DESIGN SENSITIVITY IN OBSERVATIONAL STUDIES

PAUL R. ROSENBAUM

Abstract. If an observational study were free of unobserved bias, we would not know this from the observed data. The best we could hope to say is that only very large unobserved biases could alter the conclusions of the study. What features of the design of an observational study reduce its sensitivity to unobserved biases? How can designs for observational studies be compared quantitatively in terms of their ability to resist unobserved biases? The talk is from [30], [31] and [39], the last joint with Dylan Small.

1. Notation and Review

1.1. **Review: Notation for a Paired Randomized Experiment**[11] [21] [36] [2]**.** Observed covariate $\mathbf{x}$ and an unobserved covariate u. I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, one control, matched for $\mathbf{x}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, but not matched for u, so typically $u_{i1} \neq u_{i2}$. $Z_{ij} = 1$ if j received the treatment in pair i, and $Z_{ij} = 0$ if j received the control, so $Z_{i1}+Z_{i2} = 1$. Subject (i, j) has two potential responses, (r_{Tij}, r_{Cij}), r_{Tij} observed under treatment, $Z_{ij} = 1$, r_{Cij} observed under control, $Z_{ij} = 0$, so the effect of the treatment is $r_{Tij} - r_{Cij}$; Neyman [21] and Rubin [36]. Treatments given at doses (v_{Tij}, v_{Cij}), possibly $(v_{Tij}, v_{Cij}) = (1, 0)$. Write $\mathcal{F}$ for $\{(r_{Tij}, r_{Cij}, v_{Tij}, v_{Cij}, \mathbf{x}_{ij}, u_{ij})$, $i = 1, \ldots, I$, $j = 1, 2\}$ and $\mathcal{Z}$ for the event $\{Z_{i1} + Z_{i2} = 1,\ i = 1, \ldots, I\}$; then $\mathcal{F}$ and $\mathcal{Z}$ are fixed by conditioning in Fisher's [11] theory of randomization inference. Randomization in pairs ensures $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \frac{1}{2}$, $\forall i$, with independently in distinct pairs. Observed response is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$, observed dose is $V_{ij} = Z_{ij}\, v_{Tij} + (1 - Z_{ij})\, v_{Cij}$.

1.2. **Review: Randomization inference for constant treatment effect**[18] [27]**.** If treatment effect is constant, $\tau = r_{Tij} - r_{Cij}$, then $R_{ij} = r_{Cij} + Z_{ij}\,\tau$, and the treated-minus-control difference is $D_i = (2Z_{i1} - 1)\,(R_{i1} - R_{i2}) = \tau + \epsilon_i$ where $\epsilon_i = (2Z_{i1} - 1)\,(r_{Ci1} - r_{Ci2})$. Test $H_0 : \tau = \tau_0$ by ranking $|D_i - \tau_0|$ from 1 to I; then Wilcoxon's signed rank statistic, W_{τ_0}, is the sum of the ranks for which $D_i - \tau_0 > 0$, where ties are assumed absent. If $H_0 : \tau = \tau_0$ is true, randomization ensures that $D_i - \tau_0 = \epsilon_i$ is $r_{Ci1} - r_{Ci2}$ or $r_{Ci2} - r_{Ci1}$, each with probability $\frac{1}{2}$, independently in different pairs. Given $\mathcal{Z}$, $\mathcal{F}$, if $H_0 : \tau = \tau_0$ is true, then the $|D_i - \tau_0|$ are fixed, the $D_i - \tau_0$ are independent, $\Pr(D_i - \tau_0 > 0) = \frac{1}{2}$, and each D_i is symmetric about τ_0, so W_{τ_0} is the sum of I independent random variables taking values i or 0 each with probability $\frac{1}{2}$, $i = 1, \ldots, I$. A confidence interval for τ is obtained by inverting the test, and the Hodges-Lehmann (HL) estimate $\widehat{\tau}$ of τ is (essentially) the

University of Pennsylvania, Philadelphia, PA 19104-6430 USA, rosenbaum@stat.wharton.upenn.edu. June 2007. Supported by NSF. Reprints/overheads: http://www-stat.wharton.upenn.edu/~rosenbap/.

solution to $W_{\widehat{\tau}} = I\,(I+1)/4 = \frac{1}{2}(1+2+\ldots+I)$. Null distribution of W_{τ_0} is the same for all (untied) $\mathcal{F}$, but the nonnull distribution depends on $\mathcal{F}$ or a model that generates $\mathcal{F}$. A common model for a randomized experiment has $(r_{Ci1} - r_{Ci2})/\sigma \sim_{iid} F(\cdot)$ were $\sigma > 0$ and $F(\cdot)$ is continuous and symmetric about zero, so randomization ensures $\epsilon_i/\sigma \sim_{iid} F(\cdot)$.

