (5)	50% (6)
$\widehat{\text{Corr}}[X, Y]$	1.00	.90***	.80***	.70***	.60***	.50***
	(.00)	(.02)	(.04)	(.05)	(.06)	(.08)
$\widehat{\text{Corr}}[\mathbb{E}[X], \mathbb{E}[Y]]$	1.00	.95***	.89***	.82***	.75***	.66***
	(.00)	(.01)	(.02)	(.03)	(.04)	(.06)
ORIV $\widehat{\text{Corr}}[X, Y]$	1.00	1.00	1.00	1.00	1.00	1.00
	(.00)	(.01)	(.02)	(.04)	(.06)	(.10)

NOTE.—Coefficients and standard errors (in parentheses) are averages from 10,000 simulated regressions.
* Statistical significance at the 10 percent level.
** Statistical significance at the 5 percent level.
*** Statistical significance at the 1 percent level.

ulation when measurement error accounts for only 10 percent of the variance of X. This emphasizes that issues with measurement error will not "wash out" once a study is large enough.

Correcting this is simple: elicit more controls. But questions remain: how many? and, how should these controls be used? If experimenters are certain that they can elicit the exact quantity they wish to control for (with measurement error), only two controls are necessary. One control should be entered linearly, and the other used as an instrument for the control. This will generate the proper coefficient on the control and, thus, on the variable of interest D. However, it is doubtful that this will ever be the case in practice. Even for a simple control such as risk aversion, there are multiple, imperfectly correlated elicitation methods, as we discuss in Section IV. Including multiple controls for multiple measures of multiple behaviors may not be practical.

Thus, in Section III, we explore several different ways of including controls. First, we include them linearly, as suggested by table 1. Then, we show that principal-component analysis allows for a small but informative set of controls, preserving degrees of freedom. Finally, we elicit each control twice and use the duplicate observation as an instrument. These different approaches all lead to the same conclusion: the gender gap in competitiveness can be explained by risk attitudes and overconfidence, although Niederle and Vesterlund (2007) concluded that it was a distinct phenomenon.

The problem of measurement error biasing coefficients is particularly acute when researchers estimate correlations between X and Y, as shown by the simulated results in table 2. In this table, we vary the proportion of

measurement error in both variables. Even a bit of measurement error causes significant deviations from the true correlation of 1. As measurement error accounts for 30–40 percent of the variation in our elicitations, it is extremely unlikely one would ever estimate a correlation close to 1, even if that were the true value. As in table 1, increasing N does not affect point estimates, but shrinks standard errors.[1]

To correct for measurement error, we expand on traditional instrumental-variable approaches to errors in variables. Our approach, which we call *obviously related instrumental variables* (ORIV), uses duplicate elicitations of X and Y as instruments, and produces an estimator that is more efficient than standard instrumental-variable techniques. Specifically, we obtain duplicate measures of X, denoted X^a and X^b, which are both proxies for X^* that are measured with error. If measurement error in the two elicitations is orthogonal—as we assume—then the predicted values $\hat{X}^a(X^b)$ from a regression of X^a on 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.[2] ORIV 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.[3]

ORIV produces consistent coefficients, correlations, and standard errors. This is in contrast to one common way experimenters deal with multiple noisy elicitations: averaging. As can be seen from table 2, while averaging reduces bias, it still leads to incorrect conclusions in the presence of small amounts of measurement error.

We apply ORIV, in Section IV, to show that various risk elicitation methods are more correlated than previously thought. We further use this technique to show, in Section V, 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 VI, that failing to correct for measurement error has led the field to overidentify "new" phenomena.

[1] Note that the standard errors are smaller in table 2, as $\text{Var}[\varepsilon]$, set to 1 in all columns of table 1, now varies across the columns: starting at 0 in col. 1 and climbing up to 1 in col. 6.

[2] If measurement error is positively correlated across elicitations, then instrumented coefficients will still be biased downward, although less so than without instrumenting. In our experimental design, we tried to weaken any possible correlation by varying the choice parameters, the grid of possible responses, and so on. See Sec. II and Sec. IV.D.2 for details and discussion.

[3] ORIV is equivalent to using all valid moment conditions in the generalized method of moments (GMM); however it is simpler and more transparent. See app. A.6 for a detailed comparison of ORIV and GMM.

B. Related Literature

Mismeasurement of data has been an important concern for statisticians and econometricians since the late nineneenth century (Adcock 1878). Indeed, estimating the relationship between two variables when both are measured with error is a foundational problem in the statistics literature (Koopmans 1939; Wald 1940). The use of instrumental variables to address the classical errors-in-variables problem was proposed by Reiersøl (1941; see Hausman [2001] for a review). This was first applied in economics by Friedman (1957) to estimate consumption functions. Since then, instrumental variables have been used to account for measurement error in an assortment of fields including medicine, psychology, and epidemiology.

The experimental literature has considered noise in lab data and its consequences, going back to at least Kahneman (1965). Nonetheless, in the decade from 2006 to 2015 only 9 percent of the 283 experimental (field and lab) papers in the top five economics journals explicitly tried to deal with measurement error.[4] One-fifth of these papers either used an experimental design aimed at reducing noise or averaged multiple elicitations, a technique that may do little to reduce bias, as shown in Section I.A. About one-half of these papers estimate structural models of participants' "mistakes." In particular, two-fifths estimate *quantal response equilibrium* models.[5] Most of the remaining papers use indirect methods to deal with noise, such as the elimination of outliers, or the informal derivation of additional hypotheses about the effects of noise, which are then tested.[6] Our paper is, to our knowledge, the first to offer simple, yet general, experimental techniques for mitigating the effects of measurement error.

Many of the issues in experimental work due to measurement error are present in survey research as well. For example, Bertrand and Mullainathan (2001) and Bound, Brown, and Mathiowetz (2001) highlight the potential perils of measurement error in survey research in econom-

[4] We examined full, refereed papers published in *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*.

[5] These posit a structural model in which participant mistakes are inversely related to the payoff losses they generate; see Goeree, Holt, and Palfrey (2016) for details and a review.

[6] Some particularly innovative papers of this type are worth mentioning. Battalio et al. (1973) show that even small reporting errors can lead to a rejection of the generalized axiom of revealed preferences. Castillo, Jordan, and Petrie (2015) posit a structural model of measurement error in children's risk elicitations. Coffman and Niehaus (2015) adjust for measurement error in self-interest and other-regard by projecting both on a common set of explanatory variables. Ambuehl and Li (2015) use instrumental variables to control for measurement error in one of their analyses. Like our paper, all of these consider only classical measurement error. Standard texts on the statistics of measurement error, such as Buonaccorsi (2010), may be of use in future research.

ics and social psychology. Bound et al. (2001) also note the potential usefulness of instrumental-variable approaches. The closest paper to ours in that literature is Beauchamp, Cesarini, and Johannesson (2015), which considers measurement error in survey-based risk elicitations. It uses a latent-variable model that allows inferences about the component of measured risk attitudes that is not due to measurement error. It emphasizes, as we do, the ubiquity of measurement error, and the paucity of concern about it.[7]

II. Caltech Cohort Study

We administered an incentivized, online survey to the entire undergraduate student body of Caltech in the fall of 2013 and 2014 and the spring of 2015. The survey included incentivized tasks designed to elicit an array of behavioral attributes. It also included a set of questions addressing students' lifestyle and social habits.[8]

The data used in this paper are from the fall 2014 and spring 2015 installments. In the fall of 2014, 92 percent of the student body (893/972) responded to the survey. The average payment was $24.34. In the spring of 2015, 91 percent of the student body (819/899) responded to the survey. 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.[9] Of those who had taken the survey in 2015, 96 percent (786/819) also took the survey in 2014. As Section IV requires data from both surveys, for consistency we use this subsample of 786 throughout.

