Online Appendix—Not Intended for Publication

A Sampling Properties of the ORIV Estimator

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

A.1 Assumptions for the ORIV Model

To establish the Sampling Properties of ORIV, we first detail the assumptions underlying the linear regression model in latent variables (Assumption 1) and then present the classical measurement error model for our contaminated observations (Assumption 2). The assumptions here are standard, with the usual exogeneity restrictions and bounds on moments to admit asymptotic analysis, and all results can be readily extended to allow for heteroskedasticity, autocorrelation, or other relevant features of the observed data.

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

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

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

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

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

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

Assumption 2 (Replicated Classical Measurement Error)**.** *We observe multiple replicates for noisy measurements of* Y^* *and* X^* *that are tainted by classical measurement error.*

1. $X^k = X^* + \nu^k_X,\ k = 1, \ldots K_X$
   - (a) $\mathbb{E}\left[\nu^k_X|X^*, Y^*\right] = 0$
   - (b) $\mathbb{E}\left[(\nu^k_X)^2|X^*, Y^*\right] = \sigma^2_{\nu^k_X}$
   - (c) $\mathbb{E}\left[(\nu^k_X)^4|X^*, Y^*\right] < \infty$
   - (d) $\nu^k_X \perp \nu^l_X,\ \forall k \neq l$

2. $Y^k = Y^* + \nu^k_Y\, k = 1, \ldots K_Y$
   - (a) $\mathbb{E}\left[\nu^k_Y|X^*, Y^*\right] = 0$
   - (b) $\mathbb{E}\left[(\nu^k_Y)^2|X^*, Y^*\right] = \sigma^2_{\nu^k_Y}$
   - (c) $\mathbb{E}\left[(\nu^k_Y)^4|X^*, Y^*\right] < \infty$
   - (d) $\nu^k_Y \perp \nu^l_Y,\ \forall k \neq l$

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

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

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

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

[1]The conditional independence restrictions imposed in Assumptions 1 and 2 are stronger than necessary for our purposes and can be weakened to mean independence conditions using standard arguments. We make these assumptions for ease of interpretation. In principle, the conditional independence assumption provides meaningful additional restrictions that could enhance the identification power, estimator efficiency, and testable restrictions of the model. This is an interesting direction for future research.

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

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

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

Now construct the instruments for each model:

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

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

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

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

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

A.2 ORIV Estimator Consistency

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

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

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

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

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

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

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

We now recall the formula for $\hat{\beta}^*$ from equation A.1:

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

To establish consistency of the estimator, notice that:

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

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

A.3 Asymptotic Normality and Clustered Variances

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

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

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

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

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

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

$$\sqrt{N}\left(\hat{\beta}^{*}-\beta^{*}\right)=\left(\frac{1}{N}X_{OR}'W_{OR}\left(\frac{1}{N}W_{OR}'W_{OR}\right)^{-1}\frac{1}{N}W_{OR}'X_{OR}\right)^{-1}$$
$$*\ \frac{1}{N}X_{OR}'W_{OR}\left(\frac{1}{N}W_{OR}'W_{OR}\right)^{-1}N^{-1/2}W_{OR}'\varepsilon_{OR}$$
$$\rightarrow_p N\left(0,\Sigma_{\hat{\beta}^{*}}\right),\text{ where}$$
$$\Sigma_{\hat{\beta}^{*}}=\left(S_{X'W}S_{W'W}^{-1}S_{X'W}'\right)^{-1}S_{X'W}S_{W'W}^{-1}\Omega S_{W'W}^{-1}S_{X'W}'\left(S_{X'W}S_{W'W}^{-1}S_{X'W}'\right)^{-1}.$$

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

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

Therefore, we have:

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

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

$$\begin{aligned}\Sigma_{\varepsilon_{OR}} = & \sigma^2_{\varepsilon^*}\left[\left(1_{(K_X K_Y)} 1'_{(K_X K_Y)}\right) \otimes I_N\right] \\ & + \left[S_Y \otimes \left(\left(1_{K_X} 1'_{K_X}\right) \otimes I_N\right)\right] \\ & + \beta^{*2}\left[\left(\left(1_{K_Y} 1'_{K_Y}\right) \otimes S_X\right) \otimes I_N\right].\end{aligned} \tag{A.3}$$

Importantly, the result indicates that all of the non-zero entries in $\Sigma_{\varepsilon_{OR}}$ correspond to instances where the individual representing the unit of observation is the same in two different regression matrices.

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

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

□

A.4 Inference on Correlation Coefficients

A.4.1 Consistency and Asymptotic Normality

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

Proposition 3. $\hat{\rho}^*_{XY}$ *is consistent 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.*

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

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

Given that $\hat{\beta}^*$ is consistent for $\beta^* = \frac{\text{Cov}(X^*,Y^*)}{\sigma^2_{X^*}}$, we have:

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

For the limiting distribution, assume independent sampling and suppose all observed random variables have bounded 8th moments (so their kurtosis can be consistently estimated). Then $\hat{\beta}^*$, $\widehat{\text{Cov}}[X^a, X^b]$, and $\widehat{\text{Cov}}[Y^a, Y^b]$ are jointly asymptotically normal:

$$\sqrt{n}\begin{bmatrix} \hat{\beta}^* - \beta^* \\ \widehat{\text{Cov}}[X^a, X^b] - \sigma^2_{X^*} \\ \widehat{\text{Cov}}[Y^a, Y^b] - \sigma^2_{Y^*} \end{bmatrix} \to_d N\left(0, \Psi\right),$$

where:

$$\Psi = \begin{bmatrix} \sigma^2_{\hat{\beta}^*} & \sigma_{\hat{\beta}^*,\widehat{\text{Cov}}[X^a,X^b]} & \sigma_{\hat{\beta}^*,\widehat{\text{Cov}}[Y^a,Y^b]} \\ \sigma_{\hat{\beta}^*,\widehat{\text{Cov}}[X^a,X^b]} & \sigma^2_{\widehat{\text{Cov}}[X^a,X^b]} & \sigma_{\widehat{\text{Cov}}[X^a,X^b],\widehat{\text{Cov}}[Y^a,Y^b]} \\ \sigma_{\hat{\beta}^*,\widehat{\text{Cov}}[Y^a,Y^b]} & \sigma_{\widehat{\text{Cov}}[X^a,X^b],\widehat{\text{Cov}}[Y^a,Y^b]} & \sigma^2_{\widehat{\text{Cov}}[Y^a,Y^b]} \end{bmatrix}$$

reflects the asymptotic covariance matrix of the estimators $\hat{\beta}^*$, $\widehat{\text{Cov}}[X^a, X^b]$, and $\widehat{\text{Cov}}[Y^a, Y^b]$. Denoting the kurtosis of X^* and Y^* as $\zeta_X = E\left[\left(X^* - E\left[X^*\right]\right)^4\right]$ and $\zeta_Y = E\left[\left(Y^* - E\left[Y^*\right]\right)^4\right]$, respectively, algebraic manipulation yields:

$$\begin{aligned}
\sigma^2_{\hat{\beta}^*} &= \Sigma_{\hat{\beta}_{OR}} \qquad \text{(A.4)}\\
\sigma^2_{\widehat{\text{Cov}}[X^a,X^b]} &= \zeta_X - \sigma^4_{X^*} + \sigma^2_{X^*}\left(\sigma^2_{\nu^1_X} + \sigma^2_{\nu^2_X}\right) + \sigma^2_{\nu^1_X}\sigma^2_{\nu^2_X}\\
\sigma^2_{\widehat{\text{Cov}}[Y^a,Y^b]} &= \zeta_Y - \sigma^4_{Y^*} + \sigma^2_{Y^*}\left(\sigma^2_{\nu^1_Y} + \sigma^2_{\nu^2_Y}\right) + \sigma^2_{\nu^1_Y}\sigma^2_{\nu^2_Y}\\
\sigma_{\hat{\beta}^*,\widehat{\text{Cov}}[X^a,X^b]} &= \beta\frac{X_1'X_1\sigma^2_{\nu^1_X} + X_2'X_2\sigma^2_{\nu^2_X}}{X_1'X_1 + X_2'X_2}\\
\sigma_{\hat{\beta}^*,\widehat{\text{Cov}}[Y^a,Y^b]} &= \beta\frac{\sigma^2_{\nu^1_Y} + \sigma^2_{\nu^2_Y}}{2}\\
\sigma_{\widehat{\text{Cov}}[X^a,X^b],\widehat{\text{Cov}}[Y^a,Y^b]} &= \beta^2\zeta_X
\end{aligned}$$

The delta method approximation builds on a Taylor expansion of $\hat{\rho}^*_{XY}$ at $(\beta^*, \sigma^2_{X^*}, \sigma^2_{Y^*})$:

$$\begin{aligned}
\hat{\rho}^*_{XY} = \sqrt{\frac{\widehat{\text{Cov}}[X^a,X^b]}{\widehat{\text{Cov}}[Y^a,Y^b]}}\hat{\beta}^* \approx \sqrt{\frac{\sigma^2_{X^*}}{\sigma^2_{Y^*}}}\beta^* &+ \sqrt{\frac{\sigma^2_{X^*}}{\sigma^2_{Y^*}}}\left(\hat{\beta}^* - \beta^*\right)\\
&+ \frac{\beta^*}{2\sqrt{\sigma^2_{X^*}\sigma^2_{Y^*}}}\left(\widehat{\text{Cov}}[X^a,X^b] - \sigma^2_{X^*}\right)\\
&- \frac{\beta^*}{2}\sqrt{\frac{\sigma^2_{X^*}}{\sigma^6_{Y^*}}}\left(\widehat{\text{Cov}}[Y^a,Y^b] - \sigma^2_{Y^*}\right).
\end{aligned}$$

Since the remainder in the approximation is $o_p(n)$, asymptotic normality for $\hat{\rho}^*_{XY}$ follows:

$$\begin{aligned}
\sqrt{n}\left(\hat{\rho}^*_{XY} - \rho^*_{XY}\right) =_a &\sqrt{\frac{\sigma^2_{X^*}}{\sigma^2_{Y^*}}}\sqrt{n}\left(\hat{\beta}^* - \beta^*\right)\\
&+ \frac{\beta^*}{2\sqrt{\sigma^2_{X^*}\sigma^2_{Y^*}}}\sqrt{n}\left(\widehat{\text{Cov}}[X^a,X^b] - \sigma^2_{X^*}\right)\\
&- \frac{\beta^*}{2}\sqrt{\frac{\sigma^2_{X^*}}{\sigma^6_{Y^*}}}\sqrt{n}\left(\widehat{\text{Cov}}[Y^a,Y^b] - \sigma^2_{Y^*}\right)\\
\to_d &N\left(0, V^\infty_{\hat{\rho}^*_{XY}}\right).
\end{aligned}$$

To derive the asymptotic variance, we square the right-hand side of this expression and solve for its limiting expectation. Using the expressions from (A.4), we can derive the asymptotic variance as:

$$V^{\infty}_{\hat{\rho}^*_{XY}} = \frac{\sigma^2_{X^*}}{\sigma^2_{Y^*}}\sigma^2_{\hat{\beta}^*} + \frac{\beta^{*2}}{4\sigma^2_{X^*}\sigma^2_{Y^*}}\sigma^2_{\widehat{\text{Cov}}[X^a,X^b]} + \frac{\beta^{*2}}{4}\frac{\sigma^2_{X^*}}{\sigma^6_{Y^*}}\sigma^2_{\widehat{\text{Cov}}[Y^a,Y^b]} + \frac{\beta^*}{\sigma^2_{Y^*}}\sigma_{\hat{\beta}^*,\widehat{\text{Cov}}[X^a,X^b]} - \beta^*\frac{\sigma^2_{X^*}}{\sigma^4_{Y^*}}\sigma_{\hat{\beta}^*,\widehat{\text{Cov}}[Y^a,Y^b]} - \frac{\beta^{*2}}{2\sigma^4_{Y^*}}\sigma_{\widehat{\text{Cov}}[X^a,X^b],\widehat{\text{Cov}}[Y^a,Y^b]}. \tag{A.5}$$

Since all the variance and covariance terms on the right hand side of (A.5) are finite, so long as the variances $\sigma^2_{X^*}$ and $\sigma^2_{Y^*}$ are bounded away from zero, this asymptotic variance is also finite. With asymptotic normality, we can construct confidence intervals using asymptotically pivotal t-statistics centered on the null-hypothesis that $\rho^*_{XY} = \hat{\rho}^*_{XY}$ for all $|\hat{\rho}^*_{XY}| < 1$. With this specification, the bootstrap's empirical distribution function converges to the population distribution function. In particular, standard errors can be consistently estimated using a bootstrap to construct confidence intervals. □

A.4.2 Bootstrapping Correlation Standard Errors for Confidence Regions

Given the complex formula in (A.5) and the difficulty of computing fourth moments using standard statistical tools, we propose using a bootstrap to compute standard errors for constructing confidence intervals. Proposition 3 above assures that estimated standard errors using a bootstrap are consistent. Incorporating the sampling error from $\widehat{\text{Cov}}[X^a, X^b]$ and $\widehat{\text{Cov}}[Y^a, Y^b]$ requires only a small modification to the standard bootstrap procedure, as presented for the case with two replicates of X and Y in Algorithm 1 below. We present the rescaling approach to computing the ORIV correlation in step 3 as it may facilitate bootstrap implementation in some statistical packages. STATA code implementing the bootstrapped standard errors is shown in Appendix C. A STATA package for computing ORIV correlations and bootstrapped standard errors is available from the authors upon request.

