Supplemental Web Materials: An Extended Sensitivity Analysis for Heterogeneous Unmeasured Confounding

A Construction of Valid Finite-Sample Uncertainty Sets

We now describe the construction of two $100(1-\alpha)\%$ uncertainty sets for Π^* valid for any number of pairs I. The first is based on Hoeffding's inequality, which implies that the set

$$\mathcal{H}_\beta(\Gamma, \mu_{\pi^*}) = (-\infty, \mu_{\pi^*} + I^{-1/2}\left\{1/2\log(1/\beta)(\Gamma/(1+\Gamma) - 1/2)^2\right\}^{1/2}]$$

satisfies $\mathbb{P}(\bar{\Pi} \in \mathcal{H}_\beta(\Gamma, \mu_{\pi^*})) > 1-\beta$ for all values of I. The second combines Bennett's inequality and the Bhatia-Davis inequality to create the set

$$\begin{aligned} \mathcal{B}_\beta(\Gamma, \mu_{\pi^*}) &= (-\infty, \bar{\mu}_{\pi^*} + b_\beta(\Gamma, \mu_{\pi^*}, I)] \\ b_\beta(\Gamma, \mu_{\pi^*}, I) &= \texttt{SOLVE}\{a : I^{-1}\log(1/\beta)(\Gamma/(1+\Gamma) - 1/2)^2/\nu^2(\Gamma, \mu_{\pi^*}) = \\ &\qquad h\left(a(\Gamma/(1+\Gamma) - 1/2)/\nu^2(\Gamma, \mu_{\pi^*})\right)\}, \end{aligned}$$

where $h(x) = (1+x)\log(1+x) - x$. $\mathcal{B}_\beta(\Gamma, \mu_{\pi^*})$. This set also satisfies $\mathbb{P}(\bar{\Pi} \in \mathcal{B}_\beta(\Gamma, \mu_{\pi^*})) > 1-\beta$ for any I if $\mathbb{E}[\bar{\Pi}^*] = \mu_{\pi^*}$.

In practice, the upper bound of the set based on Bennett's inequality is smaller than that based on Hoeffding's inequality when μ_{π^*} is far from $(\Gamma/(1+\Gamma) + 1/2)/2$, while the ordering reverses when μ_{π^*} is close to the midpoint. The price paid for this exactness for any I is that the upper bounds for both intervals are larger than those of $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$, the asymptotically valid uncertainty set based on the central limit theorem.

As noted in the manuscript, the general reliance of our implementation on asymptotic normality reduces the attractiveness of these finite sample uncertainty sets; however, in the case of McNemar's test with binary data, employing either $\mathcal{H}_\beta$ or $\mathcal{B}_\beta$ yields an extended sensitivity analysis for Fisher's sharp null valid for any sample size. `R` functions to compute these uncertainty set can be found in the file `multipliers.R` in the supplementary materials.

B Constructing the WLS same-sex sibling sample

Of the 10,317 in the WLS sample, 7,928 had a randomly chosen sibling who was surveyed. Of those 7,928 subjects with sibling data, 2,106 had information about sibling status (i.e. full, half or step siblings) of which 2,004 were full siblings. 1,486 of these sibling pairs were same-sex siblings of which 49.3% were men. Of the same-sex sibling pairs, there were 749 (40.6% men) where both had no more than a high school education, 265 (64.9% men) where both had at least two years of college education, and 323 (58.8% men) where one had at most a high school degree and the other had at least two years of college education. Of the same-sex pairs discordant in educational attainment, 171 (80.7% men) had complete IQ data and non-zero reported income.

C Calibrating Sensitivity Parameters to Disparities in IQ in the WLS Study

We follow a modified version of the calibration strategy introduced in Hsu and Small (2013) which involves estimating putative treatment and outcome models as a function of (X, U) under H_0 via maximum likelihood where the likelihood is marginalized over the unknown confounder U. Our modification is as follows: instead of marginalizing over the unobserved covariate we suppose that the only unobserved confounder in the Ashenfelter study is intelligence, which is measured via baseline IQ scores in the WLS study. Consequently, estimating the bias due to IQ disparities using the WLS data permits a cross-study calibration of the Ashenfelter and Rouse sensitivity analysis.

By definition, X_f is controlled automatically between siblings. We make the stylized assumption that $X_s = AGE$. Further, we assume that AGE does not affect treatment assignment. Finally, we assume that intelligence is the only unmeasured confounder in the Ashenfelter and Rouse study (i.e. $U = IQ$). Under these assumptions, a possible model for treatment assignment is

$$\mathbb{P}(Z_{ij} = 1 \mid X_{f,i},\, X_{s,ij},\, U_{ij}) = \frac{\exp(\alpha_{Z,i} + \beta_{Z,IQ} \cdot IQ_{ij})}{1 + \exp(\alpha_{Z,i} + \beta_{Z,IQ} \cdot IQ_{ij})}\,. \tag{1}$$

The pair specific intercept $\alpha_{Z,i}$ captures the $X_{f,i}$ effects. We estimate the treatment model using conditional likelihood maximization using the `R` function `clogit` in order to avoid bias arising from the fact that the number of α_i to be estimated grows with the sample size. We consider a Gaussian linear model for the outcome

$$Y_{ij} = \alpha_{Y,i} + \beta_{Y,AGE} \cdot AGE_{ij} + \beta_{Y,IQ} \cdot IQ_{ij} + \epsilon_{ij} \quad \text{such that } \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)\,. \tag{2}$$

We estimate the treatment assignment and outcome models using the 171 discordant sibling pairs that we analyze from the WLS study in the paper.