There are several advantages to using the CCS to address questions of measurement error. The large size of the study allows us to document the nonexistence of certain previously identified "distinct" behaviors with unusual precision. Furthermore, the inflation of standard errors that comes with using instrumental-variable techniques does not threaten the validity of our inferences. Last, unlike most experimental settings, there is little concern about self-selection into our experiments from the participant population, due to our more than 90 percent response rates (see Snowberg and Yariv [2018] and references therein). Thus, the issues we identify are due solely to measurement error, and not due to a small sample or self-selection.

[7] In a different context, Aguiar and Kashaev (2017) suggest a nonparametric statistical notion of rationalizability of a random vector of prices and consumption streams when there is measurement error in consumption levels reported in surveys. They use this test to assess standard exponential time discounting.

[8] For screenshots of the 2015 survey, go to leeatyariv.com/ScreenshotsSpring2015.pdf.

[9] The number of overall students was substantially lower 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.

Nonetheless, Caltech is highly selective, which may cause concern that the overall population is different from the pool used in most lab experiments. Three points should mitigate this concern. First, the raw results of the replications are virtually identical to those reported in the original papers. Second, responses from our survey to several standard elicitations—of risk, altruism in the dictator game, and so on—are similar to those reported in several other pools (see app. D for details [apps. A–E available as an online supplement]). Third, while top-10 schools account for 0.32 percent of the college age population in the United States, top-50 schools enroll only 3.77 percent of that population (using the *U.S. News & World Report* rankings). Thus, there seems to be little cause for concern that our participant 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 conclusions are likely due to our more sophisticated treatment of measurement error, rather than an artifact of the participant population.

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 $1.00.

A. *Overconfidence*

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

Overestimation and overplacement.—Participants complete two tasks: a five-question cognitive reflection test (CRT; see Frederick 2005) and five Raven matrices (Raven 1936). After each block of questions, each participant is asked how many they think they answered correctly. This, minus the participant's true performance, gives two measures of overestimation—one for each of the two tasks. Each participant is also asked where they think they are in the performance distribution of all participants. This, minus the participant's true percentile, gives a measure of overplacement for each task.

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.

B. Risk

Risk measures are used in Section III as controls and in Section IV as an outcome of interest. Further, the risk multiple-price list described below is used as an outcome of interest in Section V.

Projects.—Following Gneezy and Potters (1997), participants are asked to allocate 100 or 200 tokens between a safe option and a project that returns some multiple of the tokens with probability p, otherwise nothing. In fall 2014, two projects were used: the first returning 3 tokens per token invested with $p = .4$, and the second returning 2.5 tokens with $p = .5$. In the spring of 2015, the first project was modified to return 3 tokens with $p = .35$.

Qualitative.—Following Dohmen et al. (2011), participants are asked to rate themselves, on a scale of 0–10, in terms of their willingness to take risks. We use the spring 2015 elicitation as a duplicate measure of the fall 2014 elicitation.

Lottery menu.—Following Eckel and Grossman (2002), participants are asked to choose between six 50/50 lotteries with different stakes.[10] The first lottery contains the same payoff in each state and thus corresponds to a sure amount. The remaining lotteries contain increasing means and variances, allowing for an estimation of risk aversion.

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 pays off if a ball of the color of the participant's choosing is drawn. The first urn contains 20 balls—10 black and 10 red—and pays 100 tokens. The second contains 30 balls—15 black and 15 red—and pays 150 tokens.[11]

C. Ambiguous and Compound Lotteries

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

Compound MPL.—This follows the same protocol as the risk MPLs described above, except participants are told that the number of red balls is 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 IV.

Ambiguous MPL.—This elicitation emulates the standard Ellsberg (1961) urn. It follows the same protocol as the two other MPLs. Partici-

[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.

pants are informed that the composition of the urn was chosen by the dean of undergraduate students at Caltech.

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

III. Misspecified Controls and Measurement Error

To make the claim that an estimated effect is independent of other factors, many studies attempt to control for those other factors. If they are measured with error, one control, or even a few, may be insufficient to reliably assert the claim, as illustrated in Section I.A. Here we show that properly dealing with controls measured with error has important substantive consequences. We do so by replicating the competitiveness-and-gender study of Niederle and Vesterlund (2007) within the CCS. Like Niederle and Vesterlund, we find a robust difference in the rates at which men and women compete. However, Niederle and Vesterlund (2007, 1070) conclude as follows: "Including these controls [for overconfidence, risk, and feedback aversion], gender differences are still significant and large. Hence, we conclude that, in addition to gender differences in overconfidence, a sizable 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.

Using the notation of Section I.A, measurement error in X, in this case controls for risk aversion and overconfidence, can result in a biased estimate of the coefficient on D, in this case 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^*.[12] For illustration, consider an extreme case in which the variance of ν is very large, so that X is almost entirely noise. Ignoring that noise in X, standard regression analysis could lead to the erroneous conclusion that Y and D are correlated, even when controlling for X.[13]

To put this in terms of our substantive example, it is well known that overconfidence is correlated with gender (see, e.g., Moore and Healy 2008), and, depending on the elicitation method, risk aversion may be correlated with gender as well (see Charness, Gneezy, and Imas [2013]

[12] Note that in many of our applications, these variables will be binary or ordered multinomials. In those cases, we still model the corresponding latent variable (X^* or Y^*) as continuous and the responses (X or Y) as stochastic. Measurement error then affects the probability that a given discrete choice is made.

[13] It is well known that measurement error in left-side variables may bias estimated coefficients in discrete-choice models; see Hausman (2001). We use linear probability models as Y may also be measured with error.

and Holt and Laury [2014] for surveys, as well as our discussion in Sec. IV.F). Thus, if competitiveness is driven by overconfidence or risk aversion, mismeasurement of these traits will lead to an overestimate of the effect of gender.

What can be done to mitigate this issue? There are several approaches. The first is to include multiple measures for each of the possible controls *X*. This approach reduces the effects of measurement error, while helping to ensure that elicited controls cover the potentially different aspects of the behavioral attribute being controlled for. However, this may have two shortcomings. First, it may come at the cost of too many degrees of freedom. Second, without a large number of controls for each aspect, this approach will not entirely eliminate the effects of measurement error. We therefore consider two further approaches: including principal components of the multiple controls, and instrumenting each control with a duplicate. This final approach—spanning the space of different aspects of the behavioral attribute, while instrumenting each of these aspects—is preferable whenever feasible.

A. Measuring Competitiveness in the Caltech Cohort Study

Part of the spring 2015 survey mimicked the essential elements of Niederle and Vesterlund's design. First, participants had 3 minutes to complete as many sums of five two-digit numbers as they could. 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 were broken randomly. As in Niederle and Vesterlund, 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.

Next, in parallel to the central task of Niederle and Vesterlund's design, participants were told they would have an additional 3 minutes to complete sums. But, 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 correctly solved sum. The tournament had a similar payment scheme to the first 3-minute task. However, a participant's performance in the tournament would be compared to the performance of three randomly chosen participants in the *first* task.[14] Otherwise, the payment structure was identical to that in the first task.

There are a few ways in which our implementation differs from Niederle and Vesterlund's:

[14] 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.

Time and payments.—We gave participants 3 rather than 5 minutes to complete sums. Per-sum payments were scaled down by a factor of four. As with the rest of the CCS, participants were paid for all tasks, rather than a randomly selected one. This could have caused participants to hedge by choosing the piece-rate scheme.

Grouping of participants.—In Niederle and Vesterlund, participants were assigned to groups of four in which they could visibly see that there were two men and two women. This created an imbalance in the expected number of female competitors: two-thirds of the group for men, one-third for women. In the CCS, we randomly selected groups for each of the two tasks separately after the survey was administered. Thus, both genders faced the same expected profile of competitors.

Experimental setting.—Our tasks are on a survey, whereas Niederle and Vesterlund use a lab setting. We ran our survey several months later in a lab environment with 98 Caltech students. We saw very similar results, but with larger standard errors due to the smaller sample.

