Reinforced designs: Multiple instruments plus control groups as evidence factors in an observational study of the effectiveness of Catholic schools

Bikram Karmakar, Dylan S. Small and Paul R. Rosenbaum[1]

Abstract. Absent randomization, inference about the effects caused by treatments depends upon assumptions that can be difficult or impossible to verify. Causal conclusions gain strength from a demonstration that they are insensitive to small or moderate violations of those assumptions, especially if that happens in each of several statistically independent analyses that depend upon very different assumptions; that is, if several evidence factors concur. These issues arise with several instruments, together with the option of a direct comparison of treated and control subjects. Does each purported instrument actually satisfy the stringent assumptions required of an instrument? Is a direct comparison without instruments biased by self-selection into treated and control groups? We develop a method for constructing evidence factors, and evaluate its performance in terms of design sensitivity and simulation. In the application, we consider the effectiveness of Catholic versus public high schools, constructing three evidence factors from three past strategies for studying this question, namely: (i) having nearby access to a Catholic school as an instrument, (ii) being Catholic as an instrument for attending Catholic school, and (iii) a direct comparison of students in Catholic and public high schools. Although these three analyses use the same data, we: (i) construct three essentially independent statistical tests of no effect that require very different assumptions, (ii) study the sensitivity of each test to the assumptions underlying that test, (iii) examine the degree to which independent tests dependent upon different assumptions concur, (iv) pool evidence across independent factors. In the application, we conclude that the ostensible benefit of Catholic education depends critically on the validity of one instrument, and is therefore quite fragile.

Keywords: Evidence factors; instrumental variables; sensitivity analysis.

[1]Bikram Karmakar is assistant professor, Department of Statistics, University of Florida and Dylan S. Small and Paul R. Rosenbaum are professors, Statistics Department, University of Pennsylvania. 2/11/20.

1 Introduction: Addressing unmeasured bias in observational studies

Absent random assignment, an association between treatment received and outcome exhibited may not reflect an effect caused by the treatment, but rather some bias in the way people were selected for treatment. How can unmeasured bias be addressed?

A sensitivity analysis asks about the magnitude of bias in treatment assignment that would need to be present to materially alter the causal conclusions. See Cornfield et al. (1959) for the first sensitivity analysis in an observational study, and for related methods, see, for instance, Hosman et al. (2010), Gilbert et al. (2003), Rosenbaum (2002, §4), Rudolph and Stuart (2017), Schwartz et al. (2012) and Yu and Gastwirth (2005).

An instrument is a bit of randomized or haphazard encouragement to accept a treatment in a context in which treatment assignment is often deliberate, hence potentially biased; see Angrist, Imbens and Rubin (1996). In their example, the Vietnam War draft lottery was randomized, and it pushed some men into military service who would not otherwise have served. Typical instruments are less compelling, because they are not actually randomized, so they make assumptions no less speculative than the assumption that unmeasured biases are absent. As a single treatment might be encouraged by various instruments, it is not a foregone conclusion that analyses with different instruments will concur; so, it is informative if they do concur (Imbens and Rosenbaum 2005, §1.1). For discussion of instruments, see Angrist et al. (1996), Baiocchi, Cheng and Small (2014), Brookhart et al. (2010), Hogan and Lancaster (2004), Kang (2016), Keele and Morgan (2016), Larcker and Rusticus (2010), Li et al. (2015), Lu and Marcus (2012) and Small (2007).

Multiple analyses provide evidence about unmeasured biases if: (i) certain biases that would invalidate one analysis do not bias another analysis, (ii) each analysis is insensitive to small or moderate biases of the type that might invalidate that analysis, and (iii) these several analyses would be nearly statistically independent if the treatment had no effect.

Analyses of this type are called evidence factors (Rosenbaum 2010, 2011, 2017a). Because these analyses are affected by different types of unmeasured biases and are nearly independent despite using the same data, it is far from a forgone conclusion that the analyses will concur. See Zubizarreta et al. (2012) for an example in which two evidence factors do not concur, thereby providing evidence that at least some associations are spurious, not causal. See Zhang et al. (2011) for an example in which two evidence factors concur.

Our goal is to design observational studies to use several instruments plus direct comparisons of treated and control groups as evidence factors. Typically, when several instruments are available, investigators employ them in a joint analysis, such as two-stage least squares, so if any of the instruments fails to satisfy the many assumptions required of an instrument, then the joint analysis is compromised. However, see Kang et al. (2016) for an approach that tolerates some invalid instruments. In contrast, we use several instruments one at a time in such a way that failure of the assumptions for one instrument does not, by itself, invalidate analyses with other instruments.

A final evidence factor directly compares treated and control groups. Some scientific fields presume that direct comparisons are more easily dismissed as bias than are comparisons using instruments, but this presumption is not true in general. In a single sampling situation, a direct comparison may be insensitive to large biases, while the analysis with an instrument may be sensitive to small biases, so the direct comparison may provide the most compelling evidence. We demonstrate this theoretically by calculating design sensitivities.

Section 2 introduces the application, to which we return in §6. Section 3 defines notation. Key results in §4 demonstrate that the several subanalyses are, indeed, evidence factors. The method is evaluated in §5 in terms of design sensitivity, with some surprises about the relative safety of instruments and direct comparisons.

2 How effective are Catholic schools compared to public school?

A central question today about education in the US is whether traditional public schools are less effective than other options that offer more choice to parents and more competition among schools for students; see, for instance, Card, Dooley and Payne (2010), Garcia (2018), and Hoxby (2000). The earliest studies of this issue compared public high schools to private Catholic high schools, but the conclusions and implications of these studies remain controversial, in part for methodological reasons. Private education is typically expensive, beyond the means of most middle-class families in the US, but private Catholic education is subsidized by the Catholic Church and so is more affordable. Students with similar economic backgrounds may attend either public or Catholic high schools.

Are private Catholic high schools more effective than public high schools? Though important to the contemporary debate about school choice, this is not an easy question to answer. Paying more to attend Catholic school may signify a parent's concern or commitment to education, which may affect outcomes in many ways. Even after adjustment for educational and socioeconomic covariates, a direct comparison of students in public and Catholic school may, therefore, be biased. The empirical literature contains: (i) attempts to use the geographic accessibility of Catholic schools as an instrument for attending Catholic schools, (ii) attempts to use "being a Catholic" as an instrument, (iii) direct comparisons of students in Catholic and public high schools, (iv) sharp conflict about which, if any, of these approaches yields valid inferences about the effects caused by Catholic schools. See Altonji, et al. (2005), Coleman (1982), Goldberger and Cain (1982), Kim (2011), and Neal (1997) for several perspectives and analyses.

Rather than select one analysis and assert that it is valid, we develop three evidence factors, three (nearly) statistically independent analyses of the same data, each dependent upon very different assumptions for its validity. Because these analyses are independent,

they do not repeat one another, and concurrence among the analyses is far from a foregone conclusion. Because