War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases

Dylan S. SMALL and Paul R. ROSENBAUM

An instrument manipulates a treatment that it does not entirely control, but the instrument affects the outcome only indirectly through its manipulation of the treatment. The idealized prototype is the randomized encouragement design, in which subjects are randomly assigned to receive either encouragement to accept the treatment or no such encouragement, but not all subjects comply by doing what they are encouraged to do, and the situation is such that only the treatment itself, not disregarded encouragement alone, can affect the outcome. An instrument is weak if it has only a slight impact on acceptance of the treatment, that is, if most people disregard encouragement to accept the treatment. Typical applications of instrumental variables are not ideal; encouragement is not randomized, although it may be assigned in a far less biased manner than the treatment itself. Using the concept of design sensitivity, we study the sensitivity of instrumental variable analyses to departures from the ideal of random assignment of encouragement, with particular reference to the strength of the instrument. With these issues in mind, we reanalyze a clever study by Angrist and Krueger concerning the effects of military service during World War II on subsequent earnings, in which birth cohorts of very similar but not identical age were differently "encouraged" to serve in the war. A striking feature of this example is that those who served earned more, but the effect of service on earnings appears to be negative; that is, the instrumental variables analysis reverses the sign of the naive comparison. For expository purposes, this example has the convenient feature of enabling, by selecting different birth cohorts, the creation of instruments of varied strength, from extremely weak to fairly strong, although separated by the same time interval and thus perhaps similarly biased. No matter how large the sample size becomes, even if the effect under study is quite large, studies with weak instruments are extremely sensitive to tiny biases, whereas studies with stronger instruments can be insensitive to moderate biases.

KEY WORDS: Design sensitivity; Observational study; Sensitivity analysis.

1. WORLD WAR II VETERAN STATUS AND EARNINGS

1.1 Who Served? What Effects Did Service Have on Earnings?

As an illustration of instrumental variables methods, we perform certain alternative analyses of a very nice study by Angrist and Krueger (1994) that concerned the effects on male earnings of serving in the military during World War II (WWII). Does military service raise or lower earnings? For a specific man, military service in WWII might interrupt education or career and so cause that man to earn less than he would have if he had not served. Alternatively, for a specific man, if the labor market favored veterans, or if various veteran's programs conferred advantages, then military service in WWII might cause that man to earn more than he would have had he not served. The familiar distinction between association and causation is that these questions concern the effects caused by military service for the same man; they do not simply compare the different men who happened to be WWII veterans and those who happened to not be WWII veterans. Men may be rejected for military service for reasons of ill health or criminal behavior, and others may legally avoid or illegally evade military service, so that we would expect veterans and nonveterans of WWII to differ in earnings quite apart from any effect caused by military service.

Angrist and Krueger based their analysis on the 5% public use sample of the 1980 U.S. Census. To preserve confidentiality, the Census data do not contain dates of birth, but they do contain years and quarters of birth. Census data have certain limitations that affect the original analysis by Angrist and Krueger and our reanalysis in parallel ways. The Census data describe individuals who responded to the Census, and to do that, a man had to be alive and a respondent to the long form. Moreover, a man's description in 1980 of his childhood in the first half of the century is accepted as accurate, even though it may contain errors, and conceivably veteran status could affect these errors. The information about childhood that we use is quarter of birth, race, age, education up to 8 years, and location of birth. If veteran status altered recall or reporting of this information, then these errors would not be evident in the Census data. Serving in the military during the WWII affected survival, directly in the form of trauma during the war and also in subtler ways, such as boosting tobacco use (Bedard and Deschênes 2006), but it is unclear whether and to what extent survival was related to potential earnings. There have been some studies of twin pairs who served in WWII (e.g., Taubman 1976), but they have not addressed the issue of the earnings of the co-twin of men killed in the war. With Census data alone, the possibility cannot be eliminated that one of these limitations has distorted income comparisons.

1.2 A Picture of the Instrumental Variables Argument

Figure 1 depicts the 1980 earnings of 14,000 men born in the second half of 1926, or in quarter 3 or 4 of 1926 (henceforth Q3 or Q4), by World War II veteran status. The left boxplot for WWII veterans describes 10,571 men, and the right boxplot describes the remaining 3,429 men, so $10{,}571/14{,}000 = 75.5\%$ of these men were WWII veterans. The maximum earnings recorded by the Census in the microdata is \$75,000, a fairly high level in 1980, so the top earnings and the mean earnings are distorted, but the quartiles are not. A total of 1.5% of men were recorded as having censored incomes of \$75,000+. Table 1 gives the quartiles and the trimean (i.e., twice the median plus the extreme quartiles divided by 4). For men born in the second half of 1926, the WWII veterans earned about \$4,500

Dylan S. Small is Assistant Professor (E-mail: *dsmall@wharton.upenn.edu*) and Paul R. Roosenbaum is Professor (E-mail: *rosenbaum@stat.wharton.upenn.edu*), Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. This work was supported by grant SES-0646002 from the Methodology, Measurement, and Statistics Program and the Statistics and Probability Program of the National Science Foundation.

Journal of the American Statistical Association
September 2008, Vol. 103, No. 483, Applications and Case Studies
DOI 10.1198/016214507000001247

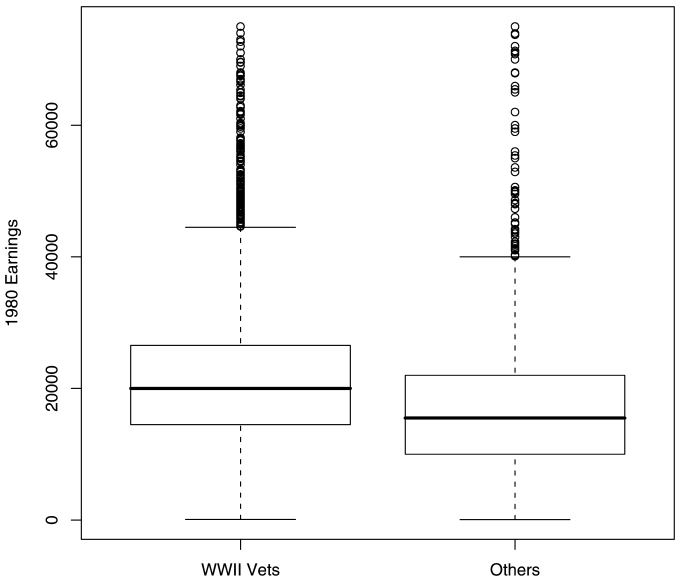

Figure 1. Earning of men born in 1926 Q3 or Q4.

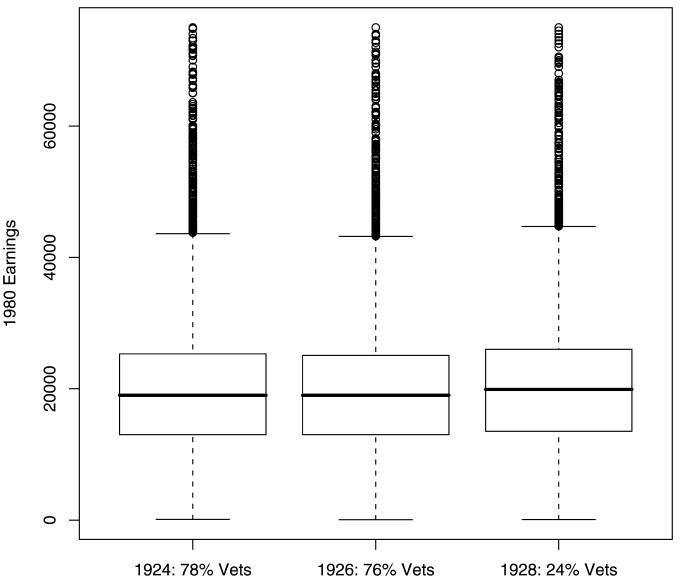

Figure 2. Earnings and WWII veteran status for men born in Q3 or Q4.

more in 1980. The $4,500 difference is association, not causation: it merges the effects of military service with the nonrandom sorting of men into WWII veterans and others. To believe that the $4,500 difference is the effect caused by service in WWII is tantamount to believing that the division of men into WWII veterans and others was effectively the same as random assignment in a clinical trial.