Algorithm 1 Bootstrap Algorithm for ORIV Correlation Standard Errors

For $m = 1, \ldots, M$:

1. Randomly draw an index of n observations from $\{1, \ldots, n\}$ with replacement. Denote this index: $T^{(m)} = \left\{t_1^{(m)}, t_2^{(m)}, \ldots, t_n^{(m)}\right\}$.
2. Construct $X^{a,(m)}, X^{b,(m)}, Y^{a,(m)}$, and $Y^{b,(m)}$ so that:
$$X^{a,(m)} = \left[X^a_{t_1^{(m)}}, X^a_{t_2^{(m)}}, \ldots, X^a_{t_n^{(m)}}\right], \ X^{b,(m)} = \left[X^b_{t_1^{(m)}}, X^b_{t_2^{(m)}}, \ldots, X^b_{t_n^{(m)}}\right]$$
$$Y^{a,(m)} = \left[Y^a_{t_1^{(m)}}, Y^a_{t_2^{(m)}}, \ldots, Y^a_{t_n^{(m)}}\right], \ Y^{b,(m)} = \left[Y^b_{t_1^{(m)}}, Y^b_{t_2^{(m)}}, \ldots, Y^b_{t_n^{(m)}}\right].$$
3. Compute ORIV Correlations $\hat{\rho}_{XY}^{(m)}$ from $X^{a,(m)}, X^{b,(m)}, Y^{a,(m)}$, and $Y^{b,(m)}$:

   3.A. Calculate the bootstrapped attenuation factors
$$\bar{X}^{a,(m)} \equiv \frac{1}{n}\sum_{\tau=1}^{n} X^a_{t_\tau^{(m)}}, \ \bar{X}^{b,(m)} \equiv \frac{1}{n}\sum_{\tau=1}^{n} X^b_{t_\tau^{(m)}}, \ \bar{Y}^{a,(m)} \equiv \frac{1}{n}\sum_{\tau=1}^{n} Y^a_{t_\tau^{(m)}}, \ \bar{Y}^{b,(m)} \equiv \frac{1}{n}\sum_{\tau=1}^{n} Y^b_{t_\tau^{(m)}}$$
$$\hat{\sigma}^2_{X^*,(m)} = \frac{1}{n}\sum_{\tau=1}^{n}\left(X^a_{t_\tau^{(m)}} - \bar{X}^{a,(m)}\right)\left(X^b_{t_\tau^{(m)}} - \bar{X}^{b,(m)}\right),$$
$$\hat{\sigma}^2_{Y^*,(m)} = \frac{1}{n}\sum_{\tau=1}^{n}\left(Y^a_{t_\tau^{(m)}} - \bar{Y}^{a,(m)}\right)\left(Y^b_{t_\tau^{(m)}} - \bar{Y}^{b,(m)}\right).$$
   3.B. Rescale the resampled data to define $\tilde{X}^{a,(m)}, \tilde{X}^{b,(m)}, \tilde{Y}^{a,(m)}$, and $\tilde{Y}^{b,(m)}$.
$$\tilde{X}^{a,(m)} \equiv \frac{X^{a,(m)}}{\sqrt{\hat{\sigma}^2_{X^*,(m)}}}, \quad \tilde{X}^{b,(m)} \equiv \frac{X^{b,(m)}}{\sqrt{\hat{\sigma}^2_{X^*,(m)}}}, \quad \tilde{Y}^{a,(m)} \equiv \frac{Y^{a,(m)}}{\sqrt{\hat{\sigma}^2_{Y^*,(m)}}}, \quad \tilde{Y}^{b,(m)} \equiv \frac{Y^{n,(m)}}{\sqrt{\hat{\sigma}^2_{Y^*,(m)}}}$$
   3.C. Compute and store $\hat{\rho}_{XY}^{(m)}$ from the ORIV Regression of $\tilde{Y}^{(m)}$ on $\tilde{X}^{(m)}$.

The bootstrap variance estimator is:
$$\hat{V}_{BS}\left(\hat{\rho}_{XY}\right) \equiv \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\rho}_{XY}^{(m)} - \hat{\rho}_{XY}\right)^2$$

A.4.3 Finite-Sample Hypothesis Tests and Confidence Interval Coverage

We performed a series of simulations to ensure hypothesis tests and confidence intervals using asymptotic and bootstrapped standard errors have appropriate size and coverage in finite samples. The simulations specify a data generating process sampling independently in t:
$$\begin{bmatrix} X_t^* \\ Y_t^* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

We generated two contaminated measures of both X_t^* and Y_t^* with independent normally distributed measurement error, defining

$$X_t^a = X_t^* + \eta_t^{X,a} \qquad X_t^b = X_t^* + \eta_t^{X,b} \qquad Y_t^a = Y_t^* + \eta_t^{Y,a} \qquad Y_t^b = Y_t^* + \eta_t^{Y,b},$$

where each $\eta_t^{\cdot} \sim N\left(0, \sigma_\eta^2\right)$ is drawn independently of all other measurement errors.

Different simulations are used to generate data sets of size $n \in \{100, 250, 500, 1000\}$. We estimate models with correlations $\rho \in \{0, 0.25, 0.5, 0.95, 0.99\}$ and measurement error variances $\sigma_\eta^2 \in \{0, 0.1/0.9, 0.25/0.75, 0.5/0.5\}$. Hypothesis tests and confidence intervals are constructed using t-Statistics at the 1%, 5%, and 10% levels, where standard errors are calculated using the asymptotic approximation in (A.5) and the bootstrap algorithm with 10,000 replications.

Table A.1 presents the finite-sample estimated standard errors from 1,000 simulations. Across all specifications, the estimated standard errors display only moderate bias. The only notable bias appears in extreme cases featuring small samples, where measurement error variability and the underlying correlation are both high. In such settings, the bootstrap procedure occasionally generates samples in which the estimated variance $\hat{\sigma}_{Y^*}^2$ is nearly zero. The explosive impact of these draws on the estimated correlations leads to bootstrapped standard errors that greatly exceed the population standard errors. Outside of these extreme cases, the bootstrap demonstrates negligible bias in estimated standard errors. In fact, for reasonable levels of measurement errors, of the order we observe in our data, we see very little bias.

To evaluate the finite-sample performance of hypothesis tests, table A.2 reports rejection frequencies using t-Tests. As with the standard errors, when measurement error constitutes less than 50% of the variance in the underlying variables or correlations are lower than 0.95, inference using bootstrapped or asymptotic standard errors appears reasonable for any sample size. We expect these to be the ranges that are relevant for most experimental work.