1.3. **Review: Randomization inference if randomization is an instrument for the dose received**[13] [17] [25] [28] [27]**.** Treatment effect is proportional to dose if $r_{Tij} - r_{Cij} = \beta\,(v_{Tij} - v_{Cij})$, which is a constant effect for doses $(v_{Tij}, v_{Cij}) = (1,0)$, $\forall i,j$. If effect is proportional to dose, then $R_{ij} - \beta\,V_{ij} = r_{Tij} - \beta v_{Tij} = r_{Cij} - \beta v_{Cij} = a_{ij}$, say, takes the same value a_{ij} whether $Z_{ij} = 1$ or $Z_{ij} = 0$, and a_{ij} is fixed by conditioning on $\mathcal{F}$. Test $H_0 : \beta = \beta_0$ using $R_{ij} - \beta_0\,V_{ij} = a_{ij} + (\beta - \beta_0)\{Z_{ij}v_{Tij} + (1 - Z_{ij})\,v_{Cij}\}$ which is fixed at a_{ij} if $H_0 : \beta = \beta_0$ is true, but otherwise varies with Z_{ij}. Write $D_i^{\beta_0}$ for the matched pair difference in $R_{ij} - \beta_0\,V_{ij}$, treated $(Z_{ij} = 1)$ minus control $(Z_{ij} = 0)$, so that $D_i^{\beta_0} = (\beta - \beta_0)\,S_i + \epsilon_i$ where $S_i = Z_{i1}\,(v_{Ti1} - v_{Ci2}) + (1 - Z_{i1})\,(v_{Ti2} - v_{Ci1})$ and $\epsilon_i = (2Z_{i1} - 1)\,(a_{i1} - a_{i2})$. In a randomized experiment, if $H_0 : \beta = \beta_0$ were true, then $S_i = \pm\,(a_{i1} - a_{i2})$ each with probability $\Pr(Z_{i1} = 1) = \frac{1}{2}$.

1.4. **Review: Sensitivity to Departures from Random Assignment in Observational Studies** [10] [33] [23] [24] [12] [16] [19] [1] [9] [27] [32]**.** (i) In population, before matching, treatment assignments were independent, with unknown probabilities $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$, (ii) subjects with same *observed* $\mathbf{x}_{ij}$ may differ in *unobserved* u_{ij} and hence in odds of treatment by factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}\,(1 - \pi_{ik})}{\pi_{ik}\,(1 - \pi_{ij})} \leq \Gamma, \;\; \forall\, i, j, k \tag{1.1}$$

and (iii) the distribution of treatments within treated/control matched pairs $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F})$ is then obtained by conditioning on $Z_{i1} + Z_{i2} = 1$. Here, $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$. If $\Gamma = 1$, then $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ ensures $\pi_{ij} = \pi_{ik}$, $i = 1, \ldots, I$, whereupon $\Pr(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}) = \pi_{i1}/(\pi_{i1} + \pi_{i2}) = \frac{1}{2}$, and the distribution of treatment assignments is again the randomization distribution: bias solely due to observed $\mathbf{x}$ can be eliminated by matching on $\mathbf{x}$. If $\Gamma > 1$ in (1.1), then matching on $\mathbf{x}$ may fail to equalize the π_{ij} in pair i. Γ is unknown. A sensitivity analysis calculates, for several values of Γ, the range of possible inferences. How large must Γ be before qualitatively different causal interpretations are possible?