Additional tasks.—Niederle and Vesterlund 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 either an additional piece-rate or tournament payment scheme for their performance in the preliminary task. This final choice served as a control for risk aversion and overconfidence. As the CCS has multiple other controls for both of these traits, we omitted these two parts to reduce complexity and length.

The first three of these factors would only change interpretation if they affected men and women differently. However, as we replicate the gender gap in tournament entry found by Niederle and Vesterlund, these do not seem to be of particular importance. The final factor implies that we cannot use an analogous control to that of Niederle and Vesterlund's with the CCS data. Nonetheless, using Niederle and Vesterlund's data and an analysis accounting for measurement error in their control, we find that the gender gap in tournament entry can be explained by risk aversion and overconfidence (see Sec. III.C). More details about our, and Niederle and Vesterlund's, implementation can be found in appendix E.1.

B. 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.[15]

[15] As noted in n. 13, discrete-choice models may produce biased estimates of coefficients when the left-side variable is measured with error. Nonetheless, in our data, probit and logit specifications produce almost identical levels of statistical significance as in table 3.

TABLE 3
GENDER, COMPETITION, AND CONTROLS

	Chose to Compete ($N = 783$)							
Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male	.19***	.13***	.11***	.11***	.048	.050	.041	.0063
	(.034)	(.030)	(.031)	(.031)	(.033)	(.034)	(.033)	(.054)
Guessed tournament rank		−.15***	$F = 29$	$F = 28$	$F = 23$	$F = 21$		$\chi^2_3 = 8.8$
		(.017)	$p = .00$	$p = .00$	$p = .00$	$p = .00$		$p = .04$
Tournament performance		.086***	$F = 1.6$	$F = 1.6$	$F = 1.6$	$F = 1.5$		$\chi^2_{30} = 36$
		(.020)	$p = .02$	$p = .03$	$p = .03$	$p = .05$		$p = .21$
Performance difference		−.021	$F = 1.4$	$F = 1.5$	$F = 1.4$	$F = 1.4$		$\chi^2_{25} = 27$
		(.017)	$p = .09$	$p = .07$	$p = .11$	$p = .11$		$p = .34$
Risk aversion: MPL #1				.042***				
				(.015)				
Overplacement: CRT				.026*				
				(.015)				
Risk aversion: project #2					.067***			
					(.016)			
Perceived performance (percentile): CRT					−.042***			
					(.016)			
All risk aversion controls						$F = 4.9$		
						$p = .00$		
All overconfidence controls						$F = 1.8$		
						$p = .05$		
First five principal components							$F = 37$	
							$p = .00$	
Instrumental variables								$\chi^2_7 = 24$
								$p = .00$
Adjusted R^2	.038	.23	.26	.27	.28	.29	.22	n.a.

NOTE.—Coefficients and standard errors (in parentheses) on all nondichotomous 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, 5 variables for risk aversion controls, and 12 variables for overconfidence controls.

* Statistical significance at the 10 percent level.

** Statistical significance at the 5 percent level.

*** Statistical significance at the 1 percent level.

Column 1 shows the baseline difference in competition: Women choose tournament incentives 21.4 percent of the time, while men choose them 40.4 percent of the time, for a difference of 19.0 percentage points. This difference is highly statistically significant. While these numbers are somewhat lower than those reported in Niederle and Vesterlund, their relative sizes are quite similar. Column 2 controls for participants' performance and estimates of their own rank linearly, as in Niederle and Vesterlund's main specification. Similarly to their results, the inclusion of these controls reduces the coefficient on gender by approximately one-third.

There is, however, a nonlinear relationship between expected rank and perceived probability of victory in a competition.[16] Therefore, column 3 enters participants' subjective ranks nonparametrically, by including a dummy variable for each possible response (three categories). This estimation confirms that the effect of perceived rank in a competition is, indeed, nonlinear, although the coefficient on gender remains unchanged.[17] Column 3 also enters performance nonlinearly (29 and 26 categories, respectively), as there is also a nonlinear relationship between performance and competition. The coefficient on gender in column 3 is lower than in column 2.[18]

Column 4 introduces controls for risk aversion and overconfidence. One control for each attribute, selected for illustrative purposes, is entered. This does not affect the gender coefficient, despite both controls having statistically significant coefficients. Two different controls are entered in column 5. Doing so 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 a trait is fully controlled for. Moreover, they suggest that measurement error in the controls themselves allows for the perception of competitiveness as a separate trait.

[16] 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

$$\sum_{i=0}^{3}(i+1)\binom{3}{i}(1-p)^i p^{3-i} = 3(1-p)+1.$$

Thus, with a reported rank of r (ignoring rounding), the probability of winning the competition is $[(4-r)/3]^3$.

[17] Rates of competition are 65.6 percent for participants who predicted they would come in first (in a random group of 4), and 31.4, 15.3, and 5.0 percent for participants predicting they would come in second, third, and fourth (last), respectively. The distribution of guessed ranks differs from that reported in Niederle and Vesterlund: our participants were better calibrated, and this likely resulted in the lower observed rates of tournament entry.

[18] This is entirely driven by including performance in the first task nonparametrically, as there are small differences in male and female performance in this task, as shown in fig. E.2 of app. E.1.

We note that these conclusions are not driven by our unusually large sample size. If anything, the size of our data set helps reduce standard errors and identify weak effects. To see this, we draw a random sample of 40 women and 40 men (the size and gender composition of Niederle and Vesterlund's experiment) from our data 10,000 times, and regress the competition decision on two overconfidence controls, two risk controls, and perceived rank in the first competition task. The coefficient on gender is significant at the 1 percent level 2.2 percent of the time, at the 5 percent level 7.6 percent of the time, and at the 10 percent level 13 percent of the time.

Column 6 enters all available controls for risk (six controls) and overconfidence (an additional 12 controls). The coefficient on gender is relatively unchanged. If we enter these controls separately, we find that much of the decrease in this coefficient, compared with column 3, is due to risk controls. We revisit the relationship between gender and risk aversion in Section IV.F.

The number of controls in column 6 (76, including categorical controls for performance) approaches the number of data points in a normally sized study—such as Niederle and Vesterlund, which had 80 participants. Thus, we examine ways to preserve degrees of freedom. The simplest is to perform a principal-components analysis of all 76 of the controls. In this case, entering just the first five principal components produces a very similar point estimate to entering all 76 controls. More on this technique can be found in appendix B.1.

As discussed in Section I.A, the potential bias in the gender coefficient comes from the fact that the coefficients on the noisy controls—assumed to be positively correlated with both gender and competitive behavior—are biased toward zero. Thus, in column 8, we instrument the risk aversion and overconfidence controls for which we have multiple elicitations. This approach combines consistent estimates of the coefficients on controls, while still ensuring we span the space of possible aspects of risk aversion and overconfidence in our data. While the point estimate of the gender coefficient is consistent, it is accompanied by higher standard errors that come with an instrumental-variables specification.

When dealing with measurement error, a key advantage of the experimental approach is the ability to design one's own controls. Using enough controls to span the space of aspects of a behavioral attribute, with a duplicate elicitation of each as an instrument, is preferable, if feasible. However, as mentioned, limits on participants' time and attention may impose a constraint on the number of controls that can be elicited. When this constraint binds, an experimenter should think carefully about the trade-off between measuring an additional aspect and controlling for it perfectly via instrumentation. We return to multiple elicitations and instrumental-variable strategies in Sections IV and V. Before doing so, we examine how to correct for measurement error when using specially designed controls.

C. Using Designed Controls

Niederle and Vesterlund control for risk aversion and overconfidence with another tournament entry choice. Namely, in the last stage of their experiment, participants are given a second opportunity to be paid for their performance in the piece-rate task from the beginning of the experiment. They can choose to be paid again as a piece rate, or to enter their performance into a tournament. The clever idea behind this additional choice is that it controls for all aspects determining tournament entry not directly related to a preference for competing—explicitly, risk aversion and overconfidence. Using their data, we show that accounting for measurement error in this control generates different conclusions from those of Niederle and Vesterlund. Henceforth, we refer to the choice in the main task as Y_i^a, and that in the final task as Y_i^b.

