Submitted to the Annals of Applied Statistics

EXTENDED SENSITIVITY ANALYSIS FOR HETEROGENEOUS UNMEASURED CONFOUNDING WITH AN APPLICATION TO SIBLING STUDIES OF RETURNS TO EDUCATION

By Colin B. Fogarty[†,*] and Raiden B. Hasegawa[‡]

Massachusetts Institute of Technology[†] *and University of Pennsylvania*[‡]

The conventional model for assessing insensitivity to hidden bias in paired observational studies constructs a worst-case distribution for treatment assignments subject to bounds on the maximal bias to which any given pair is subjected. In studies where rare cases of extreme hidden bias are suspected, the maximal bias may be substantially larger than the typical bias across pairs, such that a correctly specified bound on the maximal bias would yield an unduly pessimistic perception of the study's robustness to hidden bias. We present an extended sensitivity analysis which allows researchers to simultaneously bound the maximal and typical bias perturbing the pairs under investigation while maintaining the desired Type I error rate. We motivate and illustrate our method with two sibling studies on the impact of schooling on earnings, one containing information of cognitive ability of siblings and the other not. Cognitive ability, clearly influential of both earnings and degree of schooling, is likely similar between members of most sibling pairs yet could, conceivably, vary drastically for some siblings. The method is straightforward to implement, simply requiring the solution to a quadratic program. `R` code is provided in the supplementary materials.

1. Introduction.

1.1. *A motivating example: Returns to schooling.*

Is educational attainment a determining factor for success in the labor market? Initial interest among economists in addressing this question is attributed to the observation in the late 1950s that increases in education levels could account for much of the productivity growth in post-war US (Becker, 2009; Griliches, 1970; Card, 1999). With strong evidence of a positive association between education and earnings in a variety of political and geographic environments but little to no experimental data, a recurring theme in the subsequent pursuit of a causal relationship between education and income is that of the

[*] Both authors contributed equally.

Keywords and phrases: Observational Studies, Causal Inference, Nuisance Parameters, Quadratic Programming, Hidden Bias, Superpopulation Inference

presence of "ability bias" (Card, 1999). After controlling for family background, or considering within-family estimates of the causal effect using sibling or twin studies, can latent differences in ability influence both differences in schooling choice and earnings? A notable twin study by Ashenfelter and Rouse (1998), which we re-examine in this paper, argued cogently, albeit with limited statistical evidence, that identical twins can be regarded as truly identical in all dimensions relevant to schooling choices and future income, including latent ability. In a survey of contemporary economic investigations of returns to education, Card (1999, p.1852) addresses this hypothesis:

> Despite this evidence, and the strong intuitive appeal of the "equal abilities" assumption for identical twins, however, I suspect that observers with a strong a priori belief in the importance of ability bias will remain unconvinced.

Perhaps latent ability is truly identical for many twin pairs but markedly different in a few pairs; what would happen then? That exogeneity is not testable leaves even the most compelling observational evidence susceptible to the warranted, though often non-specific, criticism, "what if bias remains?" Should the totality of evidence assume the absence of hidden bias, the critic need merely suggest the existence of bias to cast doubt upon the posited causal mechanism. It is thus incumbent upon researchers not only to anticipate such criticism, but also to arm themselves with a suitable rejoinder. Rather than arguing for or against the presence of ability bias or any other unobserved confounding factor, in this paper we assess the sensitivity of causal conclusions to departures from truly randomized assignment while allowing for patterns of ability bias that may be highly heterogeneous across sibling pairs.

1.2. *Assessing returns to schooling with sibling comparison designs.* Sibling comparison studies are a special case of stratified designs where natural blocks are formed by family membership. These studies automatically control for genetic, socioeconomic, cultural, and child-rearing characteristics to the extent that they are shared between siblings; however, instability of familial characteristics over time for sibling pairs of different ages and non-shared genetic makeup are among threats to this premise (Donovan and Susser, 2011). Due to their natural and automatic control of stable familial factors, both observed and unobserved, sibling comparison designs have long been a popular tool for studying causal effects in both epidemiological and economic settings; see Griliches (1979) and Donovan and Susser (2011) for surveys of past and current sibling comparison studies in economics and epidemiology, respectively.

Sibling comparison designs have been particularly fruitful in the study of returns to schooling, where genetic and family background are deemed essential to both schooling choices and future income; see for example Hauser, Sheridan and Warren (1999), Stanek, Iacono and McGue (2011), and Ashenfelter and Rouse (1998). Hauser, Sheridan and Warren (1999) study sibling pairs from the Wisconsin Longitudinal Study (WLS), a random sample ($n = 10,317$) of men and women born between 1938 and 1940 who graduated from Wisconsin high schools in 1957. The size of the sample was set to be approximately a third of all Wisconsin high school graduates in 1957. Random siblings of those in the study ($n = 7,928$), born between 1930 and 1948, were also selected and interviewed. The WLS contains a rich set of baseline covariates and endpoints, including physical, cognitive, social, and occupational outcomes collected over nearly 60 years following graduation. Uniquely, the WLS dataset contains intelligence quotient (IQ) scores recorded while a given individual was in high school – a covariate rarely measured in longitudinal cohort studies.

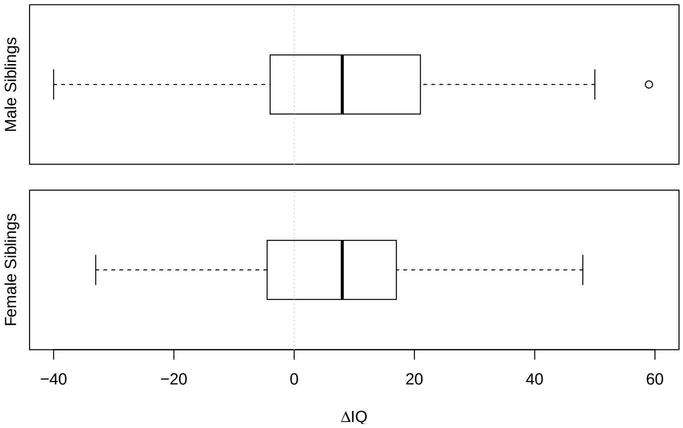
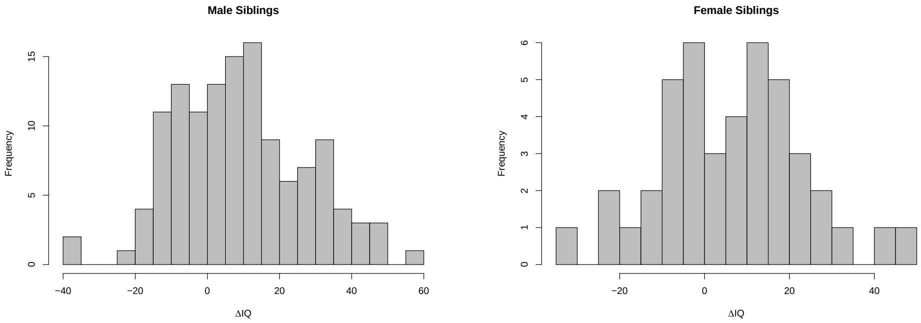


Fig 1. *Boxplots of differences in IQ scores between same-sex siblings where one attended college and the other did not. (top panel): Male same-sex sibling pairs (n* $= 128$*). (bottom panel): Female same-sex sibling pairs (n* $= 43$*).*

In other sibling studies of the returns to schooling, such as that of Ashenfelter and Rouse (1998), baseline intelligence measures such as IQ are not available, making it plausible that the siblings being compared differ in cognitive ability in unobserved ways. Furthermore, the IQ data from the WLS study suggests that, when considering same-sex sibling pairs where one sibling attended college and the other did not ($n = 171$), intellectual ability is not balanced sufficiently by shared genetics alone. The boxplots of differences in IQ between the college-attending siblings and their counterparts in Figure 1 exhibit a prominent shift in the IQ distribution between the two groups for both male and female same-sex sibling pairs. The mean (sd) is 107.1 (14.7) in the college-attending group and 97.4 (14.4) in the high school-only group for male same-sex sibling pairs. In female same-sex sibling pairs, these values are 108.1 (14.0) and 101.4 (14.2) for the college-attending and high school-only attending groups respectively. Details on the construction of the 171 same-sex sibling pairs can be found in §B of the supplemental appendix. An important inclusion criterion was that both siblings were employed when income data was collected.

1.3. *Potential for rare but extreme unmeasured biases.* Despite their analytical strengths and convenient, automatic stratification, sibling comparison designs for estimating causal effects are subject to biases arising from differences in subject-level confounders. For example, latent ability, as measured by IQ, may differ substantially within twin pairs in Ashenfelter and Rouse's twin study. This concern is magnified in sibling studies where discordant within-pair treatment assignment may actually exacerbate differences in covariates that are related to both the intervention and outcome of interest (Frisell et al., 2012). When pairs do not arise naturally, as in paired sibling studies, matching algorithms designed to minimize disparities in observed covariates may be used to construct pairs of "comparable" subjects; see, for example, Hansen and Klopfer (2006) and Stuart (2010) for discussion on various approaches to matching. Matched pairs constructed in this fashion may be comparable along observed covariates, but they are still vulnerable to unmeasured bias arising from differences in covariates not available to the matching algorithm.

While agnostic covariate adjustment within sibling sets as suggested in Rosenbaum (2002a) can help mitigate the impact of discrepancies in observed individual-specific covariates, bias arising from differences in unobserved confounders may remain and imperil the conclusions of the study. An additional inferential step known as a *sensitivity analysis* assesses the robustness of the conclusions of a study to these unmeasured biases. Sensitivity analysis was first introduced by Cornfield et al. (1959) and refined to accommodate continuous outcomes in Rosenbaum (1987). The resulting sensitivity analysis for paired studies considers the worst-case bias to which any pair may be subject and asks whether the study conclusions might change if we assumed that *all* pairs were exposed to the maximal bias in a manner adverse to the desired inference. We refer to this as the *conventional* sensitivity analysis. See Cornfield et al. (1959), Marcus (1997), Imbens (2003), Yu



Odds Ratio	$[1,2)$	$[2,3)$	$[3,6)$	$[6,7)$	$[7,9)$	$[9,10)$
Count	165	4	0	1	0	1

FIG 2. *(left panel): Histogram of between-sibling IQ disparities of same-sex male sibling pairs in the WLS study where one sibling attended college and the other did not ($n = 128$). (right panel): Histogram of between-sibling IQ disparities of same-sex female sibling pairs in the WLS study where one sibling attended college and the other did not ($n = 43$). (bottom panel): Table of the estimated increase in pairwise bias due to IQ disparities between siblings measured as an odds ratio.*

and Gastwirth (2005), Wang and Krieger (2006), Egleston, Scharfstein and MacKenzie (2009), Hosman, Hansen and Holland (2010), Zubizarreta, Cerdá and Rosenbaum (2013), Liu, Kuramoto and Stuart (2013), and VanderWeele and Ding (2017) for additional perspectives on and worked examples of sensitivity analysis.

In many paired studies, sibling or otherwise, hidden biases may strongly influence the results observed for some pairs and more modestly affect others. If the impact of unmeasured confounding were truly heterogeneous in this manner, the conventional sensitivity analysis would be conspicuously conservative. Consider, for example, discrepancies in IQ scores within sibling pairs measured in the WLS where one sibling attended college for at least two years and the other received at most a high school diploma. While existing longitudinal cohort studies rarely contain measures of intelligence (Herd, Carr and Roan, 2014), existing evidence suggests that discrepancies in IQ between sibling pairs are strongly predictive of both differences in educational attainment and differences in future income (Stanek, Iacono and McGue, 2011). In the WLS data, the between-sibling disparity in IQ scores is quite variable across sibling pairs where one sibling attended college and the other did not. The histogram of these college-minus-high school differences is shown in the left panel of Figure 2 for male sibling pairs and the right panel for female sibling pairs. Most IQ differences are modest, but a few sibling pairs have large imbalances (e.g. > 40).