Table A.1: Simulated Bias in Estimated Standard Errors for Correlations

Panel A: Bootstrap Standard Error Bias (*1,000)

$n = 100$			σ_η^2	
ρ	0	0.1/0.9	0.25/0.75	0.5/0.5
0	-5.18	-2.82	-6.30	15.47
0.25	-0.52	-0.55	0.21	3.99
0.5	-1.24	0.26	-3.79	29.09
0.95	0.47	0.28	0.25	63.98
0.99	0.50	1.00	0.45	71.29
$n = 250$			σ_η^2	
ρ	0	0.1/0.9	0.25/0.75	0.5/0.5
0	-0.61	-1.72	-0.87	-3.04
0.25	-1.66	-1.23	-2.24	-1.47
0.5	-0.73	-0.42	-0.22	-3.06
0.95	0.19	-0.03	-0.27	3.71
0.99	0.16	0.04	0.29	0.98
$n = 500$			σ_η^2	
ρ	0	0.1/0.9	0.25/0.75	0.5/0.5
0	0.72	0.92	-0.11	4.41
0.25	-0.32	0.11	0.12	-0.68
0.5	0.26	0.47	0.92	1.24
0.95	-0.17	-0.12	-0.82	0.79
0.99	0.06	0.14	0.06	0.36
$n = 1,000$			σ_η^2	
ρ	0	0.1/0.9	0.25/0.75	0.5/0.5
0	0.76	0.56	0.35	0.23
0.25	0.00	-0.18	-0.04	0.27
0.5	0.06	0.12	0.29	0.80
0.95	-0.07	-0.22	-0.72	0.98
0.99	0.03	0.04	-0.24	0.02

Panel B: Asymptotic Standard Error Bias (*1,000)

$n = 100$			σ_η^2	
ρ	0	0.1/0.9	0.25/0.75	0.5/0.5
0	-5.50	-3.07	-6.85	-6.53
0.25	-1.37	-1.38	-0.52	-7.14
0.5	-3.79	-2.15	-5.92	-2.65
0.95	-4.70	-3.15	-5.75	-4.32
0.99	4.07	7.02	0.96	-6.09
$n = 250$			σ_η^2	
ρ	0	0.1/0.9	0.25/0.75	0.5/0.5
0	-0.64	-1.70	-0.91	-3.67
0.25	-1.85	-1.42	-2.43	-2.22
0.5	-1.36	-1.01	-0.69	-4.19
0.95	-2.02	-2.16	-2.40	1.64
0.99	2.43	2.99	-0.87	-1.16
$n = 500$			σ_η^2	
ρ	0	0.1/0.9	0.25/0.75	0.5/0.5
0	0.71	0.92	-0.09	4.23
250	-0.36	0.10	0.04	-0.91
500	0.12	0.30	0.75	1.17
950	-1.83	-1.72	-2.51	0.43
990	1.46	1.77	-0.81	-0.17
$n = 1,000$			σ_η^2	
ρ	0	0.1/0.9	0.25/0.75	0.5/0.5
0	0.75	0.56	0.34	0.17
0.25	-0.03	-0.21	-0.07	0.20
0.5	0.01	0.05	0.31	0.82
0.95	-0.66	-0.71	-1.12	1.18
0.99	0.79	0.62	-1.29	-0.16

This table reports simulated finite-sample bias in estimated standard errors relative to simulated standard errors for the ORIV correlations in 1,000 simulations with normally distributed variables for a range of sample sizes (n), true correlations (ρ), and variability of measurement error (σ_η^2). Bootstrap standard errors (Panel A) are calculated using the bootstrap algorithm presented in section A.4.2 with 10,000 resampling draws. Asymptotic standard errors (Panel B) are calculated using the formula in (A.5).

For completeness, we explore more extreme cases to identify where the asymptotic approximations may not be valid. When $n = 100$, hypothesis tests typically reject the null hypothesis more often than expected based on the nominal size though the distortion is not extreme. If the true correlation is as high as 0.99 or the variance of measurement error matches or exceeds that of the underlying X^* and Y^*, the asymptotic approximation breaks down in samples with fewer than 1,000 observations. The bootstrap overstates the variance of the estimator in these settings, which results in overly conservative tests that fail to reject the null hypothesis with sufficient frequency.

Table A.2: Simulated Rejection Frequencies for Bootstrapped Correlation t-Tests

Panel A: 1% Level					Panel B: 5% Level					Panel C: 10% Level				
$n = 100$			σ_η^2					σ_η^2					σ_η^2	
ρ	0	10%	25%	50%	ρ	0	10%	25%	50%	ρ	0	10%	25%	50%
0	2.2%	1.8%	2.1%	1.5%	0	6.5%	5.2%	7.2%	5.9%	0	12.0%	10.7%	12.5%	10.5%
0.25	1.4%	1.6%	2.0%	1.4%	0.25	6.2%	6.2%	5.9%	5.9%	0.25	11.1%	11.3%	9.9%	11.1%
0.5	1.6%	1.9%	1.4%	1.1%	0.5	5.5%	5.5%	6.1%	5.1%	0.5	10.3%	11.2%	12.3%	9.1%
0.95	1.3%	1.5%	1.3%	0.3%	0.95	4.5%	5.1%	5.5%	2.1%	0.95	9.4%	9.9%	10.1%	4.9%
0.99	0.4%	0.4%	0.2%	0.2%	0.99	1.8%	2.7%	3.6%	2.0%	0.99	4.7%	6.3%	7.8%	5.3%
$n = 250$			σ_η^2					σ_η^2					σ_η^2	
ρ	0	10%	25%	50%	ρ	0	10%	25%	50%	ρ	0	10%	25%	50%
0	1.2%	1.3%	0.9%	1.3%	0	4.9%	5.9%	5.3%	6.1%	0	10.9%	11.7%	10.7%	11.5%
0.25	1.4%	1.7%	1.7%	1.8%	0.25	5.6%	5.5%	6.1%	6.0%	0.25	10.4%	10.9%	11.5%	10.7%
0.5	1.5%	1.5%	2.0%	1.7%	0.5	5.2%	5.4%	5.4%	6.0%	0.5	10.6%	10.3%	9.8%	11.4%
0.95	1.4%	2.2%	1.4%	0.5%	0.95	5.0%	5.6%	4.6%	3.0%	0.95	9.2%	10.6%	10.1%	8.3%
0.99	0.7%	0.9%	0.5%	0.5%	0.99	3.6%	4.4%	5.1%	3.0%	0.99	7.1%	10.6%	9.7%	8.6%
$n = 500$			σ_η^2					σ_η^2					σ_η^2	
ρ	0	10%	25%	50%	ρ	0	10%	25%	50%	ρ	0	10%	25%	50%
0	0.6%	1.2%	1.2%	0.6%	0	4.8%	5.4%	5.2%	3.8%	0	10.2%	9.3%	10.0%	9.3%
0.25	1.5%	1.3%	1.3%	0.9%	0.25	6.1%	5.3%	4.8%	5.8%	0.25	10.1%	10.5%	10.4%	12.2%
0.5	1.2%	1.4%	1.0%	1.0%	0.5	4.7%	4.5%	4.3%	4.2%	0.5	10.3%	9.7%	9.4%	8.9%
0.95	0.8%	0.8%	1.3%	1.1%	0.95	5.2%	5.4%	6.9%	4.6%	0.95	11.8%	10.4%	12.3%	9.1%
0.99	0.9%	0.4%	0.9%	0.6%	0.99	4.3%	4.6%	4.5%	4.7%	0.99	8.2%	10.1%	9.7%	9.8%
$n = 1,000$			σ_η^2					σ_η^2					σ_η^2	
ρ	0	10%	25%	50%	ρ	0	10%	25%	50%	ρ	0	10%	25%	50%
0	1.4%	0.6%	1.5%	1.3%	0	4.5%	5.1%	4.7%	5.2%	0	9.0%	10.1%	9.8%	9.5%
0.25	1.3%	1.2%	0.9%	1.1%	0.25	4.9%	4.7%	5.5%	4.8%	0.25	11.3%	11.3%	10.1%	11.2%
0.5	1.6%	0.9%	1.1%	1.1%	0.5	4.9%	5.0%	4.6%	4.2%	0.5	9.1%	10.4%	9.4%	9.5%
0.95	1.3%	1.4%	1.2%	0.8%	0.95	5.0%	5.4%	6.3%	4.3%	0.95	9.7%	9.9%	11.1%	9.4%
0.99	0.9%	0.6%	1.1%	0.7%	0.99	4.7%	5.4%	5.7%	5.4%	0.99	7.4%	10.0%	11.1%	9.2%