Niederle and Vesterlund regress Y_i^a on gender, performance, guessed tournament rank, and Y_i^b, as in column 3 of table 4. This reduces the coefficient on gender compared to their main specifications, which are displayed in columns 1 and 2.[19] However, suppose Y_i^b is positively correlated with both gender and competition, but contaminated with classical measurement error. This will bias the coefficient on Y_i^b downward and the coefficient on gender upward.

As Y_i^b is designed to measure every part of the tournament entry choice except a preference for competition, it should enter the regression with a coefficient of 1. This can be implemented by regressing $Y_i^a - Y_i^b$ on gender, which produces an unbiased estimate of the effect of gender on tournament entry, controlling for Y_i^b. Intuitively, the only difference between these two variables is, by construction, a desire to compete. To see whether this desire is correlated with gender, it should be regressed on gender. Doing so results in an insignificant coefficient on gender of .075. The inclusion of additional controls on the right side reduces the coefficient even further.[20]

D. Substantive Interpretation

Our analysis shows, using both new data and data from Niederle and Vesterlund, that although men are more likely to select into competition, this is not due to a distinct *preference* for performing in a competitive en-

[19] We thank Muriel Niederle and Lise Vesterlund for generously sharing their data. The coefficients in table 4 differ from Niederle and Vesterlund because of our use of ordinary least squares (OLS), which is preferred to probit for reasons described in n. 13. However, *p*-values are very similar.

[20] The inclusion of additional controls should make the test more efficient in small samples. A formal exposition of this point, and other details of this subsection, can be found in app. B.2

TABLE 4
REANALYSIS OF NIEDERLE AND VESTERLUND'S DATA

	CHOOSE TO COMPETE (Y_i^a)			$Y_i^a - Y_i^b$	
DEPENDENT VARIABLE	(1)	(2)	(3)	(4)	(5)
Male	.37***	.27***	.21**	.075	.053
	(.11)	(.11)	(.10)	(.12)	(.12)
Tournament performance	.016	−.003	−.012***		−.037
	(.019)	(.019)	(.018)		(.023)
Performance difference	.016	−.005	.012		.056**
	(.023)	(.023)	(.023)		(.027)
Guessed tournament rank		−.24***	−.20***		−.11
		(.066)	(.066)		(.080)
Y_i^b			.27**		
			(.11)		
Adjusted R^2	.13	.24	.29	.00	.054

NOTE.—Coefficients and standard errors (in parentheses) on all nondichotomous measures are standardized. $N = 80$ for all regressions.
* Statistical significance at the 10 percent level.
** Statistical significance at the 5 percent level.
*** Statistical significance at the 1 percent level.

vironment. Rather, it is driven by differences in risk aversion and overconfidence. 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 either of these factors is in explaining competition, only that together they explain much of the effect.[21]

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 shown to predict several interesting behaviors, such as choice of college major (see, e.g., Buser, Niederle, and Oosterbeek 2014). However, risk aversion and overconfidence feature in many theories, and are therefore of potential use in bringing theory to bear on gender differences.

There are also practical considerations. Niederle and Vesterlund 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 3 minutes (rather than 5), 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.

[21] Van Veldhuizen (2016) uses a clever experimental design to add refined versions of Niederle and Vesterlund's final task, and specifications following the previous subsection, to tease these effects apart.

IV. Measurement Error on Both Sides

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 independent, variables is a much more serious problem, biasing estimated coefficients 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 classical measurement error in either biases estimates toward zero. We introduce a simple method, *obviously related instrumental variables* (ORIV)—which is more efficient than standard instrumental-variable techniques—to overcome these issues. We apply this technique to the estimation of the correlation between different measures of risk attitudes in this section, and between risk and ambiguity aversion in the next. The discussion in this section focuses on implementation, with the formal properties of the estimators developed in appendix A.

A. *Risk Elicitation Techniques*

There is a substantial experimental literature assessing the validity of common experimental techniques for eliciting attitudes toward 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. This type of work has generally found small correlations between different techniques, making it difficult to study the individual correlates of risk preferences. The literature concludes that risk elicitation is a "risky business." The pun is not ours; see Friedman et al. (2014) for a survey.

However, none of the studies on which this conclusion is based account for measurement error. In what follows, we inspect several commonly used risk elicitation techniques, and estimate their within-participant correlations using an instrumental-variables strategy to account for measurement error. This generates much higher within-participant correlations than previously reported. Moreover, the corrected correlations 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 correlations with other measures, and more stability over time. Further, elicitations based on allocation decisions display substantial gender effects—which are consistent with investment behavior in the field—while certainty equivalent elicitations do not.

This section uses four measures of risk as described in Section II.B: qualitative, risk MPL, project, and lottery menu. Before proceeding, we note how we transform variables for comparison. First, when using two measures from the same form of elicitation, we put them 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).[22] Second, when comparing objects such as estimated constant relative risk aversion (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 give risk-loving answers in the risk MPLs are recoded to give a risk-neutral answer. Without this censoring, results are qualitatively similar.[23]

B. *First Take on Measurement Error Correction*

It is well known that measurement error attenuates estimated coefficients (see, e.g., Buonaccorsi 2010). Here we review that basic finding to set up a framework for our estimator.

To estimate the relationship between two variables measured with independent and identically distributed error, $Y = Y^* + \nu_Y$ and $X = X^* + \nu_X$ (with $\mathbb{E}[\nu_Y \nu_X] = 0$ and $\mathrm{Var}[\nu_k] = \sigma^2_{\nu_k}$), the ideal regression model would be $Y^* = \alpha^* + \beta^* X^* + \varepsilon^*$. Instead, we can only estimate $Y = \alpha + \beta X + \varepsilon$, where α is a constant and ε is a mean-zero random noise. Annotating finite-sample estimates with hats and population moments without hats, this results in an estimated relationship of

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

The estimated coefficient $\hat{\beta}$ is thus biased toward 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 ν^a_X, ν^b_X independent and identically distributed random variables, and $\mathbb{E}[\nu^a_X \nu^b_X] = 0$. With the additional assumption that

[22] This implicitly assumes a CRRA utility function.

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

$$\frac{\text{Var}[\nu_X^a]}{\text{Var}[X^a]} = \frac{\text{Var}[\nu_X^b]}{\text{Var}[X^b]} \equiv \frac{\text{Var}[\nu_X]}{\text{Var}[X]},$$

we have

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

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 .6, suggesting that the variance of measurement error is on the order of two-thirds of the variance of X^*.[24]

Using instrumental variables, the second noisy measure of X^* can be used to recover a consistent estimate of the true coefficient β^*. As a corollary of (4), note that $\widehat{\text{Cov}}[X^a, X^b] \approx \widehat{\text{Var}}[X^*]$ consistently estimates $\text{Var}[X^*]$ in the population. We apply two-stage least squares (2SLS) to instrument X^a with X^b,

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

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

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

a consistent estimate of β^*, the true relationship between Y^* and X^*.

C. *Two Instrumentation Strategies*

Multiple measures for X^* admit multiple instrumentation strategies that will only produce the same estimate with infinite data. The ORIV estimator consolidates the information from these different formulations, producing a more efficient estimator.[25] In our working example, we have two

[24] In practice, if one is not certain that the variation in X^a and X^b due to measurement error is identical, one can measure the proportion of measurement error in each elicitation j separately as $\text{Cov}[X^a, X^b]/\text{Var}[X^j]$. In general, it is advisable to standardize variables, at which point the formulation here coincides with that in (4). This formulation also allows a correction factor for the attenuation bias in the regression estimates from (3) dating back to Spearman (1904). Define the "disattenuated" estimator of β as $\tilde{\beta} = \hat{\beta}/\widehat{\text{Corr}}[X^a, X^b]$. The continuous-mapping theorem implies that $\tilde{\beta}$ is a consistent estimator for β. However, this approach is less efficient than ORIV.