Following the spirit (but not the details) of Angrist and Krueger's analysis, Figure 2 provides reason to doubt that military service caused an increase of $4,500 in the earnings of men born in the second half of 1926. As we discuss later, the contrast between Figures 1 and 2 depicts the instrumental variables argument.

Each boxplot in Figure 2 depicts 14,000 men born in the second half of their year of birth. The 1926 boxplot in Figure 2 merges the men and earnings in the two separate boxplots in Figure 1 and Table 1. The men in the three boxplots are matched for certain demographic variables that almost certainly were not affected by military service (i.e. matched on covariates), so the data consist of 14,000 matched triples of three men, one man born in 1924, one in 1926, and one in 1928. The matching controlled for (a) quarter of birth (Q3 versus Q4); (b) race (white, black, or other); (c) completed $\geq$8 years of education; (d) completed $\geq$7 years of education; (e) completed $\geq$6 years of education; (f) Census region of birth (Northeast, Midwest, South, West, or American territories); (g) Census division (New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, Mountain, Pacific, or the American territories); and (h) state of birth. A random sample of 7,000 men born in Q3 or Q4 of 1924 was matched to men born in Q3 or Q4 of 1926, who in turn were matched to men born in Q3 or Q4 of 1928, so the covariate distribution represents Q3 and Q4 of 1924. The 14,000 triples were formed by 28,000 pairings. Of these 28,000 pairs, more than 92% were exactly matched for all 8 covariates, and 99% were exactly matched for covariates (a)–(f), that is, for everything including the Census region except for the Census division and individual state. We did not match for education beyond 8 years, because some men may have left high school to join the military, and military service may have either disrupted or facilitated college attendance, so later education is an outcome that may be affected by the treatment (Rosenbaum 1984).

Table 1. Earnings in 1980 of 14,000 men born in the second half of 1926 by WWII veteran status

	Count	Lower quartile	Median	Upper quartile	Trimean
WWII veterans	10,571	14,510	20,010	26,540	20,268
Others	3,429	10,010	15,510	22,010	15,760
Difference		4,500	4,500	4,530	4,508

Men born in 1924, 1926, and 1928 turned 17 years old in 1941, 1943, and 1945. Men born in 1924 or 1926 were of prime age for military service in WWII, whereas those born in the last two quarters of 1928 were somewhat on the young side for military service in WWII. Indeed, in Figure 2, WWII veterans compose 78.2% of those men born in 1924, 75.5% of those born in 1926, and 24.3% of those born in 1928. Thus the two left boxplots in Figure 2 are mostly WWII veterans, whereas the boxplot on the far right shows mostly men who were not WWII veterans.

To believe that the $4,500 difference in Figure 1 is an effect caused by military service, one would need to hold rather peculiar beliefs about what happened in Figure 2. In Figure 1, the WWII veterans earned more, but in Figure 2, the mixture of 78% WWII veterans and 22% others born in 1924 earned about the same as the mixture of 76% WWII veterans and 24% others born in 1926, but both earned slightly less than a mixture of 24% WWII veterans and 76% others born in 1928. This is a bit of a surprise; if being a WWII veteran raised income by $4,500 in Table 1, then higher earnings when there were more WWII veterans would be expected. In 1980, the 1928 cohort is 2 years younger than the 1926 cohort, which is 2 years younger than the 1924 cohort, but the 1924 and 1926 cohorts earned about

the same. It is possible that Figure 1, Figure 2, or both are affected by selection biases; the men in the boxplots may differ. The selection bias in Figure 1 would be the division of men into WWII veterans and others, and because that is partially based on health and criminal record, a substantial selection bias is not implausible. In contrast, aside from 2 years of age in Figure 2, the bias would have to reflect systematic differences between the types of men born in different years.

Instrumental variables methods are widely used in economics. Two other applications of instrumental variables to study the effects of military service on earnings have been published by Angrist (1990) and Imbens and van der Klaauw (1995).

2. NOTATION AND REVIEW

2.1 The Matched Encouragement Design

The randomized encouragement design (Holland 1988) is an experimental design that serves as a prototype for the instrumental variables argument, in the specific sense that random assignment of encouragement to accept the treatment ensures that one of the assumptions of the instrumental variables argument is true. In an encouragement experiment (Holland 1988), some subjects are picked at random and encouraged to accept the treatment, but not everyone does what they are encouraged to do, and only the treatment itself, not encouragement alone, can affect the outcome. In this structure, there can be substantial selection bias in accepting the treatment, but this can removed through an instrumental variables analysis, as was carefully developed by Angrist, Imbens, and Rubin (1996) (see also Sommer and Zeger 1991; Frangakis and Rubin 1999; Tan 2006). Encouragement is an instrument for the treatment actually received. Section 2 defines notation and reviews ideas about permutation inference with instrumental variables from Rosenbaum (1996, 1999, 2002a), Greevy, Silber, Cnaan, and Rosenbaum (2004), and Imbens and Rosenbaum (2005).

There are I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, exactly matched for observed covariates, $\mathbf{x}_{ij}$, that is, for variables measured before treatment assignment and thus unaffected by treatment; exact matching ensures $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, for $i = 1, \ldots, I$. In addition to the observed $\mathbf{x}_{ij}$, there may be unobserved covariates, u_{ij}, but these cannot be controlled for by matching, so typically $u_{i1} \neq u_{i2}$. In a matched "encouragement design," one subject in each pair, denoted by $Z_{ij} = 1$, is encouraged to accept a high dose of the treatment, whereas the other subject, denoted by $Z_{ij} = 0$, is not so encouraged, so $1 = Z_{i1} + Z_{i2}$ for $i = 1, \ldots, I$. In a randomized encouragement experiment, Z_{i1} is determined by a flip of a fair coin, independently in different pairs, with $Z_{i2} = 1 - Z_{i1}$, but in an observational study, the assignments Z_{ij} may be biased so that certain subjects are more likely to receive encouragement than others.

Each subject ij has two potential responses: r_{Tij} if encouraged, $Z_{ij} = 1$, or r_{Cij} if not encouraged, $Z_{ij} = 0$, so the effect of encouragement, $r_{Tij} - r_{Cij}$, is not observed (see Neyman 1923; Rubin 1974). The observed response for ij is $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$. If encouraged, $Z_{ij} = 1$, subject ij receives treatment at dose w_{Tij}; if not encouraged, $Z_{ij} = 0$, at dose w_{Cij}, so the dose actually received is $W_{ij} = Z_{ij} w_{Tij} + (1 - Z_{ij}) w_{Cij}$. Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, w_{Tij}, w_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, 2\}$, and write $\mathcal{Z}$ for the event $\{Z_{i1} + Z_{i2} = 1, i = 1, \ldots, I\}$.

When W_{ij} is binary, Angrist et al. (1996) say that subject ij is an "always taker" if $(w_{Tij}, w_{Cij}) = (1, 1)$, a "never taker" if $(w_{Tij}, w_{Cij}) = (0, 0)$, a "complier" if $(w_{Tij}, w_{Cij}) = (1, 0)$, and a "defier" if $(w_{Tij}, w_{Cij}) = (0, 1)$. Always takers and never takers ignore encouragement, compliers do what they were encouraged to do, and defiers do the opposite of what they were encouraged to do. Angrist et al. (1996) mostly assumed there are no defiers.

In Fisher's (1935) theory of randomization inference in experiments, quantities that depend on the randomly assigned treatment, Z_{ij} (e.g., the observed response, R_{ij}, and the observed treatment, W_{ij}), are random variables, but the covariates and potential responses, $\mathcal{F}$, are fixed features of the finite population of $2I$ subjects. To speak of a quantity as fixed is to say that all probabilities are implicitly conditional probabilities given the values of fixed quantities. In particular, random assignment of encouragement within pairs ensures that $\Pr(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ for all i and j, with independent assignments in distinct pairs. Write $\mathbf{Z} = (Z_{11}, Z_{12}, Z_{21}, \ldots, Z_{I2})^T$ for the $2I$-dimensional vector of treatment assignments, and let Ω be the set of 2^I possible values $\mathbf{z}$ of $\mathbf{Z}$. In a randomized encouragement design, $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}, \mathcal{Z}) = 1/2^I$ for each $\mathbf{z} \in \Omega$. Write $\mathbf{R} = (R_{11}, \ldots, R_{I2})^T$, $\mathbf{r}_C = (r_{C11}, \ldots, r_{CI2})^T$, and so on.