1.5. **Review: Sensitivity Analysis with the Signed Rank Statistic** [23] [24]**.** If (1.1) and $H_0 : \tau = \tau_0$ are true, then the null distribution of W_{τ_0} is unknown but is bounded by two known distributions. Write $\theta = \Gamma/(1 + \Gamma)$ so $\theta \geq \frac{1}{2}$ because $\Gamma \geq 1$. Write $\overline{\overline{W}}$ for the sum of I independent random variables taking value i with probability θ and value 0 with probability $1 - \theta$, $i = 1, \ldots, I$; also, write $\overline{W}$ for the sum of I independent random variables taking value i with probability $1 - \theta$ and value 0 with probability θ. Then (1.1) and $H_0 : \tau = \tau_0$ imply the sharp bounds

$$\Pr\left(\overline{W} \geq w\right) \leq \Pr\left(W_{\tau_0} \geq w \mid \mathcal{Z}, \mathcal{F}\right) \leq \Pr\left(\overline{\overline{W}} \geq w\right), \;\; \forall w; \tag{1.2}$$

TABLE 1. Power of the Sensitivity Analysis: Heterogenity vs Sample Size.

Errors	I Pairs	τ	σ	σ^2/I	$\Gamma = 1$	$\Gamma = 1.5$	$\Gamma = 2$
Normal	120	$\frac{1}{2}$	1	1/120	1.00	0.96	0.60
Normal	30	$\frac{1}{2}$	$\frac{1}{2}$	1/120	1.00	1.00	0.96
Cauchy	200	$\frac{1}{2}$	1	1/200	0.98	0.32	0.02
Cauchy	50	$\frac{1}{2}$	$\frac{1}{2}$	1/200	0.95	0.60	0.28

e.g., [23][27]. If $\Gamma = 1$, then equality in (1.2); otherwise bounds (1.2) widen as Γ increases. For $H_0 : \tau = \tau_0$ vs $H_A : \tau > \tau_0$, the upper bound on the one-sided significance level is at most 0.05 for all π_{ij} satisfying (1.1) if $W_{\tau_0} \geq \widetilde{w}$ where $0.05 = \Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$.

For each $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{I2})$, there is an HL estimate $\widehat{\tau}_{\boldsymbol{\pi}}$ (essentially) solving $W_{\widehat{\tau}} = \mu_{\boldsymbol{\pi}}$ where the expectation $\mu_{\boldsymbol{\pi}} = E_{\boldsymbol{\pi}}\left(W_{\tau} \mid \mathcal{Z}, \mathcal{F}\right)$ is computed using $\boldsymbol{\pi}$. Then (1.1) implies $(1-\theta)\, I\, (I+1)/2 \leq \mu_{\boldsymbol{\pi}} \leq \theta\, I\, (I+1)/2$, yielding an interval of HL point estimates, $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$. With $\Gamma = 1$, $\mu_{\boldsymbol{\pi}} = I\,(I+1)/4$, and $\widehat{\tau}_{\min} = \widehat{\tau}_{\max}$ is the usual HL estimate.

2. Design Sensitivity

2.1. **Question.** In the fortunate situation, biases are confined to observed covariates, and adjustments remove these biases, yielding unbiased or consistent estimates of treatment effects. In an observational study, even if the fortunate situation arose, we would not know this from the data. In the fortunate situation, we hope to report insensitivity to small or moderate unobserved biases. How do aspects of study design affect the chance that this hope will be realized? Does a dose-response relationship strengthen causal claims [15] [41] [26] [29]? (§2.5) Does multivariate coherence or pattern specificity strengthen causal claims [15] [6] [40] [8] [26] [27] [37]? (§2.5) Does reducing heterogeneity strengthen causal claims? Is reducing heterogeneity any better than increasing the sample size? Or do both just reduce the standard error of a biased estimate [20] [31]? (§2.2-2.3) Is a weak but nearly valid instrumental variable better or worse than a stronger but possibly somewhat biased instrument [34] [4] [39]? (§2.6)

2.2. **Heterogeneity: Power of a sensitivity analysis** [30] [31]**.** For a fixed $\Gamma \geq 1$, the *power of the sensitivity analysis* is the probability that the upper bound on the significance level from (1.2) is less than, say, 0.05. Determine $\widetilde{w}$ so $0.05 = \Pr\left(\overline{\overline{W}} \geq \widetilde{w}\right)$ in (1.2); then calculate the probability that $W_{\tau_0} \geq \widetilde{w}$ under some specific alternative hypothesis. For $\Gamma = 1$, this is the usual concept of power. If the treatment has an effect and there is no hidden bias, we would not know this, but hope to report results that are insensitive to unobserved bias; the power is the probability this will happen. Therefore, we compute the power under models in which the treatment has an effect (e.g., additive effect τ) and there is no hidden bias (e.g., random assignment within pairs).