This table reports simulated rejection frequency of the null hypothesis using t-Tests with bootstrapped standard errors (presented in section A.4.2 with 10,000 resampling draws) for ORIV correlations. The results report on 1,000 simulations with normally distributed variables for a range of sample sizes (n), true correlations (ρ), and variability of measurement error (σ_η^2). Critical values are based on nominal test sizes of 1% (Panel A), 5% (Panel B), and 10% (Panel C).

Overall, these results affirm the obvious caution when performing inference in relatively small samples in which most of the variability in observations is due to measurement error. More extreme cases may require additional data or a more refined test statistic accounting for boundary issues that complicate the sampling distribution of the estimator as $\rho^*_{XY} \to 1$ (a complex statistical problem beyond the scope of the current paper).

A.5 Equivalent Estimators

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

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

As mentioned in the text, averaging of observations is common in experimental work. As it turns out, when measurement error occurs on both left- and right-side variables, a numerically equivalent estimator to ORIV can be derived by averaging observations of the left-side variable and stacking only right-side variables.

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

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

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

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

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

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

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

We complete the proof using a parallel argument to show $X_{OR}' P_{W_{OR}} Y_{OR} = K_Y \bar{X}_{OR}' P_{\bar{W}_{OR}} \bar{Y}_{OR}$.

□

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

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

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

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

Proof. We begin with the observation that $\tilde{W}^k = W^k$ and $\tilde{X}^k = X^k$ when $K_Y = 1$. From the previous subsection, we can write:

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

□

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

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

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

A.6 GMM Representation, Efficiency, and GLS Estimation

We close our discussion of the ORIV estimator with a brief comment on the estimator's efficiency. First, we present an unweighted GMM representation for the ORIV estimator to establish ORIV's efficiency in a homoskedastic model where measurement error has a constant variance across replicates. In the special case where measurement errors are homoskedastic, the ORIV estimator is asymptotically efficient. In general, when measurement errors are characterized by heteroskedastic variances, the efficiently weighted GMM estimator corresponds to a GLS implementation of ORIV that can improve efficiency.

A.6.1 GMM Representation of the ORIV Estimator

ORIV's efficiency derives from combining all available exclusion restrictions when estimating the regression model. Let W_t^j denote the t^{th} row of the matrix W_t, construct the $K_Y K_X$ identifying moment conditions:

$$g_t^{k,j} \equiv E\left[\left(Y_t^k - X_t^j\beta\right)W_t^j\right] = 0,\ k = 1,\ldots,K_Y,\ j = 1,\ldots,K_X.$$ [2]

We can consolidate these conditions using a vector representation:

$$g_t \equiv \left[g_t^{1,1},\ldots,g_t^{1,K_X},g_t^{2,1},\ldots,g_t^{2,K_X},\ldots g_t^{K_Y,1},\ldots,g_t^{K_Y,K_X}\right]'.$$

[2] This mean independence assumption is implied by the conditional independence assumptions in Assumptions 1 and 2. However, the conditional independence assumptions provide additional restrictions on higher order moments such that $E[(Y_t^k - X_t^j\beta)f(W_t^j)] = 0$ for any function f. Under conditional independence, these additional restrictions are informative and could be used to enhance estimator efficiency or test the validity of the independence assumptions.

The GMM objective function weights the quadratic loss for the sample-estimated moment conditions $\bar{g}_n \equiv \frac{1}{n}\sum_{t=1}^{n} g_t$ so that:

$$\hat{\beta}_{GMM} = \operatorname{argmin} \bar{g}_n' \mathcal{W} \bar{g}_n$$

where $\mathcal{W}$ is a $K_Y K_X \times K_Y K_X$ weighting matrix.

In our leading specification, which features homoskedastic errors and homogeneous variances across replicated measures, the efficient choice for $\mathcal{W}$ is the identity matrix (since all moment restrictions are equally informative). The GMM objective function in this case exactly matches the objective function minimized by the ORIV estimator, establishing ORIV's efficiency as demonstrated in Section A.6.2. More generally, in the presence of heteroskedasticity or heterogeneous contamination variances, the identity-weighting matrix may be inefficient. In these cases, discussed in Sections A.6.3 and A.6.4, the efficiently-weighted GMM estimator is equivalent to an FGLS implementation of ORIV that also achieves efficiency.

A.6.2 Efficiency in Homoskedastic Setting: $K_Y = 1$

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

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

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

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

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

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

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

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

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

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

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

Proposition 6 (Two-Parameter Covariance Matrix for Individual IV Estimators)**.** *Under Assumptions (1)–(3), the variance-covariance matrix for estimates from the individual IV models for β^* features constant correlations and homoskedastic variances. That is,*

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

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

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

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

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

With respect to ω_{12}, notice that

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

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

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

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

A.6.3 Efficiency in Heteroskedastic Settings: $K_Y = 1$

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

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

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

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

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

Without reweighting observations, the ORIV estimator will assign weights of $w_{i,OR} = \frac{X^{i\prime} P_{W^i} X_i}{\sum_{k=1}^{K}\left(X^{k\prime} P_{W^k} X_k\right)}$. Therefore, consider reweighting observations for the i^{th} model to achieve the optimal weights. Specifically, defining $\lambda_i \equiv \sqrt{\frac{w_i^*}{w_{i,OR}}}$ and $\lambda \equiv [\lambda_1, \ldots, \lambda_{K_X}]'$, let

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

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

A.6.4 Efficiency in Heteroskedastic Settings: $K_Y \geq 2$

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

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

A.6.5 Estimating Measurement Error Variance and FGLS Efficiency: $K_Y = 1$

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

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

Feasible implementation then requires estimating s_{10} and s_{1i}. Recalling their definition from (A.2), a feasible estimator for s_{10} would average over the cross-products between residuals corresponding to two different elicitations of X^* for the same individual. The feasible estimator for s_{1i} then just sums those squared residuals corresponding to the i^{th} model's specification.