[25] This estimator is the same as using both valid moment conditions in GMM; see app. A. ORIV offers a simpler and more transparent correction technique than GMM. In our settings, GMM and ORIV are equally efficient.

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 combine these sources of information into a single estimated relationship.

Table 5 shows estimated correlations between different elicitation techniques.[26] These are first estimated using a standard regression, and then the two different instrumental-variable strategies discussed above. Although different instrumentation strategies may produce similar results—as in columns 3 and 4 of table 5—they may also produce different results—as in columns 7 and 8. Moreover, given that estimated standard deviations—inflated by measurement error—are used to standardize the variables in table 5, neither strategy produces an accurate correlation. The next subsection deals with both of these issues.

D. Obviously Related Instrumental Variables

We construct ORIV estimates and corrected correlations in three steps. First, we consider the case in which only explanatory variables are measured with error. We then extend the analysis to the case in which 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 β. Throughout, we focus on designs in which there are at most two replications for each measure.[27] This is done for simplicity, and because it fits precisely the implementation carried out using the Caltech Cohort Study.

1. Errors in Explanatory Variables

The ORIV regression estimates a stacked model to consolidate the information from the two available instrumentation strategies. In the model of Section IV.B this can be written as

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

[26] The coefficients are from regressions in which both the left- and right-side variables are standardized, which removes scale effects and provides for easy comparison.

[27] When there are two replications of each, the ORIV estimator is twice as efficient as instrumental variables; i.e., the variance of the ORIV estimator is one-half that of instrumental-variable estimators. App. A extends the ORIV estimator to settings where more than one replicate is available.

TABLE 5
CORRELATION BETWEEN DIFFERENT RISK MEASURES IS UNDERSTATED BECAUSE OF MEASUREMENT ERROR

DEPENDENT VARIABLE	QUALITATIVE ASSESSMENT				LOTTERY MENU			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Project #1	.30***				.24***			
	(.034)				(.034)			
Project #2		.29***				.29***		
		(.034)				(.034)		
Project #1 (instr.)			.55***				.60***	
			(.067)				(.073)	
Project #2 (instr.)				.58***				.50***
				(.069)				(.070)
Risk MPL #1	.16***				.19***			
	(.036)				(.035)			
Risk MPL #2		.17***				.23***		
		(.035)				(.035)		
Risk MPL #1 (instr.)			.22***				.44***	
			(.048)				(.069)	
Risk MPL #2 (instr.)				.21***				.37***
				(.047)				(.067)

NOTE.—Coefficients are from regressions where both the right- and left-side variables are standardized, and thus are correlations. Standard errors are in parentheses. $N = 775$.
* Statistical significance at the 10 percent level.
** Statistical significance at the 5 percent level.
*** Statistical significance at the 1 percent level.

where N is the number of participants and 0_N is an $N \times 1$ zero matrix. To implement this, one should create a stacked data set and run a 2SLS regression. This is equivalent to estimating a first stage, as in (5), for both instrumentation strategies, 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.

With a single replication, the stacked regression will produce an estimate of β^* that is the average of the estimates from the two instrumentation approaches in prior subsections. Intuitively, with no theoretical reason to favor one estimate or the other, it is as likely that the smaller is too small as it is that the larger is too large.[28] The estimator splits the difference, leading to a consistent estimate of β^*, the true relationship between Y^* and X^*.

[28] Overidentification can be addressed using GMM, which allows testable restrictions that can be evaluated using a standard Sargan test.

PROPOSITION 1. ORIV produces consistent estimates of β^*.

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.[29]

PROPOSITION 2. The ORIV estimator satisfies asymptotic normality under standard conditions. The estimated standard errors, when clustered by participant, are consistent estimates of the asymptotic standard errors.

Simulations using *t*-tests show that these asymptotic standard errors reject the null slightly too often at the 1, 5, and 10 percent levels. Bootstrapped standard errors, in contrast, produce effective rejection rates equal to the nominal rate. Thus, asymptotic standard errors slightly understate the true variability of estimates in finite samples.

2. Designing Experiments for ORIV

Our analysis assumes that measurement error is independent across elicitations. To increase the chances this is true, we recommend that experimental designers follow a few standard practices. First, duplicate elicitations should use different numerical values. Second, if using an MPL, the response grid should be constructed so that implied values (usually the midpoint between two choices) are not the same. Third, duplicate items should be placed in different parts of the survey or study to alleviate any tendency for consistency of responses.

Even after adopting these practices, a common component of measurement error may still remain. Assume that this common component is in-

[29] 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

$$\begin{pmatrix} (\text{Var}[\varepsilon^*] + \beta^2\text{Var}[\nu^a])I_N & \text{Var}[\varepsilon^*]I_N \\ \text{Var}[\varepsilon^*]I_N & (\text{Var}[\varepsilon^*] + \beta^2\text{Var}[\nu^b])I_N \end{pmatrix},$$

where I_N is an $N \times N$ identity matrix. Clustering takes care of the common ε_i^* for participant i on- and off-diagonal.

As the diagonal terms of the residual covariance matrix differ in whether $\text{Var}[\nu^a]$ or $\text{Var}[\nu^b]$ remains, a different weighting of X^a and X^b is optimal for efficiency. Such a condition can be most easily identified by comparing the unconditional variances of X^a and X^b. If they differ substantially, a feasibly efficient GMM or weighted feasible-generalized least squares (FGLS) approach would yield asymptotically efficient (and equivalent) estimators. In our data set, which is an order of magnitude larger than most, FGLS weighting does not produce different results, suggesting that the assumption of homogeneous errors underlying ORIV is reasonable in our application. We present the asymptotic equivalence between feasible-efficient GMM and FGLS in app. A.6. This appendix concludes with a discussion of the numerical challenges that may arise in implementing a naive FGLS ORIV estimator due to rank deficiency in the residual covariance matrix.

dependent across measures (but not elicitations of that measure).[30] In this case, when estimating the correlation between the two measures using ORIV—or any other technique to deal with measurement error—there is a residual error in each measurement and assessed correlations would still be attenuated. Even in such settings, ORIV would be advised as it corrects for one component of measurement error. However, resulting estimates would be conservative. This illustrates a general point: survey- or session-based measurement error is difficult to estimate or interpret. In particular, if we see different patterns of responses across surveys or sessions, it would be challenging to disentangle correlated measurement error from a change in preferences.

3. 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* between two variables, classical measurement error in either will attenuate the estimated correlation.

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

$$\begin{pmatrix} Y^a \\ Y^a \\ Y^b \\ Y^b \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} + \beta \begin{pmatrix} X^a \\ X^b \\ X^a \\ X^b \end{pmatrix} + \varepsilon \text{ with instruments}$$

$$W = \begin{pmatrix} X^b & 0_N & 0_N & 0_N \\ 0_N & X^a & 0_N & 0_N \\ 0_N & 0_N & X^b & 0_N \\ 0_N & 0_N & 0_N & X^a \end{pmatrix}.$$

[30] In particular, $X^i = X^* + \eta_X + \nu_X^i$ and $Y^i = Y^* + \eta_Y + \nu_Y^i$, with $\mathbb{E}[\eta_X \eta_Y] = \mathbb{E}[\nu_X^a \nu_X^b] = \mathbb{E}[\nu_Y^a \nu_Y^b] = 0$.

4. Estimating Correlations from Consistent Coefficients

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

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

where ρ_{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 whether transforming the regression coefficient into a correlation will generate an estimate that is biased up or down. Nonetheless, we have

$$\text{Cov}[X^a, X^b] = \text{Cov}[X^* + \nu^a_X, X^* + \nu^b_X] = \text{Var}[X^*],$$

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

PROPOSITION 3. $\hat{\rho}^*_{XY}$ is consistent with an asymptotically normal distribution, where standard errors can be derived using the delta method. These standard errors can be consistently estimated using a bootstrap to construct confidence intervals.