In a sibling study on the returns of schooling where IQ was not recorded, such as Ashenfelter and Rouse's twin study, the maximal bias to which any pair is subject could be materially larger than the typical bias for any sibling pair. Evidence of this pattern's plausibility can be seen in the bottom table of Figure 2. The table shows the distribution of the estimated increase in pairwise bias due to IQ disparities between siblings measured as an odds ratio. The numerator of the odds ratio is the predicted maximum odds that the sibling who reported higher income attended college given the reported disparities in IQ while the denominator corresponds to the maximum odds had both siblings had the same IQ. (the method for estimating these odds ratios is described in §§C-D of the supplemental appendix). While the odds ratio in most pairs is close to one, there are a handful of pairs with odds ratios near 2 and two rare cases of odds ratios greater than 6. As far as the 'typical' or 'expected' pairwise bias is as interpretable a quantity as the worst-case pairwise bias, an *extended* sensitivity analysis of both maximal and expected bias may alleviate concerns that the conventional approach is overly pessimistic while providing a more flexible handling of unobserved bias.

1.4. *Accommodating varying degrees of unmeasured confounding.* We present an extended sensitivity analysis bounding both the maximal and expected bias for paired studies. The concept of expected bias is made precise in §3.1. The theoretical foundations and implementation of the extended sensitivity analysis are developed in §§2- 4, while supporting Type I error control and power simulations are presented in §5. The procedure involves two interpretable parameters, Γ and $\bar{\Gamma} \leq \Gamma$, bounding the maximal and expected bias, respectively. At one extreme, setting $\bar{\Gamma} = \Gamma$ recovers the conventional sensitivity analysis for paired studies proposed in Rosenbaum (1987, §2). At the other, setting $\Gamma = \infty$ for a fixed value of $\bar{\Gamma}$ allows one to bound the average bias while leaving the maximal bias in any given pair unbounded, subsuming the extension presented in Rosenbaum (1987, §4) where the investigator specifies a fraction β of the pairs that satisfy a constraint on the maximal bias and allows the remaining pairs to be exposed to potentially unbounded bias.

The procedure builds in two important ways on recent work by Hasegawa and Small (2017) that established an exact sensitivity analysis for the sample average bias for paired studies with binary outcomes. First, our procedure accommodates continuous outcomes while providing an asymptotically valid testing procedure for sharp null hypotheses for a large class of test statistics. While generalizing to continuous outcomes corrupts properties unique to McNemar's test statistic utilized in Hasegawa and Small (2017), these difficulties are overcome through a new formulation of the optimization problem necessitated by the sensitivity analysis as a quadratic program. Second, our procedure allows the researcher to bound the expected bias at the level of a superpopulation, rather than the average of the bias at the level of the observed study population, if a superpopulation model is deemed appropriate. This facilitates consonance between superpopulation and finite-sample modes of inference to which the researcher is automatically entitled when only bounding the maximal bias. Actualizing this harmony requires the combination of concentration inequalities with the technique presented in Berger and Boos (1994) for yielding valid p-values by maximizing over a confidence set for nuisance parameters.

To demonstrate the practical consequences of our procedure we return in §6 to the motivating example of returns to schooling. Using the availability of IQ measures in the WLS sibling data, we follow Hsu and Small (2013) to estimate the maximal and expected bias under the assumption that inherent cognitive ability is the overwhelming unobserved confounding factor in sibling studies of returns to schooling when IQ measures are not available. We compare standard and extended sensitivity analyses calibrated to these estimates of the sensitivity parameters for Ashenfelter and Rouse's twin study where IQ was not observed.

2. Sensitivity analysis for paired studies.

2.1. *An idealized construction of a paired observational study.* There are I pairs of individuals. In the i^{th} matched pair one individual receives the treatment, $Z_{ij} = 1$, and the other receives the control, $Z_{ij'} = 0$, such that $Z_{i1}+Z_{i2} = 1$ for each i. In practice, the I pairs come into being by minimizing a metric reflective of the within-pair discrepancies between the observed covariates $\mathbf{x}_{ij}$ for the treated and control individuals in a candidate pairing, such that $\mathbf{x}_{i1} \approx \mathbf{x}_{i2}$ in the resulting pairs. As an idealization of this practice, we follow Rosenbaum (1987) and imagine a generative model where the pairs are constructed, for $i = 1, ..., I$, by initially drawing, without replacement from an infinite population of treated individuals (that is, conditional upon $Z = 1$), an individual who has an observed covariate $X_i = x_i$. For each i, we then sample a control individual from the population of controls with the same value for the observed covariate, i.e. given $Z = 0, X = x_i$. Finally, randomly assign indices $(i, 1)$ and $(i, 2)$ to the two individuals in pair i, and let X_i be a random variable denoting the shared value $X_{i1} = X_{i2}$. Despite having a shared value X_i, it may be the case that $U_{i1} \neq U_{i2}$ in any pair i

for some unobserved covariate U. In §3.3, we describe the extent to which the following methodology applies to finite-sample inference in the absence of a superpopulation.

Under the stable unit-treatment value assumption (Rubin, 1980), individual j in matched set i has a potential outcome under treatment, R_{Tij}, and under control, R_{Cij} which does not depend on the treatment received by other individuals in the population. The fundamental problem of causal inference is that vector (R_{Tij}, R_{Cij}) is not jointly observable. Instead, we observe the response $R_{ij} = R_{Tij}Z_{ij} + R_{Cij}(1 - Z_{ij})$, and the observed treated-minus-control paired differences $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$. Lowercase letters denote realizations of random variables. Let $\mathcal{F}_I = \{(x_{ij}, u_{ij}, r_{Tij}, r_{Cij}),\ \ 1 \leq i \leq I,\ \ j = 1, 2\}$ be the values of the potential outcomes, measured covariates, and unmeasured covariates for the $2I$ individuals in the observational study at hand. At times it will be convenient to use boldface for vector-valued constants and random variables after the assignment of indices. For example, $\mathbf{Z}$ represents a vector of length $2I$ with elements $\mathbf{Z} = (Z_{11}, Z_{12}, ..., Z_{I2})$, while $\mathbf{R}_i$ is a vector of length two with elements $\mathbf{R}_i = (R_{i1}, R_{i2})$.

2.2. *Randomization inference under strong ignorability.* The expectation of each paired difference Y_i in the infinite population model of the preceding section is $\mathbb{E}(Y_i \mid X_{ij} = x) = \mathbb{E}(R_{Tij} \mid Z_{ij} = 1, X_{ij} = x) - \mathbb{E}(R_{Cij} \mid Z_{ij} = 0, X_{ij} = x)$ which need not equal $\tau(x) := \mathbb{E}(R_{Tij} - R_{Cij} \mid X_{ij} = x)$ without further assumptions on the relationship between the potential outcomes, the observed covariates, and the treatment indicators. A sufficient condition for equality of these expectations, strong ignorability, entails that for any point x,

$$(R_T, R_C) \perp\!\!\!\perp Z \mid X,\ \ 0 < \mathbb{P}(Z = 1 \mid X = x) < 1. \tag{1}$$

Strong ignorability facilitates far more than equality between $\mathbb{E}(Y_i \mid X_{ij} = x)$ and $\tau(x)$; indeed, it entitles the researcher to use randomization tests akin to those justified in randomized experiments. We consider general hypotheses of the form

$$H_0 :\ \ F_T(R_{Tij}) = F_C(R_{Cij})\ \ \ \forall i, j$$

for pre-specified functions $F_T(\cdot)$ and $F_C(\cdot)$. While this form accommodates flexible models for treatment effects, perhaps the most classical specification is the additive treatment effect model where the treatment effect is constant at τ for all individuals. Under this model $R_{Tij} = R_{Cij} + \tau$, which can be expressed by setting $F_T(R_{Tij}) = R_{Tij} - \tau$ and $F_C(R_{Cij}) = R_{Cij}$. From our data alone we observe $F_{ij} = F_T(R_{Tij})Z_{ij} + F_C(R_{Cij})(1 - Z_{ij})$;

let $\mathbf{F} = [F_{11}, ..., F_{I2}]$. Under H_0, the vectors $\mathbf{F}_C = [F_C(R_{C11}), ..., F_C(R_{CI2})]$ and $\mathbf{F}_T = [F_T(R_{T11}), ..., F_T(R_{TI2})]$ are known to be equal, and hence are entirely specified by the vector of observed responses $\mathbf{R}$.

Let $t(\mathbf{Z}, \mathbf{F})$ be an arbitrary test statistic that is a function of the treatment indicators Z_{ij} and the observed values F_{ij}, and let $\Omega_I = \{\mathbf{z} : z_{i1} + z_{i2} = 1, \quad 1 \leq i \leq I\}$ be the set of 2^I possible assignments of individuals to treatment and control in a paired design. Further let $\mathbf{f}_C$ be the realized value of the random variable $\mathbf{F}_C$. When H_0 holds, $\mathbf{f}_C$ is fully observed. Under the idealized model in §2.1 and under (1), Theorem 1 of Rosenbaum (1984) demonstrates that under the null hypothesis H_0,

$$(2) \qquad \mathbb{P}\{t(\mathbf{Z}, \mathbf{F}) \geq a \mid \mathcal{F}_I, H_0\} = \frac{1}{2^I} \sum_{\mathbf{z} \in \Omega_I} \chi\{t(\mathbf{z}, \mathbf{f}_C) \geq a\},$$

where $\chi\{A\}$ is an indicator that the event A occurred. Importantly, under H_0, the randomization distribution (2) is free of unknown parameters through conditioning on $\mathcal{F}_I$, and hence can be used directly to facilitate inference on H_0.

2.3. *Sensitivity analysis bounding the supremum.* In paired randomized experiments, the physical act of randomization breaks the association between potential outcomes and the intervention and thus justifies both the assumption of strong ignorability and randomization inference through the conditional distribution in (2). Paired observational studies aim to mimic an idealized randomized experiment by creating pairs where individuals are similar on the basis of their observed covariates, X, which would similarly facilitate randomization inference through (2) if strong ignorability held. In observational studies, strong ignorability, and in turn belief in (2), turns a statement of fact into a leap of faith due to the potential presence of unobserved factor U. That treatment assignment is rarely known to be strongly ignorable given observed covariates X alone necessitates a sensitvity analysis which assesses the robustness of a study's conclusions to factors not included in X. A sensitivity analysis operates under the premise that strong ignorability would have been satisfied if an additional pretreatment covariate U had been used in constructing the pairs, that is if for any x and u

$$(3) \qquad (R_T, R_C) \perp\!\!\!\perp Z \mid (X, U), \;\; 0 < \mathbb{P}(Z = 1 \mid X = x, U = u) < 1.$$

A simple model parameterizing the impact of hidden bias presented in Rosenbaum (1987, §2) relates U to the assignment mechanism through a parameter $\Gamma = \exp(\gamma) \geq 1$, which constrains the degree to which U can affect the odds of receiving the intervention through a logit model,

$$\text{logit}(\mathbb{P}(Z=1 \mid X=x, U=u)) = \kappa(x) + \gamma u, \;\; 0 \leq u \leq 1. \tag{4}$$

The bounds on u in (4) may be viewed as a restriction on the scale of the unobserved covariate that is required for the numerical value of γ to have meaning (Rosenbaum, 2002b, Chapter 4). Letting $\pi_i = \mathbb{P}(Z_{i1} = 1 \mid \mathcal{F}_I)$, (3) and (4) then imply $\pi_i = \text{expit}(\gamma(u_{i1} - u_{i2}))$ and $1 - \pi_i = \text{expit}(\gamma(u_{i2} - u_{i1}))$. As a result, the model requires that the bound $\pi_i^* = \max\{\pi_i, 1-\pi_i\} = \text{expit}(\gamma|u_{i1} - u_{i2}|) \leq \Gamma/(1+\Gamma)$ holds uniformly for all i, but imposes no additional constraints on $\boldsymbol{\pi}$, and imposes no constraint on the relationship between the unobserved covariate and the potential outcomes. Theorem 1 of Rosenbaum (1987) illustrates that (3), (4) and the generative model described in §2.1 imply that under a sharp null H_0, the distribution $t(\mathbf{Z}, \mathbf{F})$ given $\mathcal{F}_I$ takes on the modified form

$$\begin{aligned}\mathbb{P}\{t(\mathbf{Z},\mathbf{F}) \geq a \mid \mathcal{F}_I, H_0\} = \sum_{\mathbf{z}\in\Omega_I} \Bigg[& \chi\{t(\mathbf{z}, \mathbf{f}_C) \geq a\} \\ & \times \prod_{i=1}^{I} \text{expit}(\gamma(u_{i1} - u_{i2}))^{z_{i1}} \text{expit}(\gamma(u_{i2} - u_{i1}))^{z_{i2}} \Bigg].\end{aligned} \tag{5}$$

At $\Gamma = 1 \Leftrightarrow \gamma = 0$, (5) recovers (2), hence representing strong ignorability on the basis of X alone. For $\Gamma > 1$, (5) depends on the unknown values of $\mathbf{u}$. A sensitivity analysis proceeds by, for a given value of Γ, finding bounds on (5) by optimizing over the nuisance parameters $\mathbf{u} \in [0,1]^{2I}$ (or equivalently, optimizing over π_i subject to $\pi_i^* \leq \Gamma/(1+\Gamma)$).