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

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

A.6.6 Estimating Measurement Error Variance and FGLS Efficiency: $K_Y \geq 2$

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

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

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

A.6.7 A Note of Caution on FGLS Implementations

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

B Other Ways of Using Controls

B.1 Principal Components

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

A growing literature in statistics and econometrics considers inference in the presence of a high-dimensional set of sparse controls.[3] One way to cope with this issue is to perform model selection after rotating the controls into their principal components, following the strategy proposed by Belloni, Chernozhukov, and Hansen (2013). This transformation concentrates the information from the controls into relatively few factors, effectively controlling for a rich characterization of risk preferences and overconfidence without giving up too many degrees of freedom.

Table B.1 illustrates this approach. We first conduct a principal components analysis of all 76 controls used in the last column of Table 3. We then enter these principal components sequentially in the columns of Table B.1. As can be seen, the first principal component causes the coefficient on gender to fall by approximately half, the third by a further half,

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

Table B.1: Principal components can be used to save degrees of freedom.

Dependent Variable:	Chose to Compete ($N = 783$)					
Male	0.19***	0.10***	0.10***	0.054*	0.054*	0.041
	(.034)	(.034)	(.034)	(.033)	(.033)	(.033)
First Principal Component		0.14***	0.14***	0.15***	0.15***	0.15***
		(.017)	(.017)	(.016)	(.016)	(.016)
Second Principal Component			0.020	0.020	0.019	0.019
			(.016)	(.015)	(.015)	(.015)
Third Principal Component				-0.14***	-0.13***	-0.14***
				(.015)	(.015)	(.015)
Fourth Principal Component					-0.0094	-0.0093
					(.015)	(.015)
Fifth Principal Component						0.034**
						(.015)
Adjusted R^2	0.038	0.12	0.12	0.20	0.20	0.20

Notes: ***, **, * denote statistical significance at the 1%, 5%, and 10% level, with standard errors in parentheses. Coefficients and standard errors on all non-dichotomous measures are standardized. $N = 783$.

and the fifth by an additional 25%. The second and fourth have no impact on the coefficient on gender. The adjusted-R^2 begins to drop when the 12th component is entered, at which point the coefficient on male is 0.0431 (s.e. 0.0330)—that is, the coefficient on male does not change meaningfully as components 6 through 11 are added. Importantly, this suggests one can control for all relevant variation using only five controls. Moreover, this strategy allows the use of non-parametric or semi-parametric versions of the controls we enter linearly.

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

B.2 Measurement Error in Binary Choices

As mentioned in the text, Niederle and Vesterlund (2007) attempt to control for preferences with another competition choice. Namely, in the last stage of the experiment (their Task 4), participants are given a second opportunity to be paid for their performance in the piece-rate task (Task 1). They may choose whether to be paid as a piece rate, or to enter their Task 1 performance into a tournament with the other three participants in the group (this decision affected participants' payoffs when Task 4 was randomly chosen for payment by the experimenters). This choice has the same payoff features as the competitive choice given in their main task (Task 3). The idea is that the choice in Task 4 would control for risk aversion, overconfidence, and feedback aversion, so that different choices in Task 3 and Task 4 would be explained by a difference in preferences for competition per se. However, in the presence of measurement error, this control is subject to the issues highlighted in Section 3. Here, we explore the effect of measurement error in binary controls theoretically. We then show that an analysis of NV's data that properly accounts for measurement error leads to the same conclusion we arrive at, that gender differences in competition are driven by gender differences in fundamental attitudes such as risk aversion and overconfidence.

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

B.2.1 Stochastic Choice with Binary Preference Control

This section provides the formal arguments underlying Section 3.3. In the general formulation, let individual i have latent characteristics X_i (risk and overconfidence), and be identified by treatment D_i (gender). Each such individual faces two binary decision tasks in which they report $Y_i^a \in \{0, 1\}$ (competition) as well as its replicate $Y_i^b \in \{0, 1\}$ (entering piece-rate performance into a competition). Each individual answers 1 to both the decision task and the replicate with probability $p_i(X_i)$ that depends on X_i but is independent of treatment, D_i, so that

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

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

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

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

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

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

Without measurement error, the estimate of β would be zero.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B.2.2 Numerical Example

Note that choosing to compete in NV is tantamount to choosing a lottery that pays a fixed amount with some probability—the probability of winning the tournament—and zero otherwise, over a sure thing—the piece-rate payment.[4] We use this observation to present a numerical example.

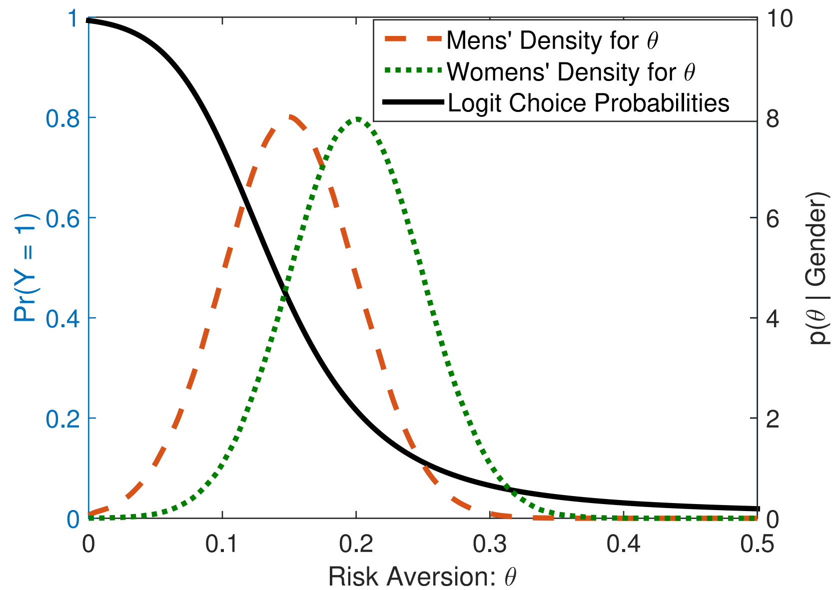
In the setup of the competition task, the choice will be driven by the interaction of subjective probabilities of winning (overconfidence) with risk aversion. To simplify, we consider only variations in risk aversion for a fixed lottery and certainty equivalent. Participants are given the choice between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

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

To synthetically calibrate preferences, participants who have $\theta_i = \frac{\log(4/5)}{\log(1/5)} \approx 0.14$ are indifferent between the lottery and the certainty equivalent, choosing each with equal probability. We assume the risk-aversion parameter is distributed normally for men and women, with

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

Figure B.1: Example Illustrating Lottery Choice Probabilities and Gender-Differences in Risk Aversion



each distribution having standard deviation 0.05. To calibrate choices to the 19.0% gap in competition we observe in our data, we assume the distribution for men has mean 0.15, and for women, 0.2. This results in men and women who will choose the lottery about 46% and 27% of the time, respectively. Figure B.1 illustrates the probability of choosing the lottery conditional on risk aversion and the distributions over risk aversion for men and women.