2.2 Randomization Inference in a Randomized Encouragement Design

A simple model says that the effect of encouragement on response is proportional to its effect on the treatment received,

$$r_{Tij} - r_{Cij} = \beta(w_{Tij} - w_{Cij}), \tag{1}$$

where in the current study, w_{Tij} and w_{Cij} are each binary but (1) is also applicable in other studies with continuous doses of treatment. If (1) is true, then

$$R_{ij} - \beta W_{ij} = r_{Tij} - \beta w_{Tij} = r_{Cij} - \beta w_{Cij} = a_{ij}, \quad \text{say,} \tag{2}$$

takes the same value whether or not the jth subject in pair i is encouraged (i.e., whether or not $Z_{ij} = 1$ or $= 0$), so a_{ij} is a function of $\mathcal{F}$ and is fixed. In a randomized encouragement experiment, exact randomization inferences about β in (1) are obtained in the following way. Consider testing $H_0 : \beta = \beta_0$ using the observed quantity $R_{ij} - \beta_0 W_{ij}$, which, using (2), is

$$\begin{aligned}
&R_{ij} - \beta_0 W_{ij} \\
&\quad = Z_{ij}(r_{Tij} - \beta_0 w_{Tij}) + (1 - Z_{ij})(r_{Cij} - \beta_0 w_{Cij}) \\
&\quad = Z_{ij}\{(r_{Tij} - \beta w_{Tij}) + (\beta - \beta_0) w_{Tij}\} \\
&\qquad + (1 - Z_{ij})\{(r_{Cij} - \beta w_{Cij}) + (\beta - \beta_0) w_{Cij}\} \\
&\quad = a_{ij} + (\beta - \beta_0)\{Z_{ij} w_{Tij} + (1 - Z_{ij}) w_{Cij}\},
\end{aligned}$$

so $R_{ij} - \beta_0 W_{ij}$ will be fixed at a_{ij} if $H_0 : \beta = \beta_0$ is true, but otherwise $R_{ij} - \beta_0 W_{ij}$ will tend to vary with Z_{ij}. Write $D_i^{\beta_0}$ for

the matched pair difference in $R_{ij} - \beta_0 W_{ij}$, encouraged ($Z_{ij} = 1$) minus control ($Z_{ij} = 0$), so that

$$\begin{aligned} D_i^{\beta_0} &= Z_{i1}\{(R_{i1} - \beta_0 W_{i1}) - (R_{i2} - \beta_0 W_{i2})\} \\ &\quad + (1 - Z_{i1})\{(R_{i2} - \beta_0 W_{i2}) - (R_{i1} - \beta_0 W_{i1})\} \\ &= (\beta - \beta_0)\{Z_{i1}(w_{Ti1} - w_{Ci2}) \\ &\quad + (1 - Z_{i1})(w_{Ti2} - w_{Ci1})\} + (2Z_{i1} - 1)(a_{i1} - a_{i2}) \\ &= (\beta - \beta_0) S_i + \epsilon_i, \end{aligned} \tag{3}$$

where

$$\begin{aligned} S_i &= Z_{i1}(w_{Ti1} - w_{Ci2}) + (1 - Z_{i1})(w_{Ti2} - w_{Ci1}) \qquad \text{and} \\ \epsilon_i &= (2Z_{i1} - 1)(a_{i1} - a_{i2}). \end{aligned} \tag{4}$$

If $H_0 : \beta = \beta_0$ is true in a randomized encouragement experiment, then $D_i^{\beta_0}$ is $\pm(a_{i1} - a_{i2})$ each with probability $\frac{1}{2}$, independently in different pairs, with $|D_i^{\beta_0}|$ fixed, so $D_i^{\beta_0}$ is symmetrically distributed about 0; therefore, the conditional distribution given $\mathcal{F}$ and $\mathcal{Z}$ of Wilcoxon's signed-rank statistic, say T^{β_0}, computed from $D_i^{\beta_0}$, has its usual null distribution. On the other hand, in a randomized encouragement design, if $\beta > \beta_0$, and if encouragement to take a higher dose tends to raise the dose received (e.g., if $w_{Tij} \geq w_{Cij}$ with some strict inequalities, then $D_i^{\beta_0}$ is the sum of a quantity $(2Z_{i1} - 1)(a_{i1} - a_{i2})$ symmetrically distributed about 0 and a quantity $(\beta - \beta_0)\{Z_{i1}(w_{Ti1} - w_{Ci2}) + (1 - Z_{ij})(w_{Ti2} - w_{Ci1})\}$ with positive expectation. The set of β_0 not rejected by the test forms a confidence set for β. The Hodges–Lehmann point estimate $\widehat{\beta}$ of β is formed by equating the test statistic to its null expectation and solving this estimating equation for $\widehat{\beta}$. If the treatment is binary and compliance with encouragement is perfect ($w_{Tij} = 1, w_{Cij} = 0, \forall i, j$), then $r_{Tij} - r_{Cij} = \beta$ in (1) is an additive treatment effect, and the procedures just described yield conventional randomization inferences in a paired randomized experiment. Greevy et al. (2004) obtained randomization inferences in a randomized encouragement experiment of a drug intended to preserve cardiac performance after chemotherapy for cancer.

The procedure need not use Wilcoxon's signed-rank statistic. Instead, it could use the randomization distribution of a combined quantile average (Rosenbaum 1999, sec. 5) or the sample mean of the $D_i^{\beta_0}$ (Imbens and Rosenbaum 2005), yielding the Wald (1940) estimator or, equivalently, two-stage least squares,

$$\widehat{\beta}_W = \frac{\sum_{i=1}^{I}(2Z_{i1} - 1)(R_{i1} - R_{i2})}{\sum_{i=1}^{I}(2Z_{i1} - 1)(W_{i1} - W_{i2})}. \tag{5}$$

Encouragement is a *weak instrument* if encouragement has very little impact on the dose of treatment, so that $w_{Tij} - w_{Cij}$ is 0 or near 0 for most subjects ij. With a weak instrument, in (3), even when β_0 is far from β, the matched pair difference $D_i^{\beta_0}$ is nearly symmetric about 0, so the data contain limited information about β. In the extreme, if encouragement is always ignored, $w_{Tij} = w_{Cij}$, then $D_i^{\beta_0}$ in (3) is symmetric about 0 for every β_0, so there is no information about β. The randomization inference correctly reflects this, with the confidence interval for β maintaining nominal coverage with weak instruments by becoming appropriately longer, perhaps infinite in length (see Imbens and Rosenbaum 2005 for detailed discussion). In contrast, the usual confidence interval associated with two-stage least squares performs very poorly with weak instruments, often having coverage much lower than the nominal level (see Bound, Jaeger, and Baker 1995). The formula (5) hints at the nature of the problem; when the instrument is weak, the denominator estimates a quantity close to 0, so $\widehat{\beta}_W$ and its "plug-in" estimate of its standard error are both highly unstable.

2.3 Sensitivity to Nonrandom Assignment of Encouragement in Observational Studies

In nonrandomized or observational studies, assignment to encouragement, $Z_{ij} = 1$, or not, $Z_{ij} = 0$, is not determined by the flip of a coin, and there may be little basis for believing that $\Pr(Z_{ij} = 1|\mathcal{F}, \mathcal{Z}) = \frac{1}{2}$. A sensitivity analysis asks how departures from random assignment of various magnitudes might alter a study's conclusions. One model for sensitivity analysis begins by assuming that in the population before matching on $\mathbf{x}_{ij}$, encouragements Z_{ij} are assigned independently with unknown probabilities, $\pi_{ij} = \Pr(Z_{ij} = 1|\mathcal{F})$, such that two subjects, say subjects ij and ik, who might be matched because they have the same value of the observed covariates, $\mathbf{x}_{ij} = \mathbf{x}_{ik}$, may differ in their odds of receiving the treatment by at most a factor of $\Gamma \geq 1$ because they differ in terms of an unobserved covariate $u_{ij} \neq u_{ik}$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ik})}{\pi_{ik}(1 - \pi_{ij})} \leq \Gamma \qquad \forall i, j, k. \tag{6}$$