We consider test statistics of the form $t(\mathbf{Z}, \mathbf{F}) = \mathbf{Z}^T\mathbf{q}$ for some function $\mathbf{q} = \mathbf{q}(\mathbf{F})$, commonly referred to as sum statistics. Examples of sum statistics in paired observational studies include Wilcoxon's signed rank test and McNemar's test among many others; see Rosenbaum (2002b, Chapter 2) for more on sum statistics. For example, were we to test the null that the treatment effect was constant at zero for all individuals (commonly referred to as Fisher's sharp null hypothesis), then a choice of $q_{ij} = (R_{ij} - R_{ij'})/I = (r_{Cij} - r_{Cij'})/I$ would amount to a choice of the average of the treated-minus-control paired differences in outcomes as the test statistic. In paired studies, arguments parallel to those in Rosenbaum (2002b, Chapter 4) yield that a tight lower bound on (5) is found by setting $u_{i1} - u_{i2} = -\text{sign}(q_{i1} - q_{i2})$ for each pair i, where sign(a) is the sign of the scalar a. Similarly, a tight upper bound on (5) is found by setting $u_{i1} - u_{i2} = \text{sign}(q_{i1} - q_{i2})$ for each i. As a further illustration, if one uses the difference in means as the test statistic, the lower (upper) bound is attained through a perfect negative (positive) correlation between the differences in unmeasured covariates and the signs of the treated-minus-control paired differences.

3. An extended sensitivity analysis.

3.1. *Average-case unmeasured confounding in paired studies.* In §§1.1-1.2, it was argued that large discrepancies in IQ within pairs of siblings, while likely uncommon, would have a large impact on both likelihood of attaining more than a high school degree and on an individual's expected earnings. Were this the only unmeasured confounder, we would then expect most of the values for $\boldsymbol{\pi}^*$, the maximal probabilities of assignment to treatment within a pair, to not deviate substantially from 0.5, while a few pairs would likely have values for π_i^* substantially larger than 0.5. The conventional model for a sensitivity analysis presented in §2.3 bounds π_i^* by $\Gamma/(1+\Gamma)$ for all pairs. Despite typical discrepancies in IQ likely being small, the smallest value of Γ for which (4) and (5) hold would be large due to the small number of extremely biased pairs. When utilized in its original form, the sensitivity analysis in §2.3 may then paint an overly pessimistic picture of the robustness of the study's findings to unmeasured confounding under this belief, as it cannot account for the 'typical' level of unmeasured confounding being different from the worst-case level.

We consider an extension of the conventional sensitivity analysis summarized in §2.3 involving two sensitivity parameters, Γ and $\bar{\Gamma}$. The first, Γ, plays a role identical to that of Γ in the conventional sensitivity analysis by bounding the supremum of the biased assignment probabilities within a pair. Explicitly, we bound the probabilities of receiving the intervention through a logit form,

$$\text{logit}(\mathbb{P}(Z = 1 \mid X, U)) = \kappa(X) + \gamma U, \;\; 0 \leq U \leq 1. \tag{6}$$

That $0 \leq U \leq 1$ trivially implies that for any pair i

$$1/2 \leq \text{expit}(\gamma|U_{i1} - U_{i2}|) \leq \frac{\Gamma}{1+\Gamma}. \tag{7}$$

Under (3) and the setup of §2.1, (6) yields that $\Pi_i^* = \max\{\Pi_i, 1 - \Pi_i\} = \text{expit}(\gamma|U_{i1} - U_{i2}|) \leq \Gamma/(1+\Gamma)$, where $\Pi_i = \mathbb{P}(Z_{i1} = 1 \mid X_i, \mathbf{U}_i, \mathbf{R}_{Ti}, \mathbf{R}_{Ci}) = \mathbb{P}(Z_{i1} = 1 \mid X_i, \mathbf{U}_i)$. We capitalize U_{ij} and Π_i^* to emphasize that they themselves are random variables with respect to the superpopulation model in §2.1, which would become deterministic by conditioning in $\mathcal{F}_I$.

The second sensitivity parameter, $\bar{\Gamma}$, serves to bound the *expectation* of the biased probabilities. We define $\mu_{\pi^*} = \mathbb{E}[\Pi_i^*] = \mathbb{E}[\text{expit}(\gamma|U_{i1} - U_{i2}|)]$, and impose that for some value $\bar{\Gamma}$ such that $1 \leq \bar{\Gamma} \leq \Gamma$,

$$1/2 \leq \mu_{\pi^*} \leq \frac{\bar{\Gamma}}{1+\bar{\Gamma}}. \tag{8}$$

Again, this expectation is taken over repeated samples in the idealized setting in §2.1, within which the fixed but unknown values π_i^* in our observational study can be modeled as *iid* realizations of the random variables Π_i^*. As with the conventional sensitivity analysis, our model makes no assumption about the relationship between the unobserved covariates and the potential outcomes.

Like the conventional sensitivity analysis, our extended procedure solves an optimization problem over a set of nuisance parameters $\boldsymbol{\pi}$ that satisfy the typical and maximal bias bounds specified in (7) and (8). Although the population-level bound $\Pi_i^* \leq \Gamma/(1+\Gamma)$ implies the corresponding sample level bound $\pi_i^* \leq \Gamma/(1+\Gamma)$, the same cannot be said about the corresponding bound on μ_π^*. If $\mu_\pi^* \leq \bar{\Gamma}/(1+\bar{\Gamma})$, a sample realization $\bar{\pi}^*$ arbitrarily close to $\Gamma/(1+\Gamma)$ is still possible, however unlikely. To address this, we translate the bound on μ_π^* to a stochastic bound on $\bar{\Pi}^*$.

In order to construct this stochastic bound, we consider properties of the random variable Π_i^* across draws from the idealized setting in §2.1. From (7) and (8), we have that for all i Π_i^* is bounded above by $\Gamma/(1+\Gamma)$, bounded below by $1/2$, and has expectation μ_{π^*} which is itself bounded above by $\bar{\Gamma}/(1+\bar{\Gamma})$. The Bhatia-Davis inequality (Bhatia and Davis, 2000) provides the variance upper bound

$$\text{var}(\Pi_i^*) \leq (\Gamma/(1+\Gamma) - \mu_{\pi^*})\,(\mu_{\pi^*} - 1/2) = \nu^2(\Gamma, \mu_{\pi^*}).$$

As the Π_i^* can further be modeled as *iid* random variables under the setting being considered, defining $\bar{\Pi}^* = I^{-1}\sum_{i=1}^I \Pi_i^*$, it follows that

$$\mathbb{E}[\bar{\Pi}^*] = \mu_{\pi^*}, \;\; \text{var}(\bar{\Pi}^*) \leq \nu^2(\Gamma, \mu_{\pi^*})/I.$$

If $\text{var}(\Pi_i^*) > 0$ the Central Limit Theorem applies to $\bar{\Pi}^*$, indicating that for any $0 < \beta \leq 0.5$

$$\lim_{I\to\infty} \mathbb{P}(\bar{\Pi}^* \in \mathcal{C}_\beta(\Gamma, \mu_{\pi^*})) \geq 1 - \beta, \tag{9}$$

where, because $\bar{\Pi}^* \geq 1/2$ by definition of Π_i^*

$$\mathcal{C}_\beta(\Gamma, \mu_{\pi^*}) = \left[1/2, \mu_{\pi^*} + I^{-1/2}\Phi^{-1}(1-\beta)\nu(\Gamma, \mu_{\pi^*})\right], \tag{10}$$

and $\Phi^{-1}(p)$ is the p-quantile of the standard normal distribution. Further, (9) is trivially true if $\mathrm{var}(\Pi_i^*) = 0$, as the upper bound of $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$ is no smaller than μ_{π^*} when $\beta \leq 0.5$. That is, knowledge of μ_{π^*} alone enables the construction of asymptotically valid uncertainty sets for $\bar{\Pi}^*$.

3.2. *Sensitivity analysis bounding the supremum and expectation.* Conditional upon $\mathcal{F}_I$, attention returns to the unmeasured confounders for the individuals in our study population, $\mathbf{u}$, and the corresponding assignment probabilities $\boldsymbol{\pi}$. For any value of $\mathbf{u}$ and value for Γ, we have that

$$\mathbb{P}\{t(\mathbf{Z}, \mathbf{F}) \geq a \mid \mathcal{F}_I, H_0\} = \sum_{\mathbf{z} \in \Omega_I} \chi\{t(\mathbf{z}, \mathbf{f}_C) \geq a\} \prod_{i=1}^{I} \pi_i^{z_{i1}}(1-\pi_i)^{z_{i2}}, \tag{11}$$

where $\pi_i = \mathrm{expit}(\gamma(u_{i1} - u_{i2}))$. As the shared notation seeks to emphasize, (11) is precisely the null distribution utilized in (5). Here as well as in (5), the unmeasured confounders $\mathbf{u}$, and hence the conditional assignment probabilities $\boldsymbol{\pi}$, are unknown constants, hindering the desired inference through their presence as nuisance parameters. The approach taken in §2.3 was to maximize or minimize (11) over $\mathbf{u} \in [0,1]^{2I}$ for a given value Γ, or equivalently over $\pi_i^* \leq \Gamma/(1+\Gamma)$. In what follows, we replace this optimization with one over a subset informed by both Γ and $\bar{\Gamma}$ while providing an asymptotically valid level-α test.