Let $Y_i^a = 1\{\text{Player } i \text{ Chooses Lottery in Main Decision}\}$ represent the observed choice for participant i in the main decision task. If we knew θ_i and could compute each individual's p_i, then $p_i \equiv \mathbb{E}[Y_i^a|\theta_i]$ would be the best predictor for Y_i^a and the ideal explanatory variable and control. As such, following NV, we describe p_i as an individual's *preferences*. A regression of Y_i^a on gender and preferences p_i would fail to find any relationship between gender and lottery take-up (competition) in a large sample.

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

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

This suggests that by forcing the coefficient on Y_i^b to be 1, the resulting regression can produce a valid test. That is, by regressing $Y_i^a - Y_i^b$ on gender, the resulting coefficient will be a conservative test of the effect of gender controlling for the second choice.[5] Following this prescription in our simulated data gives a very accurate and precise estimate of -0.000078 (s.e. 0.034, $p = 1.00$).

C Example STATA

C.1 Using Principal Components as Controls

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

#delimit;
pca ss* scc* pdd* *RiskyProjectAllocation riskyUrn*0MaxValue compoundUrn*
*Over* CRTCorrect CRTPercentile ravenCorrect ravenPercentile guess*Confidence
gpaPercentile;
#delimit cr

predict p1-p76

//stardardize principal components to make output easily interpretable
foreach x of varlist p1-p5 {
sum ‘x’
```

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

```
replace `x' = (`x'-r(mean))/r(sd)
}

reg sumsCompete male p1-p5
```

C.2 ORIV When Y is Measured without Error

This subsection and the next implement ORIV assuming that the X and Y variables are on the same scale. If they are not, one should first put them on the same scale. If there is an obvious way to do this, as in the case of certainty equivalents of lotteries with the same probabilities, but a different high option, this should be done. Otherwise standardization of the variables may be attractive.

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

```
keep id highLowValue firstProjectValue secondProjectValue
//standardize variables
for each var of varlist highLowValue-secondProjectValue {
    quietly sum `var'
    replace `var' = (`var' - r(mean))/r(sd)
}

expand 2, generate(replicant)
gen mainVar = firstProjectValue if replicant == 0
replace mainVar = secondProjectValue if replicant == 1
gen instrument = secondProjectValue if replicant == 0
replace instrument = firstProjectValue if replicant == 1

//One constant per stack
forvalues x=0/1 {
    gen constant`x' = replicant == `x'
}

ivregress 2sls highLowValue (mainVar = instrument) control*, cluster(id) nocons
```

C.3 ORIV Estimates of a Correlation

```
keep id riskyUrn20Value riskyUrn30Value compoundUrn20Value compoundUrn30Value
replace id = _n
//standardize variables
for each var of varlist riskyUrn20Value-compoundUrn30Value {
    quietly sum ‘var’
    replace ‘var’ = (‘var’ - r(mean))/r(sd)
}

//Duplicate data four times for each way of stacking data
expand 4
sort id
gen replicant = mod(_n,4)

//Created stacked variables
gen LHS = riskyUrn20Value if replicant <= 1
replace LHS = riskyUrn30Value if replicant >= 2
gen mainVar = compoundUrn20Value if replicant == 0 | replicant == 2
replace mainVar = compoundUrn30Value if replicant == 1 | replicant == 3
gen instrument = compoundUrn30Value if replicant == 0 | replicant == 3
replace instrument = compoundUrn20Value if replicant == 1 | replicant == 2

//One constant per stack
forvalues x=1/4 {
    gen constant‘x’ = replicant == ‘x’-1
}

//Compute ORIV coefficients and correlations
ivregress 2sls LHS (mainVar = instrument) constant*, cluster(id) nocons
local correctedCoefficient = _b[mainVar]
qui corr riskyUrn20Value riskyUrn30Value if replicant == 0, cov
local correctedYVar = r(cov_12)
qui corr compoundUrn20Value compoundUrn30Value if replicant == 0, cov
local correctedXVar = r(cov_12)
local correctedCorrelation = ///
    ‘correctedCoefficient’*sqrt(‘correctedXVar’/‘correctedYVar’)

display "The ORIV Correlation is: ‘correctedCorrelation’"
```

C.4 Bootstrapped Standard Errors

```
//note: should already have dataset generated above in memory
local reps = 10000

keep if replicant == 0
drop replicant

capture program drop orivbstrap
program define orivbstrap, rclass
    preserve
    bsample,cluster(id)

    quietly ivreg LHS (mainVar = instrument) control*, nocons
    scalar BSCorrectedCoefficient = _b[mainVar]

    quietly corr riskyUrn20Value riskyUrn30Value, cov
    scalar BSCorrectedYVar = r(cov_12)
    quietly corr compoundUrn20Value compoundUrn30Value, cov
    scalar BSCorrectedXVar = r(cov_12)

    return scalar BSCorrectedCorrelation = ///
    BSCorrectedCoefficient*sqrt(BSCorrectedXVar/BSCorrectedYVar)
    restore
end