Then, after pair matching of encouraged subjects ($Z_{ij} = 1$) to controls ($Z_{ij} = 0$) to ensure that $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, the distribution of $\mathbf{Z}$ in Ω is obtained by conditioning on the event $\mathcal{Z}$. Write $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \ldots, \pi_{I2})^T$. The model (6) is equivalent to a logit model of the form $\log\{\pi_{ij}/(1 - \pi_{ij})\} = \phi(\mathbf{x}_{ij}) + \gamma u_{ij}$, with $0 \leq u_{ij} \leq 1$, $\gamma = \log(\Gamma)$ and $\phi(\cdot)$ an arbitrary function (see Rosenbaum 2002b, sec. 4), and under certain assumptions, Wang and Krieger (2006) argued that the scale restriction $0 \leq u_{ij} \leq 1$ is conservative, in the sense that a binary unobserved u_{ij} creates greater sensitivity to bias than does any other unobserved u_{ij} with the same standard deviation. Model (6) is similar in spirit, although different in detail, to the model in the first sensitivity analysis of Cornfield et al. (1959). In (6), Γ is an unknown sensitivity parameter that is varied systematically to display the sensitivity of an inference to departures from random assignment. If $\Gamma = 1$ in (6), then this yields the randomization distribution, $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = 1/2^I$, for $\mathbf{z} \in \Omega$. If $\Gamma > 1$, then the distribution of treatment assignments, $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z})$, is unknown, but the magnitude of the departure from random assignment is controlled by the value of Γ. Consider an inference quantity, such as a significance level, a point estimate, or one endpoint of a confidence interval. Each value of $\boldsymbol{\pi}$ compatible with (6) yields a value of the inference quantity, say a significance level. For several values of $\Gamma \geq 1$, a sensitivity analysis computes the maximum and minimum values of the inference quantity for all $\boldsymbol{\pi}$ compatible with (6), say the range of possible significance levels. For instance, in Hammond's (1964) study of heavy smoking as a possible cause of lung cancer, the range of possible significance levels for $\Gamma = 5$ was [<.0001, .03], so a bias of magnitude $\Gamma = 5$ is insufficient to explain away, as not

causal, the association between heavy smoking and lung cancer, but for $\Gamma = 6$, the range is $[<.0001, .1]$, which includes values above the conventional .05 level. In words, even if subjects matched for observed covariates differed in terms of unobserved covariates to the extent that one might have $\Gamma = 5$ times higher odds of smoking than the other, and even if these unobserved covariates were excellent predictors of lung cancer, then the association between smoking and lung cancer is too strong to be dismissed as not an effect of smoking. Compared with many other observational studies, this is a high degree of insensitivity to unobserved biases. In general, the sensitivity analysis reveals the magnitude of unobserved bias that would need to be present to alter the conclusions of an observational study. (For a brief discussion of this approach to sensitivity analysis with specific reference to matched pairs, see Rosenbaum 1993, 1999, and for detailed discussion, including alternative but equivalent formulations of the model, varied statistical procedures, and many examples, see Rosenbaum 2002b, sec. 4. For alternative models and methods of sensitivity analysis, see, e.g., Copas and Eguchi 2001; Cornfield et al. 1959; Gastwirth 1992; Imbens 2003; Lin, Psaty, and Kronmal 1998; Robins, Rotnitzky, and Scharfstein 1999; Wang and Krieger 2006.)

Sensitivity analysis for Wilcoxon's signed-rank statistic, T^{β_0}, is straightforward (for details, see Rosenbaum 2002b, sec. 4.3). Let $\overline{\overline{T}}$ be the sum of I independent random variables taking the values 0 or i with probabilities $1/(1+\Gamma)$ or $\Gamma/(1+\Gamma)$ for $i = 1, \ldots, I$. In parallel, let $\overline{T}$ be the sum of I independent random variables taking the values 0 or i with probabilities $\Gamma/(1+\Gamma)$ or $1/(1+\Gamma)$ for $i = 1, \ldots, I$. It is straightforward to show that under the null hypothesis $H_0: \beta = \beta_0$, for all $\boldsymbol{\pi}$ satisfying (6),

$$\Pr(\overline{T} \geq t) \leq \Pr(T^{\beta_0} \geq t | \mathcal{F}, \mathcal{Z}) \leq \Pr(\overline{\overline{T}} \geq t), \tag{7}$$

which yields bounds on significance levels and thus also bounds on confidence intervals. The bounds in (7) are sharp; they are attained for particular $\boldsymbol{\pi}$ satisfying (6). For $\Gamma = 1$, the inequalities in (7) become equalities and yield the familiar null randomization distribution of Wilcoxon's signed-rank statistic. For a specific Γ, let $\widetilde{t}_\alpha$ be the value such that $\alpha = \Pr(\overline{\overline{T}} \geq \widetilde{t}_\alpha)$, so if $T^{\beta_0} \geq \widetilde{t}_\alpha$ then the upper bound on the significance level is less than or equal to α for this Γ, so β_0 is rejected at the α level. (Typically, we take $\alpha = .025$ in one tail and $\alpha = .05$ in two tails.) S–PLUS code for the exact distributions in (7) was given by Rosenbaum (2003), but for most purposes, a large-sample approximation suffices. The expectation and variance of $\overline{\overline{T}}$ and $\overline{T}$ are $E(\overline{\overline{T}}) = \zeta I(I+1)/2$, $E(\overline{T}) = (1-\zeta)I(I+1)/2$, $\text{var}(\overline{\overline{T}}) = \text{var}(\overline{T}) = \zeta(1-\zeta)I(I+1)(2I+1)/6$, where $\zeta = \Gamma/(1+\Gamma)$, which reduce to the familiar null expectation and variance of the signed-rank statistic when $\Gamma = 1$. As $I \to \infty$, the bounds in (7) may be approximated using the central limit theorem, comparing the deviates $\{T^{\beta_0} - E(\overline{\overline{T}})\}/\sqrt{\text{var}(\overline{\overline{T}})}$ and $\{T^{\beta_0} - E(\overline{T})\}/\sqrt{\text{var}(\overline{T})}$ to the standard normal distribution. By redefining $\overline{\overline{T}}$ and $\overline{T}$ slightly, it is straightforward to allow for ties in the responses. The calculations are illustrated in Section 3.

3. EXAMPLE: SOME EMPIRICAL RESULTS

3.1 Analysis Assuming That the Instrument Is Strictly Valid

If the instrument is strictly valid in the sense that $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = 1/2^I$ for $\mathbf{z} \in \Omega$ and (1) are true, then inference about β may be based on comparing Wilcoxon's signed-rank statistic, T^{β_0}, with its usual null reference distribution to test $H_0: \beta = \beta_0$. Assuming that the instrument is valid, Figure 3 plots the standardized deviate for testing $H_0: \beta = \beta_0$ against β_0 for 2 groups of 14,000 pairs of men in Figure 2; the 1928 versus 1926 pairs, where birth year is a strong instrument for service in WWII, and the 1924 versus 1926 pairs, where birth year is a very weak instrument. Figure 3 also plots horizontal lines at ± 1.96, the upper and lower .025 percentiles of the standard normal distribution. The plots of deviate values cross the horizontal lines at the endpoints of the large-sample 95% confidence limits.

Consider first the strong instrument, the 1928 versus the 1926 pairs. The deviate curve crosses ± 1.96 at $-1{,}445$ and -500, so the 95% confidence interval for the effect of WWII service is $[-1{,}445, -500]$; thus service is estimated to have depressed earnings by between \$500 and \$1,445. This negative treatment effect is in sharp contrast to the positive association of earnings with WWII service shown in Figure 1 and Table 1. The deviate for testing the hypothesis suggested by Figure 1, $H_0: \beta = \$4{,}500$, is 22.37, so it falls in the extreme tail of the standard normal distribution. If the instrument were valid for the 1928 versus 1926 pairs, then it would be very clear that the difference in Figure 1 is generated by selection bias in service in WWII, and that this service reduced earnings.