Suppose without loss of generality that we are considering a one-sided, greater than alternative. Let $\mathcal{P}_\beta(\Gamma, \mu_{\pi^*}) = \{\boldsymbol{\pi} : \bar{\pi}^* \in \mathcal{C}_\beta(\Gamma, \mu_{\pi^*}),\quad \pi_i^* \leq \Gamma/(1+\Gamma),\quad 1 \leq i \leq I\}$, and consider the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\pi}, \mu_{\pi^*}}{\text{maximize}} \quad & p(\boldsymbol{\pi}, \mu_{\pi^*}) = \sum_{\mathbf{z} \in \Omega_I} \chi\{t(\mathbf{z}, \mathbf{f}_C) \geq t(\mathbf{Z}, \mathbf{F})\} \prod_{i=1}^{I} \pi_i^{z_{i1}}(1-\pi_i)^{z_{i2}} \\ \text{subject to} \quad & \boldsymbol{\pi} \in \mathcal{P}_\beta(\Gamma, \mu_{\pi^*}) \\ & \mu_{\pi^*} \leq \bar{\Gamma}/(1+\bar{\Gamma}). \end{aligned} \tag{12}$$

Let $\mathcal{U}_\beta(\Gamma, \bar{\Gamma})$ be the set of feasible solutions to (12). Let $\boldsymbol{\pi}_{\sup,\beta}$ and $\mu_{\sup,\beta}$ be the arg max of (12), such that $p(\boldsymbol{\pi}_{\sup,\beta}, \mu_{\sup,\beta})$ is the tail probability at the solution to (12). If $\bar{\Gamma} < \Gamma$, let $p_\beta = p(\boldsymbol{\pi}_{\sup,\beta}, \mu_{\sup,\beta}) + \beta$; otherwise, let $p_\beta = p(\boldsymbol{\pi}_{\sup,\beta}, \mu_{\sup,\beta})$.

PROPOSITION 1. *Suppose we sample I pairs from an infinite population through the procedure in §2.1, that treatment assignment is strongly ignorable given (X, U), and that (7) and (8) hold at Γ and $\bar{\Gamma} \leq \Gamma$ respectively. Then,* *if H_0 is true, for $0 < \beta \leq 0.5$,*

$$\lim_{I\to\infty} \mathbb{P}(p_\beta \leq \alpha \mid H_0) \leq \alpha$$

That is, p_β is an asymptotically valid p-value for an extended sensitivity analysis testing H_0 with parameters $(\Gamma, \bar{\Gamma})$.

Proof. We first prove the result for $\bar{\Gamma} < \Gamma$. The proof is similar to that of Lemma 1 in Berger and Boos (1994), differing primarily in that the nuisance parameters given $\mathcal{F}_I$, $\boldsymbol{\pi}$, are themselves realizations of random variables in the setting of §2.1. Suppose the null hypothesis is true, and let μ_0 be the true value for μ_{π^*}. Further, for any set $\mathcal{F}_I$ let $\boldsymbol{\pi}_0$ be the true value of $\boldsymbol{\pi}$. and let $p(\boldsymbol{\pi}_0, \mu_0)$ be the value of (11) evaluated at $\boldsymbol{\pi}_0$ and μ_0.

$$\begin{aligned}
\mathbb{P}(p_\beta \leq \alpha) &= \mathbb{E}[\mathbb{P}(p_\beta \leq \alpha, \bar{\pi}_0^* \in \mathcal{C}_\beta(\Gamma, \mu_0) \mid \mathcal{F}_I)] + \mathbb{E}[\mathbb{P}(p_\beta \leq \alpha, \bar{\pi}_0^* \notin \mathcal{C}_\beta(\Gamma, \mu_0) \mid \mathcal{F}_I)] \\
&\leq \mathbb{E}[\mathbb{P}(p(\boldsymbol{\pi}_0, \mu_0) + \beta \leq \alpha \mid \mathcal{F}_I)] + \mathbb{E}[\mathbb{P}(\bar{\pi}_0^* \notin \mathcal{C}_\beta(\Gamma, \mu_0) \mid \mathcal{F}_I)] \\
&= \mathbb{E}[\mathbb{P}(p(\boldsymbol{\pi}_0, \mu_0) \leq \alpha - \beta \mid \mathcal{F}_I)] + \mathbb{P}(\bar{\Pi}^* \notin \mathcal{C}_\beta(\Gamma, \mu_0))
\end{aligned}$$

The second line follows from $p(\boldsymbol{\pi}_0, \mu_0) \leq \sup_{\boldsymbol{\pi}\in\mathcal{P}_\beta(\Gamma,\mu_0)} p(\boldsymbol{\pi}, \mu_0) \leq p_\beta - \beta$ if $\bar{\pi}_0^* \in \mathcal{C}_\beta(\Gamma, \mu_0)$. By validity of (11) at $\boldsymbol{\pi}_0$ given $\mathcal{F}_I$, the first term in the third line is less than or equal to $\alpha - \beta$, while (9) illustrates that $\lim_{I\to\infty}\mathbb{P}(\bar{\Pi}^* \notin \mathcal{C}_\beta(\Gamma, \mu_0)) \leq \beta$ for $0 < \beta \leq 0.5$, proving the result for $\bar{\Gamma} < \Gamma$.

If $\bar{\Gamma} = \Gamma$, a solution $\boldsymbol{\pi} \in \mathcal{U}(\Gamma, \Gamma)$ is $\pi_i = \Gamma/(1+\Gamma)$ if $(q_{i1} > q_{i2})$ and $\pi_i = 1/(1+\Gamma)$ otherwise, which recovers the sensitivity analysis of §2.3. Call this solution $\boldsymbol{\pi}_\Gamma$. By arguments in Rosenbaum (2002b, Chapter 4), this solution yields a tight upper bound for the probability in (11) under the constraint that $\pi_i^* \leq \Gamma/(1+\Gamma)$. Hence, $p(\boldsymbol{\pi}_{\sup,\beta}, \mu_{\sup,\beta}) = p(\boldsymbol{\pi}_\Gamma, \Gamma/(1+\Gamma))$ for any β. At $\bar{\Gamma} = \Gamma$, we simply employ the conventional sensitivity analysis which produces valid p-values without an additive increase by β. □

Prior to conducting an extended sensitivity analysis, the practitioner needs to choose a value for β. A compromise must be made, as β acts as a lower bound on the p-value reported by the extended sensitivity analysis but larger values of β correspond to tighter constraints on $\bar{\pi}^*$. Accordingly, we recommend that β be chosen to be smaller than the precision with which p-values are typically reported, but not by much. This recommendation is similar to the guidance given in Berger and Boos (1994).

p_β yields an asymptotically valid p-value for an extended sensitivity analysis with parameters $(\Gamma, \bar{\Gamma})$ because the uncertainty set $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$ defined in (10) utilizes the Central Limit Theorem. As our random variables Π_i^* are bounded, we are entitled to certain distribution-free uncertainty sets based on concentration inequalities which have the desired coverage for all sample sizes I; see §A of the supplemental appendix for two approaches using Hoeffding's inequality and Bennett's inequality. These sets, used in place of $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$ when constructing $\mathcal{P}_\beta(\Gamma, \mu_{\pi^*})$, would provide valid p-values for the extended sensitivity analysis through the solution of (12) for all values of I. Unfortunately, exact computation of p_β through (12) is itself generally intractable, with the additional constraints imposed on the value of $\bar{\pi}$ destroying the properties of the optimization problem solved by the conventional sensitivity analysis which facilitate an exact solution. In §4, we provide an implementation of our sensitivity analysis valid in large samples by approximating (11) with an appropriate normal distribution, justified under mild conditions. As we employ a normal approximation through our implementation, already implying a large-sample regime, we proceed illustrating the method using the asymptotically valid uncertainty set $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$.

3.3. *On extended sensitivity analyses for observed study populations.* Under the superpopulation model described in §2.1, Π_i^* is itself a random variable with expectation $\mathbb{E}[\bar{\Pi}^*]$. In randomized experiments and observational studies, the assumption that the individuals in the study arose as a sample from some larger target population is often specious. Such an assumption is not required for inferential statements, as the act of random assignment to intervention itself can form the basis for probabilistic statements and hypothesis tests, endowing randomized experiments with what Fisher referred to as a "reasoned basis for inference" (Fisher, 1935). Rosenbaum (1999) further argues that the most compelling observational studies are not those which are representative of a larger population, but rather those arrived upon through an active choice of the conditions of observation, seeking the "rare circumstances in which tangible evidence may be obtained to distinguish treatment effects from the most plausible biases" (Rosenbaum, 1999, p. 259).

As (5) indicates through conditioning on the study population, $\mathcal{F}_I$, the classical sensitivity analysis in §2.3 yields a null distribution for finite-sample inference whose nuisance parameters are the unknown assignment probabilities $\boldsymbol{\pi}$ for the individuals in the study at hand. The parameter Γ, which originally served to bound the supremum of the random variables Π_i^*, also bounds the supremum of the observed values π_i^*. This yields harmony between inference conducted for the finite study population and inference assuming an infinite population into existence when interest is in the hypothesis H_0. Inference given $\mathcal{F}_I$ is valid on its own, but if a superpopulation model is deemed appropriate, inference given $\mathcal{F}_I$ yields valid unconditional inference within that framework.

The motivation for formulating the extended sensitivity analysis with explicit reference to a superpopulation is that while bounds on the supremum of a random variable bound the random variable's realizations, bounds on the expectation of a random variable do not afford bounds in the sample average. The idealized model is used to formulate probabilistic bounds for the sample average $\bar{\Pi}^*$, which then entitle us to a further bound on the average of the realized vector $\boldsymbol{\pi}^*$. Proposition 1 indicates that the price to be paid for implementing this bound is the addition of an extra β term to the p-value, necessitated by the view of $\boldsymbol{\pi}^*$ as a realization of a random variable. Should a superpopulation model be deemed unreasonable, our model could instead be interpreted as placing a bound on the sample average of the parameters $\boldsymbol{\pi}^*$, $\bar{\pi}^*$, in the particular observational study being analyzed. This interpretation eliminates the need for both the uncertainty set $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$ and the increase in the p-value by β, and an option to consider study population inference is available within our `R` function. In our particular case study we proceed using superpopulation bounds, as in calibrating the sensitivity parameters in one observational study by means of another one must assume comparability of biases in the two studies.

3.4. *A special case: Binary outcomes.* Although exact computation of p_β is generally intractable, in one special but common setting it is not. When the outcomes being studied are binary and $t(\mathbf{Z}, \mathbf{F})$ is chosen to be McNemar's test statistic, computing p_β exactly under Fisher's sharp null $H_0 : \ R_{Tij} = R_{Cij}$ becomes a straightforward exercise. Recall that McNemar's test statistic counts the number of pairs where the subject under treatment has a positive outcome and the control subject does not; that is, $t(\mathbf{Z}, \mathbf{F}) = \sum_{i=1}^{I}(Z_{i1} - Z_{i2})(R_{Ci1} - R_{Ci2})/2 + 1/2$ when Fisher's sharp null is true. Since pairs that are not discordant in treatment and outcome do not contribute to McNemar's statistic it is natural to distinguish pairs that are discordant in outcome and those that are not. Let the first I_d pairs be the discordant pairs and the last I_c be the concordant pairs so that $I = I_d + I_c$. Furthermore, let the first unit of each discordant pair be the unit with positive outcome, that is $R_{i1} = 1$ for $i = 1, \dots, I_d$.

For the special case of McNemar's test, let μ_m be the value of $\mu_{\pi^*} \leq \bar{\Gamma}/(1 + \bar{\Gamma})$ that maximizes the upper bound of $C_\beta(\Gamma, \mu_{\pi^*})$ and let $\bar{\pi}_m$ be the maximized upper bound. Define $\bar{\pi}_c = 1/2$,

$$\bar{\pi}_d = \min\left\{(I\bar{\pi}_m - I_c\bar{\pi}_c)/I_d, \Gamma/(1 + \Gamma)\right\},$$

and $\boldsymbol{\pi}_m = ([\bar{\pi}_d \cdot \mathbf{1}_d, \bar{\pi}_c \cdot \mathbf{1}_c])$, where $\mathbf{1}_k$ is a vector of I_k ones. $(\boldsymbol{\pi}_m, \mu_m)$ is then a feasible solution to (12) that is designed to put as much bias on the discordant pairs as is allowed by the constraints of the optimization problem. Furthermore, since the concordant pairs do not contribute to the test statistic we have that $p(\boldsymbol{\pi}_m, \mu_m) = \mathbb{P}(B(I_d, \bar{\pi}_d) \geq t(\mathbf{Z}, \mathbf{F}))$, where $B(I_d, \bar{\pi}_d)$ is a Binomial random variable with success probability $\bar{\pi}_d$ and I_d trials. Now, let $p_\beta = p(\boldsymbol{\pi}_m, \mu_m) + \beta$ when $\bar{\Gamma} < \Gamma$ and let $p_\beta = p(\boldsymbol{\pi}_\Gamma, \Gamma/(1+\Gamma))$ otherwise. In the following proposition we show that, in this special setting, an exact solution to (12) simply requires computing this Binomial tail probability.