simulate BSCorrectedCorrelation = ///
    r(BSCorrectedRho), reps(`reps') seed(10) nodots: orivbstrap
bstat, stat(correctedCorrelation)
```

D Comparison of Participant Pools

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

D.1 The Beauty Contest

Each installment of the CCS contained a beauty contest game following Nagel (1995). Each participant is asked to choose a number between 0 and 100. They are told that they will be grouped with 9 randomly-chosen participants, and if their choice is closest to 2/3 of the average of the choices in that group, they will receive a $5 reward, otherwise nothing.[6]

This task is often viewed as a proxy for cognitive sophistication. Any choice above 66.66 is dominated. If everyone uses undominated strategies, any choice above 44.44 will not pay off. In this way, each iteration can be associated with a greater level of sophistication. In the limit, this iterated elimination of dominated strategies yields the unique equilibrium of the game—everyone should choose 0. Nagel (1995) reported numbers very far from 0 when the game was played for the first time by Bonn University students. In her data, the mean number chosen was 36.73 and the median was 33. Data from the first installment of the CCS in the Fall of 2013 yields similar results: In our sample of 806 participants, the mean was 37.17 with a median of 33.[7]

D.2 The Cognitive Reflection Test

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

The number of correct answers on the CRT is viewed as a proxy for meta-cognition. It is associated with time preferences: participants who score higher are more "patient". In the CCS, the fraction of participants with 0, 1, 2, and 3 correct answers are 19%, 25%, 28%, and 27%, respectively. Frederick (2005) reports responses from 11 participant populations

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

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

with aggregate proportions of 33%, 28%, 23%, and 17%. In the CCS, participants were less likely to answer no question correctly, and more likely to answer all three questions correctly, while the rates for 1 or 2 correct answers are similar.

D.3 Dictator Giving

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

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

D.4 Risky Projects

We conclude in Section 4 that a particularly appealing risk elicitation is the "Project" measure, described in Section 2.1, and based on Gneezy and Potters (1997). Here we focus on a particular implementation of this elicitation in the Fall 2014 installment of the CCS. This implementation allowed participants to allocate 200 tokens between a safe option and a risky project, the latter returning 2.5 tokens with 50% probability per token allocated.[8] In our data, the mean allocation to the risky project is 74%, with a median of 75%.

In comparison, Agranov and Yariv (2015) elicit precisely the same measure in 8 sessions

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

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

E Question Wordings

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

E.1 Competition

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

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

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

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

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

   [radio buttons next to numbers 1 through 4]

4. You will be given an additional three minutes to complete as many sums as possible. Please pick how you would like to be paid from the following two options:
   (a) 10 tokens per sum completed; or
   (b) When all surveys are submitted, we will randomly group you with 3 other people (so you will be in a group of 4). We will compare the number of sums you complete now with the number of sums the other 3 completed in the previous stage you just concluded. **You will be paid only if you achieve the highest number of completed sums within this group, in which case you will be paid 40 tokens per sum completed.**

      In case you tie with another participant(s), we will randomly determine whom of these gets first place, and you will be paid only if it is you who is declared the winner.
   (c) [3 minutes in which to do as many sums as the participant can]

Some other details of our implementation are worth mentioning. As these are identical to the design choices of NV, we address them here, rather than in the main text.

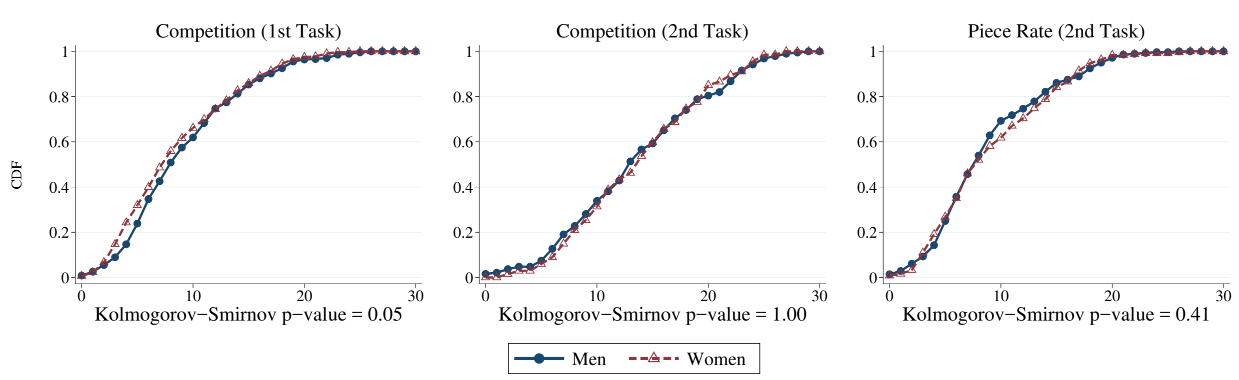
Generation of sums: The five two-digit numbers used for the tasks were generated randomly.

Payment for guessed tournament rank: If there were ties in a group, Question 3 was paid in the way that was most favorable to participants. For example, if a participant correctly solved 14 sums, while others in the group solved 14, 12, and 11 sums, the participant would get the 50 token reward with a guessed rank of either 1 or 2.

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

Figure E.2 illustrates the performance of men and women in both tasks, splitting participants in the second task into those who chose to compete and those who chose the piece-rate scheme. There is no significant difference between the performance of men and women in the second task, regardless of the compensation scheme.

Figure E.2: Performance in different tasks by gender



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

A final note on our implementation is warranted: because our experiment was conducted on a survey rather than in the lab we could not enforce our prohibition on the use of calculators. To mitigate this, sums were presented as images so they would be impossible to cut and paste into an online calculator. Participants could not use the back button during the task, and logging out and returning caused them to skip the task altogether. Participants were asked to sign an acknowledgement that they would abide by the Caltech Honor Code. Despite all this, we cannot eliminate the possibility that some students used calculators. However, in order for this to affect our results, it would have to be the case that women believed men used calculators more often, and that this belief affected not only their decision to compete, but also their choices in unrelated risk aversion and overconfidence elicitations.

Moreover, as noted in the text, we later had 98 Caltech students take the survey in the lab, where they could not use calculators. All results were substantially the same, but with larger standard errors driven by the smaller sample.

E.2 Overconfidence

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

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

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

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

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

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

   (a) No confidence at all
   (b) Not very confident
   (c) Somewhat unconfident
   (d) Somewhat confident
   (e) Very confident
   (f) Certain

[Repeat three times.]

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

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

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

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

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

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

8. This next task asks you to answer five logical questions. You will have up to 30 second to answer each question and will be paid 20 tokens for each question answered correctly.
9. A monitor and a keyboard cost $350 in total. The monitor costs $300 more than the keyboard. How much does the keyboard cost?
10. It takes 10 computers 10 minutes to run 10 simulations. How long does it take 200 computers to run 200 simulations?
11. In the pond in front of Baxter Hall, there is a patch of lily pads. The patch doubles in size every day. If it takes 36 days for the patch to cover the entire pond, how many days would it take to cover half the pond?
12. Professor Wiseman spent one-fourth of his life as a boy, one-eighth as a youth, and one-half as an active man. If Professor Wiseman spent 8 years as an old and wise man, how many years did he spend as an active man?
13. A 4 foot pole casts a shadow that is 2 feet long on the ground. If the pole was 16 feet in height, how long would the shadow be?
14. How many of the 5 questions do you think you answered correctly?
15. Out of 100 other randomly picked survey participants, what percentage do you think answered more questions correctly than you?

E.3 Risk Elicitations

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

E.3.1 Projects

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

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

   The risky project has a 40% chance of success.

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

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

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

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

   The risky project has a 50% chance of success.

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

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

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

E.3.2 Urn MPL

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

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

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

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

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

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

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

   (a) red

   (b) black

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

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

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

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

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

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

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

   (a) red

   (b) black

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

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

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

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

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

E.3.3 Lottery Menu

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

9. Which of the following gambles would you prefer?

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

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

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

E.3.4 Qualitative Risk Assessment

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

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

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

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

E.4 Compound and Ambiguous Lotteries

Compound and ambiguous lotteries follow the same format as the risky urn lotteries above. These three blocks of urn gambles were spaced out throughout the survey, and their order was randomized. The compound lotteries were also used as a control (risk-aversion measure) in the regressions regarding competitiveness.

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

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

   (a) red

   (b) black

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

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

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

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

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

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

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

   (a) red
   (b) black

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

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

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

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

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

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

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

   (a) red
   (b) black

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

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

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

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

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

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

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

   (a) red

   (b) black

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

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

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

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

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

References

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

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

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

Candes, Emmanuel and Terence Tao. 2007. "The Dantzig Selector: Statistical Estimation when p is much Larger than n." *Annals of Statistics* 35 (6):2313–2351. URL `http://dx.doi.org/10.1214/009053606000001523`.

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

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3):522–550.

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

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

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

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

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

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

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

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

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

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

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

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

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