In the Ashenfelter and Rouse twins study, AGE is controlled within twin pairs so we are interested in calibrating the sensitivity parameters to the estimated bias due to IQ disparities alone. Following Hsu and Small (2013) we estimate that, controlling for age and assuming that IQ is the only confounding factor, the probability that the sibling that went to college reported a higher income in pair i to be

$$\pi_i(\mathbf{IQ}) = \frac{\exp\{\hat{\beta}_{Z,IQ}(IQ_{i1} - IQ_{i2})\}\exp\{(\hat{\beta}_{Y,IQ}/\hat{\sigma}^2)(Y_{i(2)} - Y_{i(1)})(IQ_{i1} - IQ_{i2})\} + 1}{[1 + \exp\{\hat{\beta}_{Z,IQ}(IQ_{i1} - IQ_{i2})\}][1 + \exp\{(\hat{\beta}_{Y,IQ}/\hat{\sigma}^2)(Y_{i(2)} - Y_{i(1)})(IQ_{i1} - IQ_{i2})\}]}$$

where $Y_{i(1)} = \min\{Y_{i1}, Y_{i2}\}$ and $Y_{i(2)} = \max\{Y_{i1}, Y_{i2}\}$. Define $\boldsymbol{\pi}(\mathbf{IQ})$ to be the 171×1 vector of $\pi_i(\mathbf{IQ})$. Letting $\boldsymbol{\pi}^*(\mathbf{IQ}) = \boldsymbol{\pi}(\mathbf{IQ})$ when $\hat{\beta}_{Z,IQ}\hat{\beta}_{Y,IQ} \geq 0$ and $1 - \boldsymbol{\pi}(\mathbf{IQ})$ otherwise, one reasonable set of estimates for $(\Gamma,\ \bar{\Gamma})$ is $(\pi_{max}/(1 + \pi_{max}),\ \bar{\pi}/(1 + \bar{\pi}))$ where $\pi_{max} = \sup_i \pi_i^*(\mathbf{IQ})$ and $\bar{\pi} = (1/171)\sum_{i=1}^{171} \pi_i^*(\mathbf{IQ})$. It may concern some that $\pi_{max}/(1 + \pi_{max})$ is a downwardly-biased estimator of Γ, but due to sampling variability and possible misspecification of the treatment and outcome models, the calibration is inherently approximate and meant only to act as a guide for the researcher conducting a sensitivity analysis of the Ashenfelter and Rouse study. It should also be noted that since higher IQ does not perfectly predict higher earnings, we find ourselves in a simultaneous sensitivity framework where we simultaneously bound the dependence between IQ and education and between IQ and earnings (see Gastwirth et al. (1998) for further details). This explains the slightly different definition of π_i^* used here than the one found in the paper. Simultaneous sensitivity analysis is closely related to amplified sensitivity analysis, which we discuss briefly in §7 of the paper (see Rosenbaum and Silber (2009) for more details). For our purposes, the simultaneous framework suffices to calibrate Γ and $\bar{\Gamma}$ in the Ashenfelter and Rouse study to the WLS study.

D Details of Histogram in Right Panel of Figure 2

The figure in the right panel of Figure 2 in the paper is described as the *[h]istogram of the estimated increase in pairwise bias due to IQ disparities between siblings measured as an odds ratio.* To be specific, and using the notation introduced in Appendix C, this is a histogram of

$$\frac{\pi_i^*(\mathbf{IQ})}{1-\pi_i^*(\mathbf{IQ})} \bigg/ \frac{\pi_i^*(\mathbf{0})}{1-\pi_i^*(\mathbf{0})} \tag{3}$$

for $i = 1, \ldots, 171$ where $\pi_i^*(\mathbf{0}) = (1/2)$ is π_i^* computed for the sibling pair i had they had same IQ scores.

E Additional Simulation Results

E.1 Type I Error Control (*Unbiased*)

	Γ				
$\bar{\Gamma}$	1	1.1	1.25	1.5	2
1	0.049	0.044	0.042	0.050	0.045
1.05		0.018	0.010	0.008	0.004
1.1		0.016	0.007	0.002	0.001
1.15			0.005	0.000	0.000
1.2			0.003	0.000	0.000
1.25			0.004	0.001	0.000
1.3				0.000	0.000
1.35				0.000	0.000
1.4				0.001	0.000
1.45				0.000	0.000
1.5				0.000	0.000
1.6					0.000
1.7					0.000
1.8					0.000
1.9					0.000
2					0.000

Table 1: Rejection probability of the true null hypothesis, $H_0 : \tau = 0$, under the *unbiased* setting with target Type I error control at $\alpha = 0.05$. The Monte Carlo standard error of these probability estimates is bounded above by $\sqrt{.25/5000} \approx 0.007$.

E.2 Power at $\tau = 0.25$

	$\mathbf{\Gamma}$				
$\bar{\mathbf{\Gamma}}$	1	1.1	1.25	1.5	2
1	0.694	0.677	0.677	0.694	0.683
1.05		0.544	0.462	0.391	0.338
1.1		0.528	0.363	0.282	0.188
1.15			0.340	0.202	0.123
1.2			0.322	0.160	0.072
1.25			0.333	0.132	0.046
1.3				0.121	0.031
1.35				0.111	0.024
1.4				0.110	0.019
1.45				0.107	0.017
1.5				0.119	0.015
1.6					0.012
1.7					0.006
1.8					0.009
1.9					0.008
2					0.010

Table 2: Rejection probability of the false null hypothesis, $H_0 : \tau = 0$, under the *unbiased* setting with true alternative hypothesis $H_1 : \tau = 0.25$ and target Type I error control at $\alpha = 0.05$. The Monte Carlo standard error of these probability estimates is bounded above by $\sqrt{.25/5000} \approx 0.007$.

References

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85,** 907–920.

Hsu, J. Y. and Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69,** 803–811.

Rosenbaum, P. R. and Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association* **104,**.