PROPOSITION 2. *Consider a test of $H_0 : R_{Tij} = R_{Cij}$ with binary outcomes, and let $t(\mathbf{Z}, \mathbf{F})$ be McNemar's test statistic. Further, let $C_\beta(\Gamma, \mu_{\pi^*})$ be an exact, distribution-free $1 - \beta$ uncertainty set. Then under the same conditions as Proposition 1,*

$$\mathbb{P}(p_\beta \leq \alpha \mid H_0) \leq \alpha$$

for any I if $t(\mathbf{Z}, \mathbf{F}) \geq I_d \bar{\pi}_d$. In other words, for any value of I, computing a valid p-value for an extended sensitivity analysis testing H_0 with parameters $(\Gamma, \bar{\Gamma})$ reduces to computing the Binomial tail probability $\mathbb{P}(B(I_d, \bar{\pi}_d) \geq t(\mathbf{Z}, \mathbf{F}))$.

PROOF. When $\bar{\Gamma} = \Gamma$, the proof follows immediately from the proof of this case in Proposition 1. Hence, we restrict our attention to the case when $\bar{\Gamma} < \Gamma$. As noted in §3.2, if we replace $C_\beta(\Gamma, \mu_{\pi^*})$ with a distribution-free uncertainty set the optimal solution to (12) yields a valid p-value for an extended sensitivity analysis for all values of I. All that remains to be shown is that $(\boldsymbol{\pi}_m, \mu_m)$ is the argmax of (12).

Without loss of generality, suppose once again that the first subject of each discordant pair is the unit with a positive outcome, $R_{i1} = 1$ for all $i = 1, \ldots, I_d$. Let $(\boldsymbol{\pi}', \mu')$ be a feasible solution of (12) and define $\bar{\pi}'_d$ and $\bar{\pi}'_c$ to be the sample average of the maximal assignment probabilities for the discordant and concordant pairs, respectively. $([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu')$ is clearly also a feasible solution. Then, Theorem 1 in Hasegawa and Small (2017) implies that $p([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu') \geq p(\boldsymbol{\pi}', \mu')$ when $t(\mathbf{Z}, \mathbf{F}) \geq I_d \cdot \bar{\pi}'_d$. Hence, we need only consider feasible solutions of the form $([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu')$. An elementary fact about Binomial random variables is that $B(I_d, p_1)$ stochastically dominates $B(I_d, p_2)$ when $p_1 \geq p_2$. By construction, $(\boldsymbol{\pi}_m, \mu_m)$ yields a feasible solution such that $\bar{\pi}_d \geq \bar{\pi}'_d$ for all feasible solutions of the form $([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu')$. Consequently, $p(\boldsymbol{\pi}_m, \mu_m) \geq p([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu') \geq p(\boldsymbol{\pi}', \mu')$ for all feasible solutions $(\boldsymbol{\pi}', \mu')$ which proves the result for $\bar{\Gamma} < \Gamma$. □

For McNemar's test, the extended sensitivity analysis exhibits an interesting behavior when $\bar{\pi}_d = \Gamma/(1+\Gamma)$: the procedure returns a p-value equal to the p-value returned by the conventional sensitivity analysis at Γ *plus* the extra β term. We still pay the cost of specifying a bound on $\mathbb{E}[\Pi_i^*]$ but do not receive the benefit of a tighter constraint on the realization of $\boldsymbol{\pi}^*$ for discordant pairs. What, exactly, explains this phenomenon? A plausible scenario that may give rise to this behavior is when $I_c >> I_d$, i.e. there are many concordant pairs in the sample of I pairs. In throwing out concordant pairs when using McNemar's statistic, the uncertainty set for $\bar{\Pi}^*$, the average of Π_i^* over all pairs, tells us relatively little about the realized average $\bar{\pi}_d^*$ over discordant pairs, reflecting the cost of bounding the marginal expectation $\mathbb{E}[\Pi_i^*]$ instead of the conditional expectation $\mathbb{E}[\Pi_i^* \mid \mathbf{R}_{Ti}, \mathbf{R}_{Ci}]$.

Although this behavior indicates that the extended sensitivity analysis is, in some sense, suboptimal compared to the conventional sensitivity analysis when $I_c >> I_d$, the practical implications are mostly negligible as β is chosen to be smaller than the precision with which p-values are generally reported. Furthermore, given a choice of Γ and conditional on (I_d, I_c), we can a priori determine the value of $\bar{\Gamma}$ above which the conventional analysis is superior to the extended analysis. Because (I_d, I_c) are known conditional on $\mathcal{F}_I$, we are not at risk of using the data twice – once to choose the best test and once to perform that test. Consequently, the resulting sensitivity analyses will still have the appropriate level.

4. Implementation through quadratic programming.

The test statistics described in §2.3 can be represented as the sum of I independent random variables, $\mathbf{Z}^T\mathbf{q} = \sum_{i=1}^I T_i$, where $T_i = (q_{i1}+q_{i2})/2 + (Z_{i1} - Z_{i2})(q_{i1}-q_{i2})/2$. This suggests that, under mild regularity conditions, a central limit theorem would be applicable to the distribution of $\mathbf{Z}^T\mathbf{q}$ for any value of $\boldsymbol{\pi}$ in (11) for almost every sample path $\mathcal{F}_I$. One sufficient condition proposed in the special central limit theorem of Hájek, Šidák and Sen (1999, §6.1.2) is that, almost surely,

$$\frac{\sum_{i=1}^I (q_{i1}-q_{i2})^2}{\max\limits_{1\leq i\leq I}(q_{i1}-q_{i2})^2} \to \infty,$$

which requires that no one term $(q_{i1}-q_{i2})^2$ dominates the sum as the number of pairs increases. (An aside: the central limit theorem in Hájek, Šidák and Sen (1999, §6.1.2) as originally stated applies to sums of the form $\sum_{i=1}^I a_i X_i$ where X_i are *iid* random variables; however, the proof can readily be extended to settings where $I\sigma^2 \leq \sum_{i=1}^I \text{var}(X_i) \leq Ic\sigma^2$ for $c > 1$ while dropping the requirement of identical distribution, which encompasses the setting of our extended sensitivity analysis). Under a normal approximation, the problem of finding the worst-case p-value is equivalent to finding the worst-case deviate.

Recall that a sensitivity analysis is typically conducted only if the null hypothesis is rejected under the assumption of no unmeasured confounding ($\Gamma = \bar{\Gamma} = 1$), and then proceeds by iteratively increasing the sensitivity parameters until the test fails to reject. Having proceeded to sensitivity analysis only after rejecting the null under no unmeasured confounding, even with one-sided alternatives we can safely consider rejection or failure to reject for sequentially larger values of Γ and $\bar{\Gamma}$ based on the minimal squared deviate, an objective function which is preferred for computational reasons alluded to below. Recalling that under (11) we condition on $\mathcal{F}_I$ and hence treat the vector $\mathbf{q}$ as fixed, minimizing the squared deviate can be expressed as an optimization problem over the unknown probabilities $\boldsymbol{\pi}$ as

$$\min_{\boldsymbol{\pi}\in\mathcal{U}_\beta(\Gamma,\bar{\Gamma})} \frac{(t - \mathbb{E}_{\boldsymbol{\pi}}[\mathbf{Z}^T\mathbf{q} \mid \mathcal{F}_I])^2}{\text{var}_{\boldsymbol{\pi}}(\mathbf{Z}^T\mathbf{q} \mid \mathcal{F}_I)}, \tag{13}$$

where t is the observed value of the statistic $t(\mathbf{Z}, \mathbf{F})$, and the expectation and variance are for the test statistic $t(\mathbf{Z}, \mathbf{F})$ under the randomization distribution (11) for a given vector $\boldsymbol{\pi}$. Under a normal approximation for $t(\mathbf{Z}, \mathbf{F})$, the squared deviate follows a χ^2_1 distribution. By the argument of the previous section, we then reject the null at level α if (13) is greater than or equal to $G^{-1}(1-2(\alpha-\beta))$ for one-sided alternatives or $G^{-1}(1-(\alpha-\beta))$ for two-sided alternatives, where $G^{-1}(p)$ is the p quantile of a χ^2_1 distribution.

The expectation and variance of the contribution of T_i can be expressed as a function of the unknown vector $\boldsymbol{\pi}$ as

$$\mathbb{E}_{\boldsymbol{\pi}}[T_i \mid \mathcal{F}_I] = \mathbf{q}_i^T\boldsymbol{\pi}_i \tag{14}$$

$$\begin{aligned}\text{var}_{\boldsymbol{\pi}}(T_i \mid \mathcal{F}_I) &= \pi_i(1-\pi_i)(q_{i1}-q_{i2})^2 \\ &= (\mathbf{q}_i^2)^T\boldsymbol{\pi}_i - (\mathbf{q}_i^T\boldsymbol{\pi}_i)^2\end{aligned} \tag{15}$$

where $\boldsymbol{\pi}_i$ and $\mathbf{q}_i$ are vectors of length two with elements $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2})$ and $\mathbf{q}_i = (q_{i1}, q_{i2})$, respectively. Suppose without loss of generality that we are considering a one-sided, greater than alternative and that we rejected the null at $(\Gamma, \bar{\Gamma}) = (1, 1)$, which implies that $t \geq (2I)^{-1}\sum_{i=1}^I\sum_{j=1}^2 q_{ij}$ (i.e. that the observed value of t exceeded its null expectation). Sort each vector $\mathbf{q}_i$ in descending order such that $q_{i1} \geq q_{i2}$. Then, $\text{var}_{\boldsymbol{\pi}}(T_i \mid \mathcal{F}_I) = \text{var}_{\boldsymbol{\pi}^*}(T_i \mid \mathcal{F}_I)$ from (15), while from (14) $\mathbb{E}_{\boldsymbol{\pi}}[T_i \mid \mathcal{F}_I] \leq \mathbb{E}_{\boldsymbol{\pi}^*}[T_i \mid \mathcal{F}_I] = \mathbf{q}_i^T\boldsymbol{\pi}_i^*$ and $(q_{i1} + q_{i2})/2 \leq \mathbb{E}_{\boldsymbol{\pi}^*}[T_i \mid \mathcal{F}_I]$. Hence, any feasible solution $\boldsymbol{\pi}'$ to (13) has an objective value that is no smaller than that of $(\boldsymbol{\pi}^*)'$, as the variance will be the same while, recalling the iterative nature of a sensitivity analysis, the distance $(t - \mathbb{E}_{(\boldsymbol{\pi}^*)'}[\mathbf{Z}^T\mathbf{q}' \mid \mathcal{F}_I])^2$ will be smaller than $(t - \mathbb{E}_{\boldsymbol{\pi}'}[\mathbf{Z}^T\mathbf{q}' \mid \mathcal{F}_I])^2$. Maintaining this ordering of the vectors $\mathbf{q}_i$, we can express our optimization problem as a function of the maximal probabilities $\boldsymbol{\pi}_i^*$.