The second deviate curve in Figure 3 is for the 1924 versus 1926 pairs, where the instrument is very weak, because the fraction of men serving in these two years is very similar. The second deviate curve does not cross ± 1.96 in Figure 3, so none of

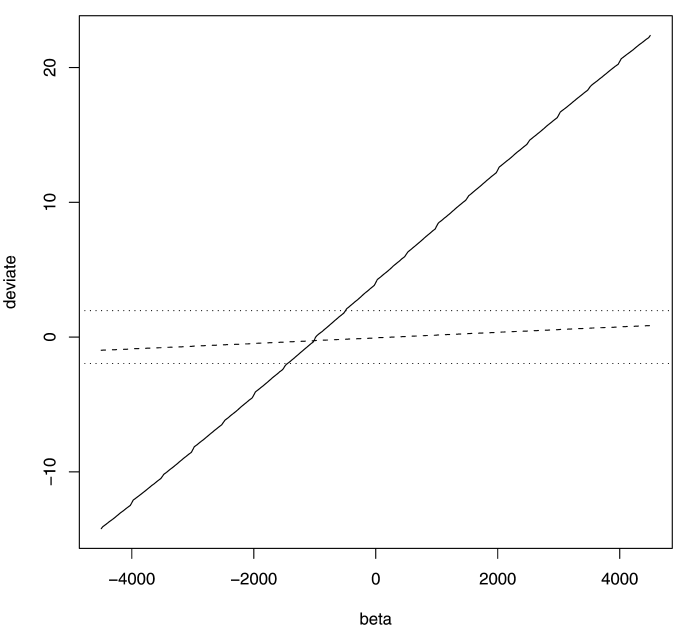

Figure 3. Strong and weak instruments, assuming no bias (——— strong IV: 1926 vs. 1928; – – – weak IV: 1924 vs. 1926; · · · · · 1.96 or -1.96).

these hypotheses are rejected. The deviate curve crosses ±1.96 far outside the limits of Figure 3, at −$10,130 and $10,750, so the 95% confidence interval is [−10,130, 10,750]. This interval is quite long; the effect of roughly ±$10,000 is large when compared with the median earnings for veterans of roughly $20,000 in Table 1.

It should be emphasized that in these pairs, there was only a slight shift, if any, in earnings from the men born in 1924 to the men born in 1926. The conventional 95% confidence interval for a location shift, 1926 minus 1924, in 1980 earnings, using Wilcoxon's signed-rank statistic, is −$250 to $240. Stated informally, the second deviate curve in Figure 3 is nearly flat because it is trying to apportion a tiny, perhaps zero, shift in earnings to a tiny shift in the fraction of WWII veterans.

In this example, we know when, where, and why the instrument is strong or weak; the strong instrument is created by the ending of WWII. In some other situation, we might have pairs that mixed strong and weak instruments, with no way to distinguish them. What would happen if the instrument were sometimes strong, sometimes weak? Just to illustrate what can happen, we combined the two sets of pairs in two different ways. In the first way, we formed 14,000 triples and built the confidence interval for β using the aligned-rank statistic of Hodges and Lehmann (1993), which generalizes the signed-rank test for matched pairs to matched sets. With 14,000 triples, this yields a 95% confidence interval for β of [−1,971, −227], which is longer, not shorter, than the interval [−1,445, −500] based on the 14,000 pairs from 1928 to 1926. Our first approach is strictly correct, because it takes into account the fact that the same man born in 1926 appears in a 1924–1926 pair and in a 1926–1928 pair, creating dependence between these two pairs. Our second approach incorrectly applies Wilcoxon's signed-rank test to the merged 28,000 pairs, ignoring the dependence. Dependence can invalidate the level of Wilcoxon's signed-rank test (see Gastwirth and Rubin 1971). If this is ignored, then the 95% confidence interval based on 28,000 pairs is [−1,524, −261], which is again longer, not shorter, than the interval [−1,445, −500] based on the 14,000 pairs from 1928 to 1926. The 14,000 weak pairs were able to produce a finite confidence interval of [−10,130, 10,750] on their own, so they contain some information; however, their addition to the strong pairs is a loss, not a gain, and this is true whether the (correct) aligned-rank test or the (incorrect) signed-rank test is used.

3.2 Sensitivity Analysis for a Nonrandom Instrument

Aside from veterans status and two years of age, men born in the last half of 1926 are not expected to differ dramatically from men born in the last half of 1928, but some differences are possible if not likely. There are many gradual, long-term trends in fertility, education, apprenticeship, and employment that affect earnings, and a small part of the long-term trend may bias comparisons of workers born 2 years apart.

There is little point in conducting a sensitivity analysis for the weak instrument 1924 versus 1926, because this would not provide useful information even assuming that it were valid, $\Gamma = 1$. For the strong instrument 1926 versus 1928, Figure 4 depicts a part of the sensitivity analysis, plotting the two deviates, $\{T^{\beta_0} - E(\overline{\overline{T}})\}/\sqrt{\text{var}(\overline{\overline{T}})}$ and $\{T^{\beta_0} - E(\overline{T})\}/\sqrt{\text{var}(\overline{T})}$, for

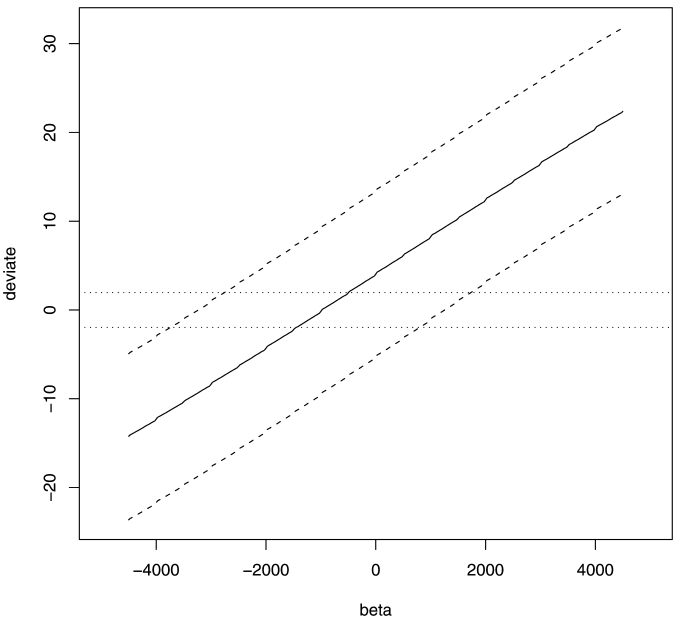

Figure 4. Sensitivity bounds, 1926 versus 1928 (—— $\Gamma = 1$; - - - $\Gamma = 1.2$).

the bounds in (7) when $\Gamma = 1.2$, with additional values of Γ considered in Table 2. The two bounding curves describe, in effect, the most extreme possible disturbance that could be produced by an unobserved binary covariate u_{ij} whose odds ratio with year of birth, 1926 to 1928, is at most $\Gamma = 1.2$. With no unobserved bias, $\Gamma = 1$, the 95% confidence interval for the effect of WWII service on 1980 earnings in dollars, β, was [−1,445, −500], but with $\Gamma = 1.2$, the lower endpoint of the interval could be as low as −$3,745 and the upper endpoint could be as high as $1,735; thus the $4,500 difference in Figure 1 remains implausible as an effect of WWII service, but substantial gains or losses are possible. Even with $\Gamma = 1.05$ just barely different from $\Gamma = 1$, the null hypothesis of no effect, $H_0 : \beta = 0$, cannot be rejected, because the minimum standardized deviate is $\{T^{\beta_0} - E(\overline{\overline{T}})\}/\sqrt{\text{var}(\overline{\overline{T}})} = 1.55$.

Table 2 considers other values of Γ and four specific hypotheses. The randomization distribution, $\Gamma = 1$, rejects at the .001 level all positive effects of service in WWII, but with a fairly small bias of $\Gamma = 1.2$, neither $H_0 : \beta = 0$ nor $H_0 : \beta = 1{,}000$ is rejected at the one-sided .05 level. The dramatic difference in Figure 1 is rejected as too large to be an effect for $\Gamma = 1.5$, but this is no longer true for $\Gamma = 1.6$. An enormous benefit, $\beta = 10{,}000$, is rejected for $\Gamma \leq 2.2$.

Table 2. Upper bound on one-sided significance level in the 14,000 1928 versus 1926 pairs

Γ	$H_0 : \beta = 0$	$H_0 : \beta = 1{,}000$	$H_0 : \beta = 4{,}500$	$H_0 : \beta = 10{,}000$
1	.001	.001	.001	.001
1.2	1.000	.860	.001	.001
1.5	1.000	1.000	.027	.001
1.6	1.000	1.000	.904	.001
2.2	1.000	1.000	1.000	.016
2.3	1.000	1.000	1.000	.476

In short, a naive analysis based on the comparison in Figure 1 might lead to the conclusion that military service in WWII increased earnings by perhaps $4,500. A less naive analysis, based on the comparison in Figure 2 and assuming a perfect instrument, $\Gamma = 1$, might lead to the conclusion that military service in WWII *reduced* earnings, perhaps by −$500 to −$1,445. A very small departure, $\Gamma = 1.05$, from a perfect instrument in Figure 2 would make it plausible that WWII service had no effect on earnings. For the naive analysis in Figure 1 to be correct as an estimate of the effect of WWII service on earnings, Figure 2 would need to be distorted by a moderate bias greater than $\Gamma = 1.5$.