An example of how to estimate correlations and bootstrapped standard errors using ORIV in Stata can be found in appendix C. A Stata program to estimate correlations and bootstrapped standard errors is available from the authors upon request.

E. Corrected Correlations between Risk Elicitation Techniques

We now use ORIV estimators to examine the correlations between different risk measures. Table 6 contains both the raw and corrected correlations. Both the risk MPL and the project measures were elicited twice. The qualitative risk assessment was elicited once in the fall of 2014 and again in the spring of 2015, the latter serving as the duplicate elicitation. The lottery menu measure was elicited only once.[31]

Previous work comparing different risk elicitation techniques often transforms them into a common scale (see Deck et al. [2010] for an exam-

[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 errors in the measures of X and Y are equal—i.e., $\text{Corr}[X^a, X^b] = \text{Var}[Y^*]/\text{Var}[Y]$, as shown in (4). Having only one measure of Y, we have no measure of $\text{Var}[Y^*]$, and cannot test this. However, the correlation between elicitations for the three that we do have are .67 (standard error .027, projects), .62 (standard error .028, qualitative), and .59 (standard error .29, risk MPLs), so this seems reasonable.

TABLE 6
CORRELATION MATRICES BEFORE AND AFTER ACCOUNTING FOR MEASUREMENT ERROR ($N = 775$)

	RAW CORRELATIONS			CORRECTED FOR MEASUREMENT ERROR		
	Project	Qualitative	Lottery Menu	Project	Qualitative	Lottery Menu
	In Units Given by the Questions					
Qualitative	.26***			.40***		
	(.029)			(.043)		
Lottery menu	.47***	.25***		.71***	.40***	
	(.029)	(.032)		(.046)	(.052)	
Risk MPL	.19***	.13***	.22***	.30***	.19***	.38***
	(.032)	(.033)	(.030)	(.048)	(.047)	(.053)
	Measured in CRRA Coefficients					
Qualitative	.20***			.36***		
	(.032)			(.047)		
Lottery menu	.27***	.24***		.55***	.38***	
	(.040)	(.033)		(.078)	(.053)	
Risk MPL	.18***	.070**	.22***	.37***	.10**	.42***
	(.037)	(.033)	(.037)	(.078)	(.047)	(.071)
	Measured in Certainty Equivalent of a 50/50 Lottery over 0/100 Tokens					
Qualitative	.25***			.44***		
	(.029)			(.046)		
Lottery menu	.38***	.23***		.73***	.37***	
	(.026)	(.032)		(.077)	(.051)	
Risk MPL	.24***	.13***	.20***	.43***	.19***	.34***
	(.043)	(.033)	(.027)	(.067)	(.047)	(.047)

NOTE.—Bootstrapped standard errors from 10,000 simulations are in parentheses.
* Statistical significance at the 10 percent level.
** Statistical significance at the 5 percent level.
*** Statistical significance at the 1 percent level.

ple). For comparability, we do the same in table 6. This is not theoretically advisable as it introduces a nonlinear 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 of a lottery with a 50 percent probability of 100 tokens and a 50 percent 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 .5 (and uniformly below .27 when considering imputed CRRA coefficients), corrected correlations are dramatically higher, reaching levels as high as .73. Whether this correlation is "high" or "low" is largely a judgment call. However, the literature seems to consistently suggest that correlations above .7 are very high (see, e.g., Cohen 1988; Evans 1996). Moreover, many perceived strong links correspond to correlations that are .7 or less. For example, the correlation between parents' and their children's heights hovers around .5 (Wright and Cheetham 1999); the correlation between average parents' education and their children's education ranges from around .30 in Denmark to .54 in Italy, with most western countries falling somewhere in between (Hertz et al. 2007). Second, some measures are noticeably more correlated. Namely, the project measure is most correlated with our other elicitation techniques. It is most highly correlated with the lottery menu measure, with corrected correlations of .55–.73, depending on measurement units. The lottery menu also exhibits relatively high correlations with other measures. The lowest correlations are observed between the risk MPL and the qualitative measure.

F. 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 1.[32] These differences account for differences in the explanatory power of risk controls on the gender gap in competitiveness in Section III.B.

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, e.g., Byrnes, Miller, and Schafer [1999] for a review of the relevant experimental work in psychology, and Eckel and Grossman [2008b] and Croson and Gneezy [2009] for reviews of related experimental work in economics). On the other hand, the finance literature has found a more consistent difference between men and women when considering risky financial investments (see, e.g., Embrey and Fox 1997; Farrell 2011; Barber and Odean 2013). 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

[32] In addition to the usual concerns with the Kolmogorov-Smirnov test, it is not valid for discrete distributions. In such cases, the p-value is only approximate, and may lead to extreme implications, such as the p-values of 1.000 found in fig. 2. As such, we provide p-values only to aid visual inspection.

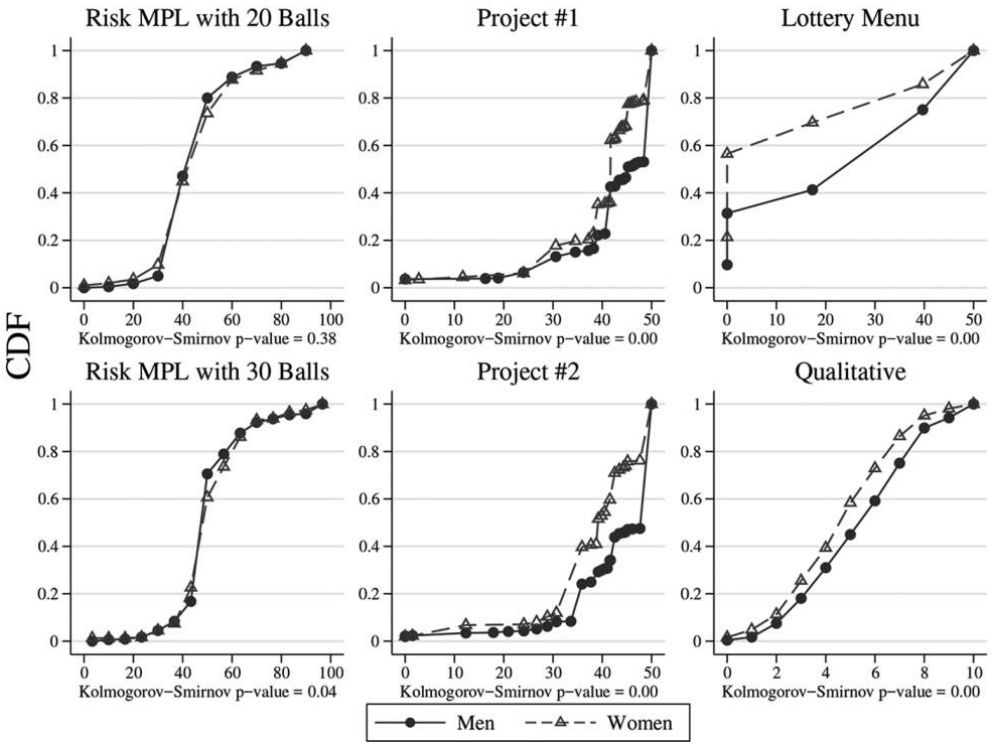

FIG. 1.—Risk aversion differs by gender, except when elicited using risk MPLs. All panels except for the qualitative assessment are put on the same scale: a certainty equivalent of a 50/50 lottery paying 0/100 tokens, via a CRRA utility function. Color version available as an online enhancement.

seen in real financial investments: men invest more aggressively than women.[33]

There is also variation in the consistency of responses to the different risk elicitation techniques across time. The project task, 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 .65 (.038) for the project measure(s), compared with .42 (.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 .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 (Slovic 1964; see also Kruger, Wang, and Wilke [2007] and references therein for more recent work). 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]

V. 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 III, due to controls measured with error. 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 a known probability distribution of rewards over a similar lottery in which the probability distribution of rewards is unknown. This behavior is expressed in

[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 percent 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 participants, which is consistent with the lottery menu task exhibiting similar patterns to the project measure.