For any candidate $\boldsymbol{\pi}^*$, we reject under a normal approximation with a one-sided, greater than alternative at level $\alpha - \beta$ if the corresponding squared deviate exceeds its critical value, $G^{-1}(1 - 2(\alpha - \beta))$ i.e. if $\zeta(\boldsymbol{\pi}^*, \alpha - \beta) = (t - \mathbb{E}_{\boldsymbol{\pi}^*}[\mathbf{Z}^T\mathbf{q} \mid \mathcal{F}_I])^2 - G^{-1}(1 - 2(\alpha - \beta))\text{var}_{\boldsymbol{\pi}^*}(\mathbf{Z}^T\mathbf{q} \mid \mathcal{F}_I) \geq 0$. We write $\zeta(\boldsymbol{\pi}^*, \alpha - \beta)$ explicitly as a function of $\boldsymbol{\pi}^*$ as

$$\zeta(\boldsymbol{\pi}^*, \alpha - \beta) = (t - \mathbf{q}^T\boldsymbol{\pi}^*)^2 - G^{-1}(1 - 2(\alpha - \beta))\sum_{i=1}^{I}\left((\mathbf{q}_i^2)^T\boldsymbol{\pi}_i^* - (\mathbf{q}_i^T\boldsymbol{\pi}_i^*)^2\right)$$

If we find that $\zeta(\boldsymbol{\pi}^*, \alpha - \beta) \geq 0$ for all feasible $\boldsymbol{\pi}^* \in \mathcal{U}_\beta(\Gamma, \bar{\Gamma})$, we can reject the null while asymptotically controlling the size of the extended sensitivity analysis with parameters $(\Gamma, \bar{\Gamma})$ at α. The function $\zeta(\boldsymbol{\pi}^*, \alpha - \beta)$ is convex and quadratic in $\boldsymbol{\pi}^*$. Meanwhile, we explicitly write the constraints determining membership in $\mathcal{U}_\beta(\Gamma, \bar{\Gamma})$ as

$$1/2 \leq \pi_i^* \leq \Gamma/(1 + \Gamma), \quad 1 \leq i \leq I \tag{16}$$

$$I^{-1}\sum_{i=1}^{I}\pi_i^* \leq \mu_{\pi^*} + I^{-1/2}\Phi^{-1}(1 - \beta)\left\{(\Gamma/(1 + \Gamma) - \mu_{\pi^*})(\mu_{\pi^*} - 1/2)\right\}^{1/2} \tag{17}$$

$$\mu_{\pi^*} \leq \bar{\Gamma}/(1 + \bar{\Gamma}). \tag{18}$$

For a fixed value of $\mu_{\pi^*} \leq \bar{\Gamma}/(1+\bar{\Gamma})$ the constraints are linear in the unknown maximal probabilites π_i^*. Hence, for fixed μ_{π^*}, the problem $\min_{\boldsymbol{\pi}^*}\zeta(\pi^*, \alpha - \beta)$ subject to (16) and (17) can be written as a quadratic program. With a one-sided alternative, an asymptotically level-α extended sensitivity analysis with parameters $(\bar{\Gamma}, \Gamma)$ simply requires checking whether the solution to that quadratic program is greater than or equal to zero, rejecting the null if so and failing to reject otherwise. For a two-sided alternative, simply replace $\zeta(\pi^*, \alpha - \beta)$ with $\zeta(\pi^*, (\alpha - \beta)/2)$ to control the level of the procedure at α. See Rosenbaum (1992) and Fogarty and Small (2016) for similar formulations of sensitivity analyses as convex programs.

A minor complication is that for small values of I or for small values for β, the right-hand side of (17) need not be monotone increasing in μ_{π^*} if $2\bar{\Gamma}/(1 + \bar{\Gamma}) \geq \Gamma/(1 + \Gamma) + 1/2$, as decreasing μ_{π^*} may lead to an increase in the component dependent on the variance bound which exceeds the corresponding decrease in the additive term μ_{π^*}. To remedy this, one can simply find the value for μ_{π^*} over the range $[(\Gamma/(1 + \Gamma) + 1/2)/2, \bar{\Gamma}/(1 + \bar{\Gamma})]$

which maximizes the right-hand side of (17) through a bisection algorithm, and then proceed with the quadratic program using this single value. If $2\bar{\Gamma}/(1+\bar{\Gamma}) < \Gamma/(1+\Gamma)+1/2$, the right-hand side of (17) is, subject to (18), maximized at $\mu_{\pi^*} = \bar{\Gamma}/(1+\bar{\Gamma})$, so one can proceed by replacing μ_{π^*} with $\bar{\Gamma}/(1+\bar{\Gamma})$ and solving the required quadratic program. Importantly, the method only requires solving a single quadratic program. Quadratic programs can be solved by many free and commercially available solvers; in the supplementary materials, we provide code implementing our method using the `R` package for the solver `Gurobi`, which is free for academic use. We also provide options to replace the constraint (17), justified by the Central Limit Theorem, with bounds described in §A of the supplemental appendix which are valid for any I through distribution-free concentration inequalities.

5. Simulations.

5.1. *Type I error control.* In the following simulations, we demonstrate that the extended sensitivity analysis introduced in §3 has the correct level. We consider two important cases: (1) when no unmeasured bias is present and (2) when the there is unmeasured bias but the sensitivity analysis is conducted at the true values of Γ and $\bar{\Gamma}$. In both settings we test Fisher's sharp null that $\tau = 0$ using the difference in means test with desired Type I error control at $\alpha = 0.05$. We set $\beta = \alpha/10 = 0.005$ for conducting the extended sensitivity analysis. The following treatment model, outcome model, and simulation settings were used to conduct the Type I error control simulations:

1. **Treatment model:** $\Pi_i^* = 1/2$ with probability $p = 2(\Gamma - \bar{\Gamma})/\{(\Gamma - 1)(\bar{\Gamma}+1)\}$ and $\Pi_i^* = \Gamma/(1+\Gamma)$ with probability $1-p$.
2. **Outcome model:**
   - *unbiased:* $Y_i = \tau \cdot (Z_{i1} - Z_{i2}) + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$,
   - *biased:* $Y_i = \tau \cdot (Z_{i1} - Z_{i2}) + \{2 \cdot \chi(\pi_i > 1 - \pi_i) - 1\} \cdot |\epsilon_i|$ where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$.
3. **Sensitivity parameters:**
   - $\Gamma \in \{1, 1.1, 1.25, 1.5, 2\}$,
   - $\bar{\Gamma} \in \{1, 1.05, 1.1, 1.15, 1.2, 1.25, 1.3, 1.35,$ $1.4, 1.45, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0\}$,
   - $\bar{\Gamma} \leq \Gamma$.
4. **Study and simulation size:** $I = 100$ pairs, $N_{sim} = 5000$ simulations.

In the biased setting, the unit with higher potential outcome under control has higher probability of receiving treatment. When $\Gamma = \bar{\Gamma} = 1$ we use the convention that $p = 0/0 = 0$. The value of $p = \mathbb{P}(\Pi_i^* = 1/2)$ was chosen so that the population treatment model satisfies $\mathbb{E}[\bar{\Pi}^*] = \bar{\Gamma}/(1 + \bar{\Gamma})$. The results of the simulation study for the biased and unbiased settings are shown in Table 1 and the table in §E.1 of the supplemental appendix, respectively. The extended sensitivity procedure correctly controls the Type I error rate for all pairs of sensitivity parameters $(\Gamma, \bar{\Gamma})$ tested. The first row of each table, where $\bar{\Gamma} = 1$, corresponds to tests under the absence of unmeasured confounding. The pairs where $\Gamma = \bar{\Gamma}$ correspond to the conventional worst-case sensitivity analysis. Under the unbiased treatment model, the extended sensitivity analysis is typically more conservative as we increase Γ or $\bar{\Gamma}$. In the biased setting, we observe the same pattern as we vary Γ, but as $\bar{\Gamma}$ approaches Γ, the level of the extended sensitivity analysis does not decrease monotonically. In fact, at a certain value of $\bar{\Gamma}$, the extended sensitivity analysis becomes less conservative as we approach Γ. In short, the solution $\boldsymbol{\pi}_{sup,\beta}$ to the optimization problem in (12) tends to more closely approximate the true allocation $\boldsymbol{\pi}_0$ when $\bar{\Gamma}$ is close to either 1 or Γ in the biased setting. When $\bar{\Gamma}$ is close to 1, the feasible set of $\boldsymbol{\pi}$'s is closely bounded around $\boldsymbol{\pi}_0 \approx \mathbf{1} \cdot 1/2$. When $\bar{\Gamma}$ is close to Γ the true allocation is $\boldsymbol{\pi}_0 \approx \boldsymbol{\pi}_\Gamma$ and the extended sensitivity analysis behaves like the conventional sensitivity analysis, where $\boldsymbol{\pi}_{\text{sup},\beta} = \boldsymbol{\pi}_\Gamma$ yields a tight upper bound on the probability in (11). In between these edge cases, when the feasible set of $\boldsymbol{\pi}$ is relatively large and the trade-off between maximizing expectation and variance is more nuanced, (12) may produce solutions $\boldsymbol{\pi}_{\text{sup},\beta}$ that yield appreciably more conservative inference than if had we known the true $\boldsymbol{\pi}_0$.

5.2. *The power of an extended sensitivity analysis.* The power of a sensitivity analysis quantifies the ability of an observational study design to distinguish treatment effects from unmeasured bias. Formally, it reports for a given study design the probability of rejecting a false null hypothesis for a chosen level α and sensitivity parameter Γ under 'favorable' conditions, defined in Rosenbaum (2010, Chapter 14), as the presence of a treatment effect that causes meaningful effects and absence of unmeasured biases. The investigator cannot determine from observable data alone whether or not such favorable conditions hold. An attractive study design would be highly insensitive to unmeasured confounding if she was lucky enough to find herself in this favorable setting. The power of an extended sensitivity analysis extends this formalism to the triplet $(\alpha, \Gamma, \bar{\Gamma})$. Power simulations for $\alpha = 0.05$ and several pairs of $(\Gamma, \bar{\Gamma})$ are reported in Table 2 and the table in §E.2 in the

			$\mathbf{\Gamma}$		
$\bar{\mathbf{\Gamma}}$	1	1.1	1.25	1.5	2
1	0.047	0.047	0.045	0.046	0.044
1.05		0.022	0.011	0.007	0.005
1.1		0.032	0.010	0.004	0.003
1.15			0.012	0.002	0.002
1.2			0.017	0.004	0.001
1.25			0.025	0.004	0.001
1.3				0.006	0.000
1.35				0.009	0.001
1.4				0.011	0.001
1.45				0.014	0.001
1.5				0.025	0.001
1.6					0.003
1.7					0.004
1.8					0.006
1.9					0.011
2					0.021

TABLE 1

Rejection probability of the true null hypothesis, $H_0 : \tau = 0$, under the biased setting with target Type I error control at $\alpha = 0.05$. The Monte Carlo standard error of these probability estimates is bounded above by $\sqrt{0.05 \times 0.95/5000} \approx 0.003$ if the true Type I error rate is 0.05.

supplemental appendix for $\tau = 0.5$ and $\tau = 0.25$, respectively. Other than the presence of a 'meaningful' treatment effect τ, the simulation settings are identical to the unbiased setting in §5.1.

Unsurprisingly, the power of the extended sensitivity analysis decreases as $\bar{\Gamma}$ approaches Γ. If the investigator has reason to believe that unmeasured confounding is heterogeneous and that extreme pairwise unmeasured confounding is possible but relatively rare, the conventional sensitivity analysis is likely unduly conservative. Further, the extended sensitivity analysis allows the investigator to compare the power of competing study designs under different assumptions about the maximal and expected degree of unmeasured confounding.