An alternative approach to sensitivity analysis with an instrumental variable was developed by Small (2007). When that method is applied to the data in Figure 2, qualitatively similar conclusions are obtained.

4. POWER OF A SENSITIVITY ANALYSIS WITH AN INSTRUMENTAL VARIABLE

If Z_{ij} were in fact a valid instrument, so in fact $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = 1/2^I$ for each $\mathbf{z} \in \Omega$, then we could not be sure of this from empirical data, but we would hope to be able to report that conclusions were not extremely sensitive to small biases. More precisely, if $\beta - \beta_0$ were large, so that the hypothesis $H_0 : \beta = \beta_0$ was substantially in error, and if the potential magnitude of bias Γ in the assignment of encouragement were not very large, then we would hope to be able to reject the null hypothesis at level α, which we could do if the upper bound on the significance level in (7) were less than α, and this would happen if $T^{\beta_0} \geq \widetilde{t}_\alpha$. Whether this hope is likely to be realized depends on features of the study design, particularly features of the design that determined $\mathcal{F}$. Using simple models for $\mathcal{F}$, we ask: How do sample size and strength of the instrument affect the sensitivity to unobserved biases? If we knew that the instrument was valid, if encouragement had actually been assigned at random, then we would not need to study sensitivity to nonrandom assignment of encouragement, and a somewhat weak instrument could be offset by an increase in sample size. If we were not certain that encouragement had been assigned at random, and so planned to conduct a sensitivity analysis, then this strategy of offsetting weakness in the instrument by a larger sample size might or might not continue to work. It is this issue that the current section seeks to clarify. Specifically, we extend the concepts of design sensitivity and the power of a sensitivity analysis (Rosenbaum 2004, 2005) to cover observational studies with an instrumental variable.

The discussion in Section 2 and the analysis in Section 3 were conditional given $\mathcal{F}$, so tests had correct levels and confidence intervals had correct coverage rates no matter how $\mathcal{F}$ was generated; specifically, (7) and resulting tests and confidence intervals are correct no matter how $\mathcal{F}$ was generated. In this section we consider different models that might generate $\mathcal{F}$ and how they affect sensitivity. For considerations of power, a model that generates $\mathcal{F}$ is needed, and, as is usually done with power computations, we consider very simple models that are too simple to rely on in inference about β as alternative hypotheses.

If the instrument is valid, then $D_i^{\beta_0} = (\beta - \beta_0)S_i + \epsilon_i$, where $\Pr(Z_{ij} = 1|\mathcal{F}, \mathcal{Z}) = \frac{1}{2}$, ϵ_i in (4) is symmetric about 0 and S_i in (4) reflects the strength of the instrument. Using standard results of Lehmann (1998, sec. 4.2) about the expected power of Wilcoxon's signed-rank statistic, $E_{\mathcal{F}}\{\Pr(T^{\beta_0} \geq \widetilde{t}_\alpha|\mathcal{F}, \mathcal{Z})\}$ is approximately determined by four numbers, namely I and, in Lehmann's notation, $p = \Pr(D_i^{\beta_0} > 0)$, $p_1' = \Pr(D_i^{\beta_0} + D_j^{\beta_0} > 0)$, $i \neq j$, and $p_2' = \Pr(D_i^{\beta_0} + D_j^{\beta_0} > 0 \text{ and } D_i^{\beta_0} + D_k^{\beta_0} > 0)$, with $i < j < k$. A valid instrument and a fully specified model for $\mathcal{F}$ imply values for p, p_1', and p_2' that may be determined by direct calculation, numerical calculation, or simulation. We used simulation, computing half a million independent triples, $(D_i^{\beta_0}, D_j^{\beta_0}, D_k^{\beta_0})$, with each triple providing three correlated 1-or-0 indicators to estimate each of p, p_1', and p_2'.

For ϵ_i, we consider normal, Cauchy, and logistic distributions symmetric about 0. In power calculations, the simple model that we consider for compliance (w_{Tij}, w_{Cij}) has no defiers, and the other three compliance types are purely random, independent of ϵ_i and covariates, with a multinomial distribution with probabilities π_A for always takers, π_C for compliers and π_N for never takers, $\pi_A + \pi_C + \pi_N = 1$. To repeat, this purely random compliance model is never assumed in the inference in Sections 2 and 3, where $\mathcal{F}$ is fixed; rather, the calculations evaluate the power of methods that do not assume random compliance when applied to $\mathcal{F}$'s in which compliance happens to be random. Under this model, an always taker with $(w_{Tij}, w_{Cij}) = (1, 1)$ is paired to another always taker with probability π_A^2, yielding $S_i = 0$. Continuing in this way, $\Pr(S_i = 1) = \pi_C^2 + \pi_A\pi_C + \pi_N\pi_C + \pi_A\pi_N$, $\Pr(S_i = 0) = \pi_A^2 + \pi_N^2 + \pi_A\pi_C + \pi_N\pi_C$, $\Pr(S_i = -1) = \pi_A\pi_N$, and $1 = \Pr(S_i = 1) + \Pr(S_i = 0) + \Pr(S_i = -1)$.

Table 3 gives power when $\epsilon_i \sim_{\text{iid}} \text{N}(0, \sigma^2)$ and the null hypothesis is far from correct, $(\beta - \beta_0)/\sigma = 1$. Table 4 is similar, except the errors have a Cauchy distribution. In Table 5 the errors are again $\epsilon_i \sim_{\text{iid}} \text{N}(0, \sigma^2)$, but now $(\beta - \beta_0)/\sigma = \frac{1}{2}$. In these tables the case of $\Gamma = 1$ and 100% compliance, $(\pi_A, \pi_C, \pi_N) = (0, 1, 0)$, is the power of a randomized experiment, with perfect compliance and an additive treatment effect, so it is exactly the power calculation of Lehmann (1998, sec. 4.2). When there is noncompliance, when $\pi_C < 1$, there is no longer an effect β that appears in all pairs; rather, β appears with probability $\Pr(S_i = 1)$, 0 appears with probability $\Pr(S_i = 0)$, and $-\beta$ appears with probability $\Pr(S_i = -1) = \pi_A\pi_N$. The case of $\Gamma > 1$ asks: Would the results be significant in a one-sided .025 level test if the signed rank statistic were compared with the upper bound in (7); that is, what is the probability that $T^{\beta_0} \geq \widetilde{t}_\alpha$? The case where $\Gamma = 1.2$ is a nontrivial but fairly small bias, whereas $\Gamma = 2$ is a moderate bias.

There are $2I$ subjects in I pairs, so a large study with low compliance, $I = 100{,}000$ and $\pi_C = \frac{1}{10}$, is expected to have $2I\pi_C = 20{,}000$ compliers and 180,000 noncompliers, whereas a small study with moderate compliance, $I = 100$ and $\pi_C = \frac{1}{2}$, is expected to have $2I\pi_C = 100$ compliers and 100 noncompliers.