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

the *Ellsberg paradox*, in which participants prefer a bet on the draw of a black ball from an urn with, say, 10 red and 10 black balls than on an urn with 20 balls of 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 in which the number of red balls is uniformly determined. In his results, Halevy reports correlations of around .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 implies 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. Correcting for measurement error using ORIV, ambiguity aversion and reaction to compound lotteries appear almost identical.

A. Ambiguity Aversion and Reaction to Compound Lotteries

As described in Section III.A, the risk MPL, compound MPL, and ambiguous MPL are all implemented similarly. 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 associated with positive payment. All have the same number of balls and the same payoff. The only difference is how the distribution of balls in the urn is specified: half black and half red for risk, drawn from a uniform distribution for compound, or unknown for ambiguous. Each measure is replicated twice: once with a 20-ball urn and a payoff of 100 tokens if the correct color ball is drawn, and once with a 30-ball urn and a 150-token payoff.

Our data show 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 percentage points lower than those of the risky urns (standard error of 0.48 for the 20-ball urn and 0.46 for the 30-ball urn), and the certainty equivalents of the compound lotteries are 2.9 (0.51) and 2.8 (0.51) percentage points lower than those of the risky urns. Note that these differences are statistically significant, but not significantly different from *each other*. On average, ambiguity aversion and reaction to compound lotteries are identical.

TABLE 7
CORRELATION BETWEEN CERTAINTY EQUIVALENTS IS SUBSTANTIAL ($N = 774$)

	RAW CORRELATIONS			CORRECTED FOR MEASUREMENT ERROR		
	Risk CE	Compound CE	Compound Reaction	Risk CE	Compound CE	Compound Reaction
Compound CE	.55***			.74***		
	(.035)			(.043)		
Ambiguous CE	.60***	.65***		.78***	.85***	
	(.033)	(.027)		(.037)	(.029)	
Ambiguity aversion			.44***			.86***
			(.039)			(.059)

NOTE.—Bootstrapped standard errors from 10,000 simulations are in parentheses.
* Statistical significance at the 10 percent level.
** Statistical significance at the 5 percent level.
*** Statistical significance at the 1 percent level.

Table 7 reports the raw and corrected correlations between the three measures. The raw correlation between ambiguous and compound certainty equivalents is .65. This is in line with Halevy (2007), who reports a correlation of .45 ($N = 104$) in the first round of his experiment, and a correlation of .71 in his robustness round ($N = 38$). However, once measurement error is corrected for in our data, the correlation is much higher: .85.

Corrected correlations between certainty equivalents of risky and compound or ambiguous urns are substantial as well: .74 and .78, respectively. This leads to an important point: the high correlation between ambiguity aversion and reaction to compound lotteries may be 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 measures of ambiguity aversion and (negative) reaction to compound lotteries. This results in a smaller raw correlation of .44, but the same correlation of .86 when measurement error is taken into account. Moreover, the 95 percent confidence interval for this correlation is (.74, .98).[35] As mentioned in the introduction, the correlations observed by Halevy have allowed various scholars to maintain the assumption that ambiguity aversion and reaction to compound lotteries are separate phenomena (see Epstein [2010] and Ahn et al. [2014], among others). Our corrected correlations suggest that the difference between the two attitudes is, in fact, extremely small.

Here, unlike in Section III, our large sample size is likely the reason we find any difference between these behaviors. 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 percent of the time at the 1 percent level, 4.8 percent

[35] The 99 percent confidence interval is (.71, 1.01).

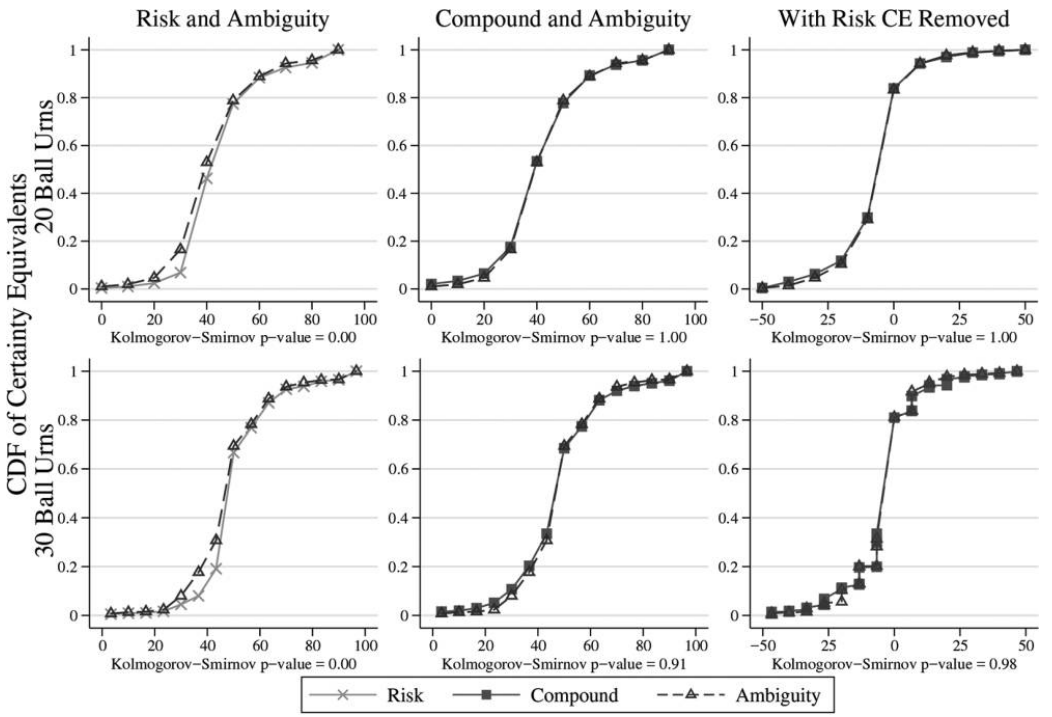

FIG. 2.—Population responses to risky and ambiguous lotteries differ, but are identical for ambiguous and compound lotteries. The first two columns are certainty equivalents for a lottery paying 0/100 points. This is the natural scale for the 20-ball urns; for the 30-ball urns the expressed certainty 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. Color version available as an online enhancement.

of the time at the 5 percent level, and 9.0 percent of the time at the 10 percent level. It should be noted that these results are from standard confidence intervals that are well calibrated. While correlations have an upper bound of 1 and thus suffer from the Andrews (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, it should be interpreted as strong evidence that the correlation is actually 1.

Finally, we plot the cumulative distribution functions (CDFs) of the various measures in figure 2. The leftmost 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 appear identical. The final panels show the distributions of ambiguity aversion and reaction to compound lotteries. Once again, these appear identical.

B. Substantive Implications

Ambiguity aversion and reaction to compound lotteries appear remarkably similar once measurement error is accounted for. It is worth noting that this is compatible with the original description of Knightian uncertainty (Knight 1921, 19): "The 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." That is, compound lotteries may be no more "susceptible to measurement" for individuals—even Caltech students, who are mathematically inclined—than those that are ambiguous.

This suggests to us that the defining characteristic, in addition to risk attitudes, in determining valuations of these lotteries is some notion of complexity. Regardless of the philosophical interpretation of these results, it is clear that any successful model of ambiguity aversion should also predict behavior in complex, but fully specified, risky environments.

VI. 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 false positives. 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 overidentification 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 distinct. Our results show that both of these beliefs 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 than previously thought. Given that these are the only three results we examined in trying to understand the influence of measurement error in experiments, it seems likely that the overidentification 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 nonreplicability 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

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

Aguiar, Victor, and Nail Kashaev. 2017. "Stochastic Revealed Preferences with Measurement Error: Testing for Exponential Discounting in Survey Data." Manuscript, Univ. Western Ontario.