6. Extended sensitivity analysis for returns to schooling.

6.1. *A model for returns to schooling.* How does going to college affect job earnings? The question and the implications of the many putative answers are important to education policy experts and parents alike. It has been empirically demonstrated that log earnings are nearly a linear function of schooling (see, for instance, Card and Krueger, 1992). In the idealized paired observational setting introduced in §§2.1-2.2 where the treatment condition is attending college for at least two years and the control condi-

			$\boldsymbol{\Gamma}$		
$\bar{\boldsymbol{\Gamma}}$	1	1.1	1.25	1.5	2
1	0.998	0.999	0.998	0.999	0.999
1.05		0.994	0.990	0.984	0.978
1.1		0.996	0.984	0.965	0.941
1.15			0.977	0.947	0.896
1.2			0.978	0.928	0.833
1.25			0.979	0.907	0.759
1.3				0.890	0.719
1.35				0.884	0.664
1.4				0.879	0.626
1.45				0.874	0.578
1.5				0.882	0.541
1.6					0.505
1.7					0.478
1.8					0.463
1.9					0.472
2					0.486

TABLE 2

Rejection probability of the false null hypothesis, $H_0 : \tau = 0$, under the unbiased setting with true alternative hypothesis $H_1 : \tau = 0.5$. The Monte Carlo standard error of these probability estimates is bounded above by $\sqrt{0.5 \times 0.5/5000} \approx 0.007$.

tion is receiving at most a high school diploma, a hypothesized treatment effect $\tau \times 100$ would describe the percentage increase in earnings associated with attending at least two years of college, the minimum number of years to receive an associates degree. Formally, we consider the multiplicative treatment effect hypothesis $H_\tau : R_{Tij} = \tau R_{Cij}$ where (R_{Tij}, R_{Cij}) are potential earnings after attending college or not. Choosing $t(\mathbf{Z}, \mathbf{F}) = \mathbf{Z}^T\mathbf{q}$ to be the adjusted difference-in-means test comparing log earnings, q_{ij} would take the form $q_{ij} = (\log R_{Tij} - \log R_{Cij'}) - \log(\tau)$ and $q_{ij'} = -q_{ij}$ under H_τ.

Let $X = [X_f, X_s]$ where X_f and X_s are familial and subject level covariates. In an idealized sibling comparison design, the strong ignorability condition in (1) would hold with respect to X_f; that is, if for all x_f,

$$(R_T, R_C) \perp\!\!\!\perp Z \mid X_f, \;\; 0 < \mathbb{P}(Z = 1 \mid X_f = x_f) < 1. \tag{19}$$

If X_s does not affect treatment assignment but does predict potential outcomes, this sibling version of strong ignorability will still hold. For example, in the sibling pairs from the WLS data that we consider in the following section, the age at which income is measured (AGE) is different between siblings. If $X_s = AGE$, then it is conceivable that X_s does not affect whether a sibling went to college or not. This would not be the case for people who went to college later in life or whose family characteristics may have changed over time, in which case AGE would be a proxy for those changes. Regardless, model-agnostic adjustment for X_s and X_f can improve the power of the resulting sensitivity analysis (Rosenbaum, 2002a). For example, we can use simple linear regression to adjust for X by replacing $\mathbf{q}$ with $(I - H_{X_s})\mathbf{q}$ where H_{X_s} is the orthogonal projection onto X_s without an intercept.

6.2. *Ashenfelter: Conventional versus extended sensitivity analysis.* To illustrate the differences between the conventional and extended sensitivity analyses, we return to the twin study of Ashenfelter and Rouse (1998) (AR). AR collected survey data on 680 monozygotic twins (340 pairs) attending the Twinsburg Twins Festival in Twinsburg, Ohio during the summers of 1991, 1992, and 1993. We consider the 40 pairs of twins where one twin attend at least two years of college and the other had no more than a high school education, and where both twins were employed at the time of data collection. Assuming no unmeasured confounding, testing Fisher's sharp null H_0 yields a p-value of ≈ 0.0001. We obtain a 95% confidence interval for $\log(\tau)$ of [0.16,0.43] by inverting H_τ for $\tau \in \mathbb{R}_+$ at $\alpha = 0.05$ with a two-sided alternative. Exponentiating the endpoints, attending at least two years of college versus receiving at most a high school diploma increased wages by between 17% and 53% with 95% confidence.

Being a retrospective study neither baseline IQ nor any other intelligence scores were collected, and a critical reader may point to the possible presence of ability bias as a basis to call the conclusions of the study into question. Conducting a sensitivity analysis produces a quantitative rejoinder to this type of criticism in the form of a *sensitivity value* Γ^* for the conventional analysis and a *sensitivity curve* $(\Gamma^*, \bar{\Gamma}^*)$ for the extended analysis. The sensitivity value is the largest bound on the maximal bias such that the qualitative conclusions of the study do not change (i.e., such that we reject H_0). The sensitivity curve is the two-dimensional analog of the sensitivity value and can be seen as the threshold between the gray region (reject H_0) and the white region (retain H_0) in Figure 3. At the limits of the sensitivity curve, we recover two separate single-parameter sensitivity analyses. The sensitivity value returned by the conventional analysis corresponds to the point where the sensitivity curve intersects the $y = x$ line ($\Gamma^* \approx 2.36$). The limit of the sensitivity curve as $\Gamma \to \infty$ is the sensitivity value of a single-parameter sensitivity analysis that bounds the typical bias ($\bar{\Gamma} \approx 1.22$).

6.3. *Ability Bias: Cross-study sensitivity analysis calibration.* Without context, the sensitivity curve and values from the Ashenfelter analysis may be difficult to interpret. In response to the critic of the "equal abilities" hypothesis for twins, we would ideally like to report whether or not the Ashen-

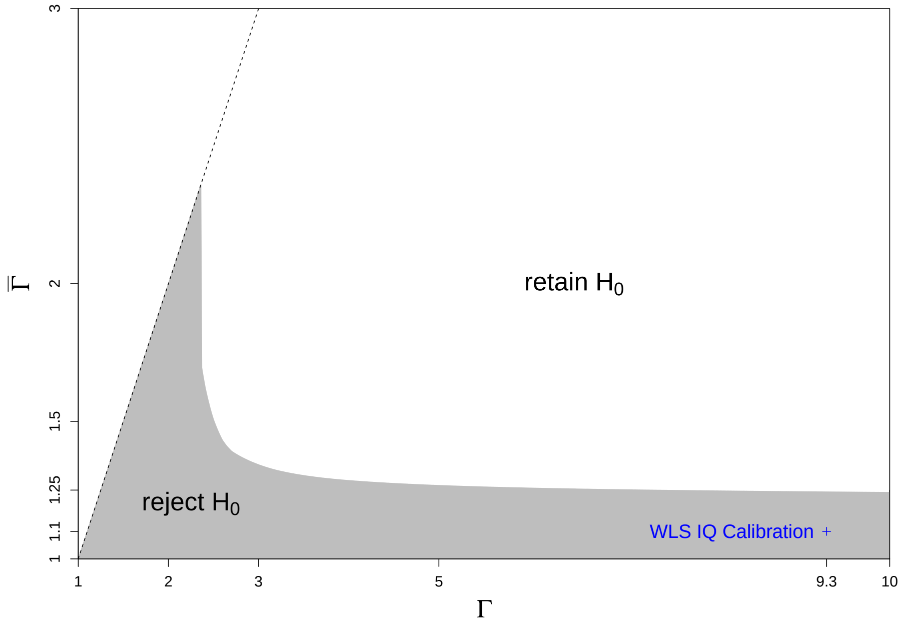


FIG 3. *Extended sensitivity curve from the AR study calibrated to the estimates of ability bias from the WLS study (cross). The gray region indicates the sensitivty parameter pairs* $(\Gamma, \bar{\Gamma})$ *for which* H_0 *can still be rejected. The point where the sensitivity curve intersects the* $y = x$ *line corresponds to the sensitivity value returned by conventional sensitivity analysis (*$\Gamma^* \approx 2.36$*). The limit of the curve as* $\Gamma \to \infty$ *corresponds to the sensitivity value returned by the single-parameter sensitivity analysis that bounds the typical bias (*$\bar{\Gamma}^* \approx 1.22$*).*

felter study is sensitive to plausible patterns of ability bias. One strategy for addressing this is to estimate the bias due to ability from a *calibration study* that has a comparable design and information on baseline ability such as IQ. We can then *calibrate* the sensitivity analysis to these estimates of Γ and $\bar{\Gamma}$. To implement this *cross-study calibration*, we modify the procedure established in Hsu and Small (2013) to calibrate sensitivity parameters to observed covariates. In brief, one fits ostensible treatment and outcome models – for instance, via linear and logistic regression – and uses the resulting model fits to estimate $\boldsymbol{\pi}^*$, $\bar{\Gamma}$, and Γ. The details of this step can be found in §C of the supplemental appendix. Calibrating the sensitivity analysis to estimates of ability bias provides the context relevant to the critic's concerns.

To assess the robustness of the AR study to ability bias, we use the sibling data from the WLS study introduced in §1.2 to design a calibration study. We constructed a set of 171 same-sex, full-sibling pairs that received discordant treatment. We let $Z_{ij} = 0$ if sibling j in pair i received 12 or fewer years of education and $Z_{ij} = 1$ if he or she received 14 or more years of education (at least two years of college). Log income for the previous year was collected for WLS participants and their siblings in 1975 and 1977, respectively. To more closely approximate the superpopulation from which the AR twins came, we only consider siblings where both had non-zero income at the time of collection (i.e. were employed). As outlined in the previous section, we let $X_s = AGE$ and use regression to adjust $\mathbf{q}$ for the age at which income was collected. This calibration analysis is stylized to some extent to avoid obscuring the primary contribution of our method. Many other subject-level covariates are available for adjustment via regression. A detailed analysis including treatment modification with respect to gender and more thorough covariate adjustment would not preclude the use nor usefulness of our method.

Using the 171 WLS sibling pairs, we estimate that $\Gamma \approx 9.3$ and $\bar{\Gamma} \approx 1.1$, summarizing the information we have about maximal and typical biases due to IQ disparities. Heterogeneneity of ability bias can explain the considerable difference between these two measures of confounding. The histogram of the estimated $\boldsymbol{\pi}^*$ in Figure 4 indicates that most sibling pairs have modest differences in intelligence in high school but in a few rare cases the disparity in sibling IQ exposes pairs to high levels of bias. Calibrating the conventional sensitivity analysis of AR to the WLS study would suggest that our conclusions are likely not robust to plausible patterns of ability bias since $\Gamma^* < 9.3$. However, calibration of the extended sensitivity analysis suggests otherwise. In Figure 3, the WLS IQ calibration point $(9.3, 1.1)$ is indicated by the blue cross and falls below the sensitivity curve. The single-parameter sensitivity analysis that bounds the typical bias agrees with the extended analysis that the conclusions are robust to plausible patterns of ability bias ($\bar{\Gamma}^* \geq 1.1$). Incorporating information about the heterogeneity of ability bias by bounding both the maximal and typical biases promotes a less pessimistic assessment of an observational study's robustness to unmeasured confounding. When information on the heterogeneity of potential confounders is available, as in the above cross-study calibration analysis, the extended sensitivity analysis provides a richer picture of the study's robustness to hidden bias.