2.3. **Heterogeneity: Limiting Uncertainty** [31]**.** Whether or not (1.1) is true, for each fixed $\Gamma \geq 1$, as $I \to \infty$, the range of HL estimates, $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$, converges in probability

Table 2. Dose Response and Coherence. Left: Power at $\Gamma = 2$, $I = 200$, $k = 3$. Right: Design Sensitivity $k = 5$. k controls per matched set, p outcomes with correlation ρ. Stratified rank sum statistic.

$I = 200$		Power	at $\Gamma = 2$	$I \to \infty$		$\widetilde{\Gamma}$ =Design	Sensitivity
$k = 3$		$\rho = 0$	$\rho = \frac{1}{2}$	$k = 5$		$\rho = 0$	$\rho = \frac{1}{2}$
Doses	p			Doses	p		
$\left(\frac{1}{2}, 1, \frac{3}{2}\right)$	1	0.54	0.54	$\left(\frac{1}{2}, 1, \frac{3}{2}\right)$	1	3.0	3.0
	3	1.00	0.92		3	6.4	3.8
$(1, 1, 1)$	1	0.28	0.28	$(1, 1, 1)$	1	2.6	2.6
	3	1.00	0.73		3	5.1	3.2
$\left(\frac{3}{2}, \frac{3}{2}, \frac{3}{2}\right)$	1	0.98	0.98	$\left(\frac{3}{2}, \frac{3}{2}, \frac{3}{2}\right)$	1	4.1	4.1
	3	1.00	1.00		3	11.7	5.6

to $[\tau_{\min}, \tau_{\max}]$, with $\tau_{\max} = \tau_{\min}$ if $\Gamma = 1$ and $\tau_{\max} > \tau_{\min}$ if $\Gamma > 1$. If (1.1) *were true* with $\Gamma = 1$, then $\tau = \tau_{\max} = \tau_{\min}$; that is, the HL estimate $\widehat{\tau} = \widehat{\tau}_{\min} = \widehat{\tau}_{\max}$ is consistent for τ in a randomized experiment. If (1.1) *were true* with a specific $\Gamma > 1$, then $\tau \in [\tau_{\min}, \tau_{\max}]$, but the uncertainty about $\boldsymbol{\pi}$ prevents a more precise statement even as $I \to \infty$. Let $\Phi(\cdot)$ and $\Upsilon(\cdot)$ be, respectively, the standard Normal and standard Cauchy cumulative distributions. Proposition 1 indicates what a sensitivity analysis yields, as $I \to \infty$, when, unknown to us, there actually is no unobserved bias: the length of the limiting interval $[\tau_{\min}, \tau_{\max}]$ is strongly affected by the heterogeneity of the experimental units σ.

Proposition 1. [31] *If* $(D_i - \tau)/\sigma \sim_{iid} \Phi(\cdot)$ *then* $[\tau_{\min}, \tau_{\max}]$ *is* $\tau \pm \sigma\, \Phi^{-1}(\theta)/\sqrt{2}$, *where* $\theta = \Gamma/(1+\Gamma)$. *If* $(D_i - \tau)/\sigma \sim_{iid} \Upsilon(\cdot)$ *then* $[\tau_{\min}, \tau_{\max}]$ *is* $\tau \pm \sigma\, \Upsilon^{-1}(\theta)$.

2.4. **Design sensitivity** [30]**.** There is a value $\widetilde{\Gamma}$ of Γ such that the power of the sensitivity analysis tends to 1 for $\Gamma < \widetilde{\Gamma}$ and to 0 for $\Gamma > \widetilde{\Gamma}$; this value, $\widetilde{\Gamma}$, is called the *design sensitivity*. No matter how large the sample size becomes, the design will always be sensitive to biases larger than $\widetilde{\Gamma}$. Akin to Pitman efficiency — competing designs or methods compared for same task in large samples.

2.5. **Design sensitivity: dose-response and coherence** [30]**.** I matched sets, each with one treated person with dose v_i matched to k untreated controls with dose zero. Will consider power for $\Gamma = 2$, $k = 3$, $I = 200$, and design sensitivity for $k = 5$, $I \to \infty$, and will use the stratified Wilcoxon rank sum, with dose weights and coherence among outcomes; details in [30]. Power evaluated with $p = 1$ or $p = 3$ outcomes, linearly related to dose with slope $\beta = \frac{1}{2}$, symmetrically correlated multivariate Normal errors, with variances 1 and intercorrelations ρ. Consider three possible patterns: $\left(\frac{1}{2}, 1, \frac{3}{2}\right)$, $(1, 1, 1)$, $\left(\frac{3}{2}, \frac{3}{2}, \frac{3}{2}\right)$, where $\left(\frac{1}{2}, 1, \frac{3}{2}\right)$ yields dose-response, $(1, 1, 1)$ has the same average dose without dose-response, and $\left(\frac{3}{2}, \frac{3}{2}, \frac{3}{2}\right)$ has larger doses without dose response.