Ahn, David, Syngjoo Choi, Douglas Gale, and Shachar Kariv. 2014. "Estimating Ambiguity Aversion in a Portfolio Choice Experiment." *Quantitative Econ.* 5 (2): 195–223.

Ambuehl, Sandro, and Shengwu Li. 2015. "Belief Updating and the Demand for Information." Manuscript, Stanford Univ.

Andrews, Donald W. K. 2001. "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis." *Econometrica* 69 (3): 683–734.

Barber, Brad M., and Terrance Odean. 2013. "The Behavior of Individual Investors." In *Handbook of the Economics of Finance*, vol. 2B, edited by George M. Constantinides, Milton Harris, and Rene M. Stulz, 1533–70. Oxford: North-Holland.

Battalio, Raymond C., John H. Kagel, Robin C. Winkler, Edwin B. Fisher, Robert L. Basmann, and Leonard Krasner. 1973. "A Test of Consumer Demand Theory Using Observations of Individual Consumer Purchases." *Western Econ. J.* 11 (4): 411–28.

Beauchamp, Jonathan, David Cesarini, and Magnus Johannesson. 2015. "The Psychometric and Empirical Properties of Measures of Risk Preferences." Manuscript, Univ. Toronto.

Bertrand, Marianne, and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *A.E.R. Papers and Proc.* 91 (2): 67–72.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, vol. 5, edited by James J. Heckman, 3705–3843. Amsterdam: Elsevier.

Buonaccorsi, John P. 2010. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: CRC Press.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, Competitiveness, and Career Choices." *Q.J.E.* 129 (3): 1409–47.

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

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

Charness, Gary, Uri Gneezy, and Alex Imas. 2013. "Experimental Methods: Eliciting Risk Preferences." *J. Econ. Behavior and Org.* 87 (1): 43–51.

Coffman, Lucas, and Paul Niehaus. 2015. "Pathways to Persuasion." Manuscript, Ohio State Univ.

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

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

Dave, Chetan, Catherine C. Eckel, Cathleen A. Johnson, and Christian Rojas. 2010. "Eliciting Risk Preferences: When Is Simple Better?" *J. Risk and Uncertainty* 41 (3): 219–43.

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

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

Eckel, Catherine C., and Philip J. Grossman. 2002. "Sex Differences and Statistical Stereotyping in Attitudes toward Financial Risk." *Evolution and Human Behavior* 23 (4): 281–95.

———. 2008a. "Forecasting Risk Attitudes: An Experimental Study Using Actual and Forecast Gamble Choices." *J. Econ. Behavior and Org.* 68 (1): 1–17.

———. 2008b. "Men, Women and Risk Aversion: Experimental Evidence." In *Handbook of Experimental Economics Results*, vol. 1, edited by Charles R. Plott and Vernon L. Smith, 1061–73. Amsterdam: North-Holland.

Einav, Liran, Amy Finkelstein, Iuliana Pascu, and Mark Cullen. 2012. "How General Are Risk Preference? Choices under Uncertainty in Different Domains." *A.E.R.* 102 (6): 2606–38.

Ellsberg, Daniel. 1961. "Risk, Ambiguity, and the Savage Axioms." *Q.J.E.* 75 (4): 643–669.

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

Epstein, Larry G. 2010. "A Paradox for the Smooth Ambiguity Model of Preference." *Econometrica* 78 (6): 2085–99.

Evans, James D. 1996. *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove, CA: Brooks/Cole.

Farrell, James. 2011. "Demographics of Risky Investing." *Res. Bus. and Econ. J.* (special ed.; May). http://www.aabri.com/manuscripts/FSC-11-2.pdf.

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

Friedman, Daniel, R. Mark Isaac, Duncan James, and Shyam Sunder. 2014. *Risky Curves: On the Empirical Failure of Expected Utility*. New York: Routledge.
Friedman, Milton. 1957. *A Theory of the Consumption Function*. Princeton, NJ: Princeton Univ. Press.
Gneezy, Uri, and Jan Potters. 1997. "An Experiment on Risk Taking and Evaluation Periods." *Q.J.E.* 112 (2): 631–45.
Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey. 2016. *Quantal Response Equilibrium: A Stochastic Theory of Games*. Princeton, NJ: Princeton Univ. Press.
Halevy, Yoram. 2007. "Ellsberg Revisited: An Experimental Study." *Econometrica* 75 (2): 503–36.
Hausman, Jerry A. 2001. "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *J. Econ. Perspectives* 15 (4): 57–67.
Hertz, Tom, Tamara Jayasundera, Patrizio Piraino, Sibel Selcuk, Nicole Smith, and Alina Verashchagina. 2007. "The Inheritance of Educational Inequality: International Comparisons and Fifty-Year Trends." *BE J. Econ. Analysis and Policy* 7 (2). https://doi.org/10.2202/1935-1682.1775.
Holt, Charles A., and Susan K. Laury. 2014. "Assessment and Estimation of Risk Preferences." In *Handbook of Economics of Risk and Uncertainty*, vol. 1, edited by Mark J. Machina and W. Kip Viscusi, 135–202. Amsterdam: North-Holland.
Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *Chance* 18 (4): 40–47.
Kahneman, Daniel. 1965. "Control of Spurious Association and the Reliability of the Controlled Variable." *Psychological Bull.* 54 (5): 326–29.
Knight, Frank H. 1921. *Risk, Uncertainty, and Profit*. New York: Hart, Schaffner, & Marx.
Koopmans, Tjalling Charles. 1939. *Tanker Freight Rates and Tankship Building: An Analysis of Cyclical Fluctuations*. Haarlem, Neth.: De erven F. Bohn.
Kruger, Daniel J., Xiao-Tian Wang, and Andreas Wilke. 2007. "Towards the Development of an Evolutionarily Valid Domain-Specific Risk-Taking Scale." *Evolutionary Psychology* 5 (3): 555–68.
Moore, Don A., and Paul J. Healy. 2008. "The Trouble with Overconfidence." *Psychological Rev.* 115 (2): 502–17.
Mukerji, Sujoy. 2000. "A Survey of Some Applications of the Idea of Ambiguity Aversion in Economics." *Internat. J. Approximate Reasoning* 24 (2–3): 221–34.
Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Q.J.E.* 122 (3): 1067–1101.
Ortoleva, Pietro, and Erik Snowberg. 2015. "Overconfidence in Political Behavior." *A.E.R.* 105 (2): 504–35.
Raven, James C. 1936. "Mental Tests Used in Genetic Studies: The Performance of Related Individuals on Tests Mainly Educative and Mainly Reproductive." PhD diss., Univ. London.
Reiersøl, Olav. 1941. "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis." *Econometrica* 9 (1): 1–24.
Segal, Uzi. 1987. "The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach." *Internat. Econ. Rev.* 28 (1): 175–202.
———. 1990. "Two-Stage Lotteries without the Reduction Axiom." *Econometrica* 58 (2): 349–77.
Simonsohn, Uri. 2015. "Small Telescopes: Detectability and the Evaluation of Replication Results." *Psychological Sci.* 26 (5): 559–69.
Slovic, Paul. 1964. "Assessment of Risk Taking Behavior." *Psychological Bull.* 61 (3): 220–33.

Snowberg, Erik, and Leeat Yariv. 2018. "Testing the Waters: Behavior across Subject Pools." Working Paper no. 24781 (June), NBER, Cambridge, MA.
Spearman, Charles. 1904. "The Proof and Measurement of Association between Two Things." *American J. Psychology* 15 (1): 72–101.
van Veldhuizen, Roel. 2016. "Gender Differences in Tournament Choices: Risk Preferences, Overconfidence, or Competitiveness?" Manuscript, WZB Berlin.
Wald, Abraham. 1940. "The Fitting of Straight Lines if Both Variables Are Subject to Error." *Ann. Math. Statis.* 11 (3): 284–300.
Wright, Charlotte M., and Tim D. Cheetham. 1999. "The Strengths and Limitations of Parental Heights as a Predictor of Attained Height." *Archives Disease in Childhood* 81 (3): 257–60.