6.4. *Sensitivity intervals: Interval estimates with hidden bias.*

For a fixed bound on the worst-case bias, incorporating heterogeneous bias through the extended sensitivity can also produce narrower *sensitivity intervals* than

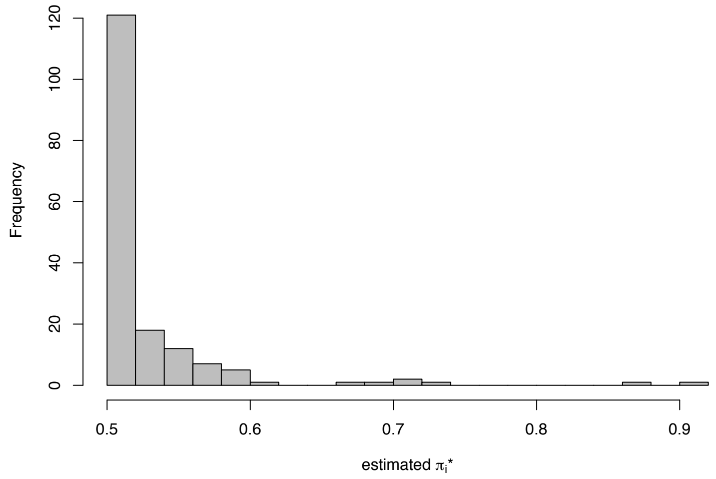


FIG 4. *Histogram of $\boldsymbol{\pi}^*$ estimated for 171 same-sex, full-sibling pairs from the WLS study.*

those attained through the conventional analysis. Representing a natural extension of confidence intervals to inference in the presence of unmeasured confounding, a $100(1-\alpha)\%$ sensitivity interval is constructed by inverting a level-α extended sensitivity analysis with a two-sided alternative at a given pair of values $(\Gamma, \bar{\Gamma})$. Explicitly, let $p_\beta(\Gamma, \bar{\Gamma}, \tau)$ be the two-sided p-value bound returned by the extended sensitivity analysis in (12) for particular values of Γ and $\bar{\Gamma}$. Then, a $100(1-\alpha)\%$ sensitivity interval can be written as $\mathcal{I}(\{\tau : p_\beta(\Gamma, \bar{\Gamma}, \tau) \leq \alpha\})$, where $\mathcal{I}(A)$ is the smallest interval containing the set A. At $\Gamma = \bar{\Gamma} = 1$, the sensitivity interval is simply the corresponding confidence interval found by inverting H_τ using the randomization p-value given in (2) as would be justified in a paired experiment. Setting $\Gamma = \bar{\Gamma} > 1$ returns sensitivity intervals produced through the conventional sensitivity analysis, while setting $\Gamma > \bar{\Gamma} > 1$ employs the extended sensitivity analysis in constructing the sensitivity intervals.

Table 3 illustrates the potential for reduced interval lengths through accommodating heterogeneity in unmeasured confounding. It reports 95% sensitivity intervals for $\log(\tau)$ in the AR study with three pairs of values for Γ and $\bar{\Gamma}$. The first, denoted by $\mathcal{I}_{rand}$, is the 95% sensitivity interval assuming no unmeasured confounding previously reported in §6.2. The second, $\mathcal{I}_{sup}$, is the 95% sensitivity interval derived by setting $\Gamma = \bar{\Gamma} = 9.3$, the calibrated value of the maximal bias parameter from the WLS study. This is precisely the sensitivity interval that the conventional sensitivity analysis bounding only the worst-case confounding would return. The final interval, $\mathcal{I}_{ext}$, is the 95% sensitivity interval setting $\Gamma = 9.3$, $\bar{\Gamma} = 1.1$ in accord with the calibrated values of the maximal and typical bias from the WLS study. We see that $\mathcal{I}_{ext}$ is more than 80% shorter than $\mathcal{I}_{sup}$. Further, both $\mathcal{I}_{rand}$ and $\mathcal{I}_{ext}$ exclude zero while $\mathcal{I}_{sup}$ does not. The positive finding in the unconfounded setting can be explained away by bias calibrated to the WLS study using the conventional sensitivity model, but not when using the extended sensitivity model. Once again, we see that when it is plausible that the typical bias to which pairs are subject is materially smaller than the worst-case bias, the conventional analysis may be overly pessimistic about how informative the data is.

Interval Type	95% Sensitivity Interval
$\mathcal{I}_{rand}$	[0.16,0.43]
$\mathcal{I}_{sup}$	[-0.88,1.63]
$\mathcal{I}_{ext}$	[0.06,0.53]
$100 \times (1 - \lvert\mathcal{I}_{ext}\rvert/\lvert\mathcal{I}_{sup}\rvert)$	81%

TABLE 3

95% sensitivity intervals for $\log(\tau)$ in the AR study constructed by inverting H_τ for different values of Γ and $\bar{\Gamma}$. $\mathcal{I}_{rand}$ is the 95% confidence interval for $\log(\tau)$ in the unconfounded setting, $\Gamma = \bar{\Gamma} = 1$. $\mathcal{I}_{sup}$ and $\mathcal{I}_{ext}$ are 95% sensitivity intervals derived from the conventional sensitivity analysis and the extended sensitivity analysis respectively. These intervals are formed using the sensitivity parameters calibrated from the WLS data, $(\Gamma, \bar{\Gamma}) = (9.3, 1.1)$. The percentage reduction in interval length from accommodating heterogeneous unmeasured confounding, $100 \times (1 - |\mathcal{I}_{ext}|/|\mathcal{I}_{sup}|)$, is reported in the last row.

7. Concluding remarks.

While convenient for ease of calculation, the low-dimensional sensitivity analysis bounding the supremum may fail to address specific concerns with unmeasured confounding in certain contexts. Rosenbaum and Silber (2009) present an amplification of the conventional sensitivity analysis, where the one-dimensional analysis based on Γ is mapped to a curve of two-dimensional analyses which simultaneously bound the extent to which differences in unobserved covariates can influence the odds of being treated and the odds of having a higher potential outcome under control by the pair (Λ, Δ). This amplificiation provides an aid to interpretation, allowing the researcher to posit bounds on the extent to which unmeasured confounding can affect treatment decisions and the outcome variable. Rather than amplifying the conventional sensitivity analysis, the extended sensitivity analysis provides the researcher a way to further control the distribution of the unmeasured confounders beyond bounding the supremum. In fact, amplification and extension can be viewed as complementary tools available to the researcher. It is straightforward to employ both: the conventional supremum bound Γ that appears in the extended sensitivity analysis may be amplified yielding yet an even richer analysis, with $\bar{\Gamma}$ bounding the typical probability that the treated individual in a pair has the larger (smaller) potential outcome under control for greater-than (less-than) alternatives.

Framing sensitivity analysis in terms of the typical bias is not a new idea, but has been largely unaddressed in the literature; the idea of expected bias appears briefly in Wang and Krieger (2006) in the context of population-level inference for binary outcomes but is not the focus of the paper. In a particular sense, Cornfield et al. (1959) anticipated the duality of both amplified and extended sensitivity analyses in their seminal work on sensitivity analysis. In their smoking and lung cancer example, the authors considered a hypothetical hormone X which increases the probability of developing lung cancer among those exposed from r_2 to r_1 and due to a positive correlation between exposure to X and smoking, appears in a higher proportion among smokers than non-smokers (i.e $p_1 > p_2$). At once, Cornfield et al. (1959) captures the spirit of an amplified analysis in specifying how X is related to both treatment assignment and outcome and that of an extended analysis by imagining that hormone X is not completely absent among non-smokers and completely present among smokers, leading to exposure to bias that is heterogeneous across subjects within both groups.

The concept of heterogeneous unmeasured confounding appeared naturally, if not intentionally, in Cornfield's original example. The extended sensitivity analysis introduced in this paper brings this idea into a modern light and provides the researcher with a way to conduct a sensitivity analysis while bounding both maximal and typical biases in matched pair studies. Using two sibling studies on the returns of schooling to income, we demonstrated that a sensitivity analysis bounding the maximal *and* typical bias is both natural and less susceptible to an overly pessimistic view of the study's robustness to hidden bias. When a researcher believes that most, if not all, pairs are exposed to the worst-case bias, our procedure can recover the conventional analysis by setting $\bar{\Gamma} = \Gamma$. If however, the researcher is worried that some, though few, pairs may be exposed to arbitrarily large biases all is not lost; by letting Γ tend to ∞ the extended sensitivity analysis recovers a single-parameter sensitivity analysis that bounds the typical bias.

SUPPLEMENTARY MATERIAL

Supplement to "An Extended Sensitivity Analysis for Heterogeneous Unmeasured Confounding with Application to Sibling Stud-

ies of Returns to Education"
(doi: COMPLETED BY THE TYPESETTER; .zip). We include in the supplementary material appendices illustrating the construction of valid finite-sample uncertainty sets and describing the calibration of the sensitivity parameters and R code that contains the function that implements the extended sensitivity procedure and scripts that produce the analysis in the figures.

References.

Ashenfelter, O. and Rouse, C. (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *The Quarterly Journal of Economics* **113** 253-284.

Becker, G. S. (2009). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89** 1012-1016.

Bhatia, R. and Davis, C. (2000). A better bound on the variance. *The American Mathematical Monthly* **107** 353-357.

Card, D. (1999). The Causal Effect of Education on Earnings. (O. C. Ashenfelter and D. Card, eds.). *Handbook of Labor Economics* **3** 1801 - 1863. Elsevier.

Card, D. and Krueger, A. B. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy* **100** 1–40.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (1959). Smoking and lung cancer:recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22** 173-203.

Donovan, S. J. and Susser, E. (2011). Commentary: Advent of sibling designs. *International Journal of Epidemiology* **40** 345.

Egleston, B. L., Scharfstein, D. O. and MacKenzie, E. (2009). On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics* **65** 497–504.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd.

Fogarty, C. B. and Small, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *Journal of the American Statistical Association* **111** 1820–1830.

Frisell, T., Öberg, S., Kuja-Halkola, R. and Sjölander, A. (2012). Sibling comparison designs: Bias from non-shared confounders and measurement error. *Epidemiology* **23** 713-720.

Griliches, Z. (1970). Notes on the role of education in production functions and growth accounting. In *Education, Income, and Human Capital. NBER Chapters* 71-127. National Bureau of Economic Research, Inc.

Griliches, Z. (1979). Sibling models and data in economics: Beginnings of a survey. *Journal of Political Economy* **87** S37–S64.

Hájek, J., Sĭdák, Z. and Sen, P. K. (1999). *Theory of Rank Tests*. Academic Press, San Diego.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* **15** 609-627.

Hasegawa, R. and Small, D. (2017). Sensitivity analysis for matched pair analysis of binary data: From worst case to average case analysis. *Biometrics* **73** 1424-1432.

Hauser, R. M., Sheridan, J. T. and Warren, J. R. (1999). Socioeconomic achievements of siblings in the life course. *Research on Aging* **21** 338-378.

Herd, P., Carr, D. and Roan, C. (2014). Cohort profile: Wisconsin longitudinal study (WLS). *International Journal of Epidemiology* **43** 34–41.

Hosman, C. A., Hansen, B. B. and Holland, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Annals of Applied Statistics* **4** 849–870.

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

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* **93** 126–132.

Liu, W., Kuramoto, S. J. and Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science* **14** 570–580.

Marcus, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational and Behavioral Statistics* **22** 193–201.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association* **79** 565–574.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13-26.

Rosenbaum, P. R. (1992). Detecting bias with confidence in observational studies. *Biometrika* **79** 367–374.

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science* **14** 259–278.

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17** 286–327.

Rosenbaum, P. R. (2002b). *Observational Studies*. Springer, New York.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer, New York.

Rosenbaum, P. R. and Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association* **104** 1398–1405.

Rubin, D. B. (1980). Comment (on D. Basu, Randomization analysis of experimental data: The Fisher randomization test). *Journal of the American Statistical Association* **75** 591–593.

Stanek, K. C., Iacono, W. G. and McGue, M. (2011). Returns to education: What do twin studies control? *Twin Research and Human Genetics* **14** 509–515.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statist. Sci.* **25** 1–21.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine* **167** 268–274.

Wang, L. and Krieger, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Statistics in Medicine* **25** 2257–2271.

Yu, B. and Gastwirth, J. L. (2005). Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics* **6** 201–209.

Zubizarreta, J. R., Cerdá, M. and Rosenbaum, P. R. (2013). Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design. *Epidemiology* **24** 79-87.

C. B. Fogarty
Operations Research and Statistics Group
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142
USA
E-mail: cfogarty@mit.edu

R. B. Hasegawa
Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania 19104
USA
E-mail: raiden@wharton.upenn.edu