2.6. **Design sensitivity: instrument strength and validity** [39]**.** Instrument strength characterized by compliance probabilities: π_A, π_C, π_N for, respectively, the proportion of 'always takers' $(v_{Tij}, v_{Cij}) = (1,1)$, 'compliers' $(v_{Tij}, v_{Cij}) = (1,0)$, and 'never takers' $(v_{Tij}, v_{Cij}) = (0,0)$, with no defiers $(v_{Tij}, v_{Cij}) = (0,1)$.

Design Sensitivity $\widetilde{\Gamma}$ For Instruments with Varying Strength. Effect size $\lambda = (\beta_0 - \beta)/\sigma$.

$(v_{Tij}, v_{Cij}) = (1,0)$

Compliance		100%	50%	20%	10%
π_A, π_C, π_N		$0,1,0$	$\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$	$\frac{2}{5}, \frac{1}{5}, \frac{2}{5}$	$\frac{9}{20}, \frac{2}{20}, \frac{9}{20}$
ϵ_i	λ				
Normal	1	11.7	2.7	1.5	1.2
Normal	$\frac{1}{2}$	3.2	1.7	1.2	1.1
Cauchy	1	3.0	1.7	1.2	1.1
Cauchy	$\frac{1}{2}$	1.8	1.4	1.1	1.1

References

[1] Aakvik A. (2001). Bounding a matching estimator: the case of a Norwegian training program. *Oxford Bull. Econ. Statist.* **63**, 115-43. *A sensitivity analysis.*

[2] Angrist, J., Imbens, G., and Rubin, D. (1996), "Identification of causal effects using instrumental variables," *JASA*, **91**, 444-469. *Links IV with randomization.*

[3] Angrist, J. and Krueger, A. B. (1994), "Why do World War II veterans earn more than nonveterans?" *J. Labor Econ.*, 12, 74-97. *Example in talk.*

[4] Angrist, J. D. and Krueger, A. B. (2001), "Instrumental variables and the search for identification," *J. Econ. Perspec.*, 15, 69-85. *Advice about design.*

[5] Ashenfelter, O. & Rouse, C. (1998), "Income, schooling and ability: Evidence from a new sample of identical twins," *Quart. J. .Econ.*, 113, 253-284. *Example of reducing heterogeneity: twins.*

[6] Campbell, D. T. (1988) *Methodology and Epistemology for Social Science: Selected Papers.* Chicago: University of Chicago Press, pp315-333. *Advice about design.*

[7] Card, D. & Krueger, A. (1994), "Minimum wages and employment," *Am. Econ. Rev.,* 84 772-793. *Example of reducing heterogeneity: retail chains.*

[8] Cook, T. D. & Shadish, W. R. (1994), "Social experiments: Some developments over the past fifteen years," *Ann. Rev. Psych.* **45**, 545-80. *Advice about design.*

[9] Copas, J. & Eguchi, S. (2001), "Local sensitivity approximations for selectivity bias," *JRSS,* B 63, 871-96. *Alternative sensitivity analysis.*

[10] Cornfield, J., et al. (1959), "Smoking and lung cancer," *J. Nat. Cancer Inst.,* 22, 173-203. *Alternative sensitivity analysis.*

[11] Fisher, R. A. (1935), *Design of Experiments*, Edinburgh: Oliver and Boyd. *Invention of randomized experimentation.*

[12] Gastwirth, J. L., Krieger, A. M. & Rosenbaum, P. R. (1998), "Dual and simultaneous sensitivity analysis for matched pairs," *Biometrika*, 85, 907-920. *Alternative sensitivity analysis.*

[13] Greevy, R., Silber, J. H., Cnaan, A. & Rosenbaum, P.R. (2004), "Randomization inference with imperfect compliance in the AAA randomized trial," *JASA*, 99, 7-15. *Randomization inference with IV.*