When the instrument is valid, in the sense that $\Gamma = 1$, the power is, of course, greater with greater compliance, but there is good power in all cases in Tables 3–5 with a sample size of $I = 10{,}000$, and the tests are consistent, with limiting power of 1 as $I \to \infty$. Under the simple models and methods considered here, valid but weak instruments eventually get it right. When $\Gamma > 1$, the power can tend to either 1 or 0, depending on the

Table 3. Power of a two-sided, .05-level sensitivity analysis with normal errors, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, and noncentrality parameter $(\beta - \beta_0)/\sigma = 1$

Compliance	(π_A, π_C, π_N)	Γ	100	1,000	10,000	100,000	$\lim_{I\to\infty}$
100%	(0, 1, 0)	1	1.00	1.00	1.00	1.00	1
50%	(.25, .5, .25)	1	.99	1.00	1.00	1.00	1
20%	(.4, .2, .4)	1	.37	1.00	1.00	1.00	1
10%	(.45, .1, .45)	1	.12	.73	1.00	1.00	1
100%	(0, 1, 0)	1.2	1.00	1.00	1.00	1.00	1
50%	(.25, .5, .25)	1.2	.92	1.00	1.00	1.00	1
20%	(.4, .2, .4)	1.2	.13	.77	1.00	1.00	1
10%	(.45, .1, .45)	1.2	.03	.03	.04	.10	1
100%	(0, 1, 0)	2	1.00	1.00	1.00	1.00	1
50%	(.25, .5, .25)	2	.18	.97	1.00	1.00	1
20%	(.4, .2, .4)	2	0	0	0	0	0
10%	(.45, .1, .45)	2	0	0	0	0	0

strength of the instrument, (π_A, π_C, π_N), the size of the noncentrality parameter, $(\beta - \beta_0)/\sigma$, and the distribution of errors ϵ_i; that is, as $I \to \infty$, the probability that $T^{\beta_0} \geq \widetilde{t}_\alpha$ tends to 0 or 1 (see Sec. 5 for detailed discussion). For instance, in Table 3 with $\pi_C = \frac{1}{10}$, the probability of rejecting $H_0 : \beta = \beta_0$, because $T^{\beta_0} \geq \widetilde{t}_\alpha$ tends very slowly to 1 when $\Gamma = 1.2$ but tends to 0 when $\Gamma = 2$. For a moderate bias, $\Gamma = 2$, in Table 3, there is excellent power of 97% in a moderate-sized study with moderate compliance, $I = 1{,}000$ and $\pi_C = \frac{1}{2}$, and no power at all in a very large study with fairly poor compliance, $I = 100{,}000$ and $\pi_C = \frac{1}{5}$, even though the smaller study contains $2I\pi_C = 1{,}000$ compliers and the larger study contains $2I\pi_C = 40{,}000$ compliers.

5. DESIGN SENSITIVITY

As has been seen, the power of the sensitivity analysis tends either to 0 or 1 as $I \to \infty$, depending on the value of Γ. The shift from limiting power of 1 to limiting power of 0 occurs at a value of Γ called the design sensitivity (Rosenbaum 2004, 2005), which may be shown to be the limiting solution, as $I \to \infty$, of the equation $E_\Gamma(\overline{\overline{T}}) = E_1(T^{\beta_0})$, where $E_\Gamma(\overline{\overline{T}})$ is the expectation of $\overline{\overline{T}}$ in (7) assuming that the null hypothesis is true and a bias of magnitude Γ, and $E_1(T^{\beta_0})$ is the expectation of T^{β_0} under some alternative hypothesis without unobserved bias, $\Gamma = 1$. For the signed-rank statistic,

$$E_\Gamma(\overline{\overline{T}}) = \frac{\Gamma}{1+\Gamma}\frac{I(I+1)}{2} \qquad \text{and}$$

$$E_1(T^{\beta_0}) = \frac{I(I-1)p_1'}{2} + Ip.$$

As $I \to \infty$,

$$\frac{2}{I^2}E_\Gamma(\overline{\overline{T}}) \to \frac{\Gamma}{1+\Gamma} \qquad \text{and} \qquad \frac{2}{I^2}E_1(T^{\beta_0}) \to p_1',$$

so the design sensitivity solves $p_1' = \Gamma/(1+\Gamma)$; therefore, $\Gamma = p_1'/(1-p_1')$.

Table 6 shows the design sensitivity for various situations. Again, in a sensitivity analysis, the power tends to 0 as $I \to \infty$ for Γ greater than the design sensitivity, and the power tends to 1 for Γ less than the design sensitivity. For instance, in Table 3, with $(\pi_A, \pi_C, \pi_N) = (.45, .1, .45)$ and $\Gamma = 1.2$, the power increases very slowly to 1 as $I \to \infty$, because $p_1'/(1-p_1')$ is just a hair above 1.2.

In Table 6, in the most favorable circumstances—normal errors, 100% compliance, and a large gap between null and alternative hypotheses, $(\beta_0 - \beta)/\sigma = 1$—the design sensitivity is

Table 4. Power of a two-sided, .05-level sensitivity analysis with Cauchy errors and noncentrality parameter $(\beta - \beta_0)/\sigma = 1$

Compliance	(π_A, π_C, π_N)	Γ	100	1,000	10,000	100,000	$\lim_{I\to\infty}$
100%	(0, 1, 0)	1	1.00	1.00	1.00	1.00	1
50%	(.25, .5, .25)	1	.64	1.00	1.00	1.00	1
20%	(.4, .2, .4)	1	.15	.84	1.00	1.00	1
10%	(.45, .1, .45)	1	.07	.31	1.00	1.00	1
100%	(0, 1, 0)	1.2	.96	1.00	1.00	1.00	1
50%	(.25, .5, .25)	1.2	.33	1.00	1.00	1.00	1
20%	(.4, .2, .4)	1.2	.04	.07	.33	1.00	1
10%	(.45, .1, .45)	1.2	.01	0	0	0	0
100%	(0, 1, 0)	2	.33	1.00	1.00	1.00	1
50%	(.25, .5, .25)	2	.01	0	0	0.00	0
20%	(.4, .2, .4)	2	0	0	0	0	0
10%	(.45, .1, .45)	2	0	0	0	0	0

Table 5. Power of a two-sided, .05-level sensitivity analysis with normal errors, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, and noncentrality parameter $(\beta - \beta_0)/\sigma = 1/2$

Compliance	(π_A, π_C, π_N)	Γ	100	1,000	10,000	100,000	$\lim_{I\to\infty}$
100%	(0, 1, 0)	1	1.00	1.00	1.00	1.00	1
50%	(.25, .5, .25)	1	.64	1.00	1.00	1.00	1
20%	(.4, .2, .4)	1	.15	.84	1.00	1.00	1
10%	(.45, .1, .45)	1	.07	.32	1.00	1.00	1
100%	(0, 1, 0)	1.2	.98	1.00	1.00	1.00	1
50%	(.25, .5, .25)	1.2	.32	1.00	1.00	1.00	1
20%	(.4, .2, .4)	1.2	.03	.07	.30	1.00	1
10%	(.45, .1, .45)	1.2	.01	0	0	0	0
100%	(0, 1, 0)	2	.38	1.00	1.00	1.00	1
50%	(.25, .5, .25)	2	.01	0	0	0	0
20%	(.4, .2, .4)	2	0	0	0	0	0
10%	(.45, .1, .45)	2	0	0	0	0	0

$\Gamma = 11.7$, which is far less sensitive to unobserved biases than Hammond's (1964) study of heavy smoking as a cause of lung cancer, which became sensitive at $\Gamma = 6$. In the second column of Table 6 with a strong instrument and 50% compliance, results are sensitive to moderate biases but not to extremely small biases. In the fourth column of Table 6 with a weak instrument, results are consistently sensitive to quite small biases. Table 6 shows that even with a very large departure from the null hypothesis, $(\beta_0 - \beta)/\sigma = 1$, no matter how large the sample size becomes, a study with a weak instrument, say 10% compliance, inevitably will be sensitive to quite small biases.

6. DISCUSSION: PRACTICAL ADVICE

Studies that use weak instruments face three problems: a common but easily fixed problem; a more serious problem that can be addressed by a sufficient increase in the sample size, I; and a third problem that cannot be fixed no matter how large I becomes. Beginning with Bound et al. (1995), it has been recognized that the most common method of inference with instrumental variables, two-stage least squares, gives highly misleading inferences when the instrument is weak even when the instrument is perfectly valid (i.e., $\Gamma = 1$). This first problem is easily fixed by using other methods of inference, such as those of Imbens and Rosenbaum (2005). The second problem is also well known; as can be seen in Tables 3–5, with appropriate methods and a perfectly valid instrument (i.e., $\Gamma = 1$), the power is lower with a weaker instrument, but it rises to 1 as $I \to \infty$. In point of fact, very large sample sizes are sometimes available from the Census or Social Security or Medicare, so this second issue restricts the use of weak instruments but does not eliminate their usefulness.