[14] Hamermesh, D. S. (2000), "The craft of labormetrics," *Indust. Labor Relat. Rev.*, 53, 363-380. *Advice about design.*

[15] Hill, A. B. (1965), "The environment and disease: Association or causation?" *Proc. R. Soc. Med.* 58, 295-300. *Advice about design.*

[16] Imbens, G. W. (2003), "Sensitivity to exogeneity assumptions in program evaluation," *Am. Econ. Rev.*, 93, 126-132. *Alternative sensitivity analysis.*

[17] Imbens, G. and Rosenbaum, P. R. (2005), "Robust, accurate confidence intervals with a weak instrument," *J. Roy. Statist. Soc.*, A1**68**, 109-126. *Randomization inference with IV.*

[18] Lehmann, E. L. (1998), *Nonparametrics,* NJ: Prentice Hall. *Randomization inference, nonparametrics.*

[19] Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998), "Assessing the sensitivity of regression results to unmeasured confounders," *Biometrics*, 54, 948-963. *Alternative sensitivity analysis.*

[20] Mill, J. S. (1867), *A System of Logic*, Indianapolis. *Reducing heterogeneity (method of difference).*

[21] Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Stat. Sci.*, 5, 463-480. *Causal effects.*

[22] Norvell, D. C. & Cummings, P. (2002), "Association of helmet use with death in motorcycle crashes: A matched-pair cohort study," *Am. J. Epidem.*, 156, 483-487. *Example of reducing heterogeneity: two riders of one motorcycle.*

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

[24] Rosenbaum, P. R. (1993), "Hodges-Lehmann point estimates of treatment effect in observational studies," *JASA,* 88 1250-1253. *Sensitivity analysis.*

[25] Rosenbaum, P. R. (1996), "Comment," *JASA* **91**, 465-468. *Randomization inference with IV.*

[26] Rosenbaum, P. R. (1997), "Signed rank statistics for coherent predictions," *Biometrics*, 53, 556-566. *Coherence, doses.*

[27] Rosenbaum, P. R. (2002), *Observational Studies* (2^{nd} ed). NY: Springer. *General.*

[28] Rosenbaum, P.R. (2002), "Covariance adjustment in randomized experiments and observational studies (with Discussion)," *Stat. Sci.,* 17, 286-327. *Randomization inference with IV.*

[29] Rosenbaum, P. R. (2003), "Does a dose-response relationship reduce sensitivity to hidden bias?" *Biostatistics*, 4, 1-10. *Doses.*

[30] Rosenbaum, P. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-164. *Basis for this talk.*

[31] Rosenbaum, P. R. (2005), "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies," *Am. Statist.*, 59, 147-152. *Basis for this talk.*

[32] Rosenbaum, P. R. (2007), "Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies," *Biometrics*, 63, 456-464. *Sensitivity analysis.*

[33] Rosenbaum, P. R. & Rubin, D. B. (1983), "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *JRSS,* B **45**, 212–8. *Alternative sensitivity analysis.*

[34] Rosenzweig, M. R. and Wolpin, K. I. (2000) Natural "natural experiments" in economics. *J. Econ. Lit.*, **38**, 827–874. *Advice about design.*

[35] Rothman, K. J. (1986), *Modern Epidemiology,* Boston: Little, Brown. *Critical of Hill 1965.*

[36] Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Ed. Psych.*, 66, 688-701. *Causal effects.*

[37] Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin. *Advice about design.*

[38] Sierra-Torres, M., et al. (2004), "Chromosome aberrations among cigarette smokers in Columbia," *Mutation Research*, 562, 67-75. *Example in talk.*

[39] Small, D. and Rosenbaum, P. R. (2007), "War and wages: The strength of instrumental variables and their sensitivity to unobserved biases," *JASA*, to appear. *Basis for this talk.*

[40] Trochim, W. M. K. (1985), "Pattern matching, validity and conceptualization in program evaluation," *Eval. Rev.*, 9, 575-604. *Advice about design.*

[41] Weiss, N. (1981). Inferring causal relationships: Elaboration of the criterion of 'dose-response,' *Am. J. Epidem.,* **113**, 487–90. *Advice about design.*

[42] Wright, P. H., & Robertson, L. S. (1976), "Priorities for roadside hazard modification," *Traffic Engineering*, 46, 24-30. *Clever example of reducing heterogeneity.*