Table 6. Design sensitivity $\Gamma = p_1'/(1 - p_1')$ for instruments with varying strength, three error distributions and two noncentrality parameters

	Compliance (π_A, π_C, π_N)	100% (0, 1, 0)	50% $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$	20% $(\frac{2}{5}, \frac{1}{5}, \frac{2}{5})$	10% $(\frac{9}{20}, \frac{2}{20}, \frac{9}{20})$
ϵ_i	$(\beta_0 - \beta)/\sigma$				
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
Logistic	1	3.9	1.9	1.3	1.1
Logistic	$\frac{1}{2}$	2.0	1.4	1.1	1.1

The third problem persists no matter how large the sample size becomes; weak instruments are sensitive to quite small biases ($\Gamma > 1$ yet $\Gamma \doteq 1$, say $\Gamma = 1.1$), even when the effect size $(\beta_0 - \beta)/\sigma$ is quite large. Unless one is confident that a weak instrument is perfectly valid (i.e., $\Gamma = 1$), its extreme sensitivity to small biases in Table 6 is likely to limit its usefulness to the study of enormous effects, $(\beta_0 - \beta)/\sigma \gg 1$. In contrast, a strong instrument may provide useful information even if moderate biases are plausible. Several practical consequences follow:

1. A small study with a stronger instrument is likely to be much less sensitive to bias than a vastly larger study with a weak instrument. For instance, for $\Gamma = 1.2$ in Table 3, the sensitivity analysis has power .92 for $I = 100$ pairs with 50% compliance and power .10 for $I = 100{,}000$ with 10% compliance.
2. A slightly biased but strong instrument may be preferable to a less biased but weak instrument. For instance, with $I = 1{,}000$ pairs in Table 3, the sensitivity analysis for a moderately biased but strong instrument ($\Gamma = 2$, 50% compliance) has power .97, but for a much less biased but weak instrument ($\Gamma = 1.2$, 10% compliance), the power is .03, and even for a perfectly valid weak instrument ($\Gamma = 1$, 10% compliance), the power is only .73.
3. In Table 6, for strong instruments, the sensitivity to unobserved biases is meaningfully affected by the magnitude of the effect size, $(\beta_0 - \beta)/\sigma$, whereas for a weak instrument, there is barely any difference between $(\beta_0 - \beta)/\sigma = 1$ and $(\beta_0 - \beta)/\sigma = \frac{1}{2}$. Sensitivity to unobserved bias sometimes can be reduced by increasing the effect size, $(\beta_0 - \beta)/\sigma$, say by reducing the unexplained heterogeneity σ of experimental subjects (Rosenbaum 2005). For instance, Ashenfelter and Rouse (1998) studied the effects of additional education on earnings using identical twins, and Kim (2007) compared the earnings of veteran siblings to estimate the effect of being drafted. Strategies of this sort may be helpful with strong instruments and largely ineffective with weak instruments.

In the WWII veterans data, birth date is a strong instrument near the end of the war and a weak instrument in the middle of the war. The effect of WWII veteran status on earnings is less sensitive to unobserved biases when the instrument is strong near the end of the war compared to when the instrument is weak in the middle of the war. The strong instrument was able to reject a substantial gain in earnings of $4,500 suggested by Figure 1 even in the presence of an unobserved bias of magnitude $\Gamma = 1.5$, whereas the weak instrument could not reject a gain of $10,000 even without bias, $\Gamma = 1$. Moreover, the larger sample, formed by combining the strong and weak periods, increases sensitivity to unobserved biases; it is better to have the smaller sample with the consistently strong instrument.

[Received August 2006. Revised April 2007.]

REFERENCES

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records," *American Economic Review*, 80, 313–336.

Angrist, J., and Krueger, A. B. (1994), "Why Do World War II Veterans Earn More Than Nonveterans?" *Journal of Labor Economics*, 12, 74–97.

Angrist, J., Imbens, G., and Rubin, D. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–469.

Ashenfelter, O., and Rouse, C. (1998), "Income, Schooling and Ability: Evidence From a New Sample of Identical Twins," *Quarterly Journal of Economics*, 113, 253–284.

Bedard, K., and Deschênes, O. (2006), "The Long-Term Impact of Military Service on Health: Evidence From World War II and Korean War Veterans," *American Economic Review*, 96, 176–194.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995), "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Copas, J., and Eguchi, S. (2001), "Local Sensitivity Approximations for Selectivity Bias," *Journal of the Royal Statistical Society*, Ser. B, 63, 871–896.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959), "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute*, 22, 173–203.

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

Frangakis, C. E., and Rubin, D. B. (1999), "Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes," *Biometrika*, 86, 365–379.

Gastwirth, J. L. (1992), "Methods for Assessing the Sensitivity of Statistical Comparisons Used in Title VII Cases to Omitted Variables," *Jurimetrics*, 33, 19–34.

Gastwirth, J. L., and Rubin, H. (1971), "Effect of Dependence on the Level of Some One-Sample Tests," *Journal of the American Statistical Association*, 66, 816–820.

Greevy, R., Silber, J. H., Cnaan, A., and Rosenbaum, P. R. (2004), "Randomization Inference With Imperfect Compliance in the ACE–Inhibitor After Anthracycline Randomized Trial," *Journal of the American Statistical Association*, 99, 7–15.

Hammond, E. C. (1964), "Smoking in Relation to Mortality and Morbidity," *Journal of the National Cancer Institute*, 32, 1161–1188.

Hodges, J. L., and Lehmann, E. L. (1962), "Rank Methods for Combination of Independent Experiments in the Analysis of Variance," *The Annals of Mathematical Statistics*, 33, 482–497.

Holland, P. W. (1988), "Causal Inference, Path Analysis, and Recursive Structural Equations Models," *Sociological Methodology*, 18, 449–484.

Imbens, G. W. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132.

Imbens, G., and Rosenbaum, P. R. (2005), "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society*, Ser. A, 168, 109–126.

Imbens, G. W., and van der Klaauw, W. (1995), "Evaluating the Cost of Conscription in the Netherlands," *Journal of Business & Economic Statistics*, 13, 207–215.

Kim, H. (2007), "The Military Draft and Career Disruption," manuscript, Dept. of Economics, University of Wisconsin at Madison, available at *http://www.ssc.wisc.edu/~hikim/*.

Lehmann, E. L. (1998), *Nonparametrics*, Englewood Cliffs, NJ: Prentice-Hall.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics*, 54, 948–963.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," reprinted in English with discussion by T. Speed and D. B. Rubin in *Statistical Science*, 1990, 5, 463–480.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference," in *Statistical Models in Epidemiology*, eds. E. Halloran and D. Berry, New York: Springer, pp. 1–94.

Rosenbaum, P. R. (1984), "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment," *Journal of the Royal Statistical Society*, Ser. A, 147, 656–666.

——— (1993), "Hodges–Lehmann Point Estimates of Treatment Effect in Observational Studies," *Journal of the American Statistical Association*, 88, 1250–1253.

——— (1996), Comment on "Identification of Causal Effects Using Instrumental Variables," by J. Angrist, G. Imbens, and D. Rubin, *Journal of the American Statistical Association*, 91, 465–468.

——— (1999), "Using Combined Quantile Averages in Matched Observational Studies," *Applied Statistics*, 48, 63–78.

——— (2002a), "Covariance Adjustment in Randomized Experiments and Observational Studies," *Statistical Science*, 17, 286–327.

——— (2002b), *Observational Studies* (2nd ed.), New York: Springer.

——— (2003), "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed-Rank Test," *American Statistician*, 57, 132–138.

——— (2004), "Design Sensitivity in Observational Studies," *Biometrika*, 91, 153–164.

——— (2005), "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies," *American Statistician*, 59, 147–152.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Small, D. S. (2007), "Sensitivity Analysis for Instrumental Variables Regression With Overidentifying Restrictions," *Journal of the American Statistical Association*, 102, 1049–1058.

Sommer, A., and Zeger, S. L. (1991), "On Estimating Efficacy From Clinical Trials," *Statistics in Medicine*, 10, 45–52.

Tan, Z. (2006), "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101, 1607–1618.

Taubman, P. (1976), "Earnings, Education, Genetics and Environment," *Journal of Human Resources*, 11, 447–461.

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error," *The Annals of Mathematical Statistics*, 11, 284–300.

Wang, L. S., and Krieger, A. (2006), "Causal Conclusions Are Most Sensitive to Unobserved Binary Covariates," *Statistics in Medicine*, 25, 2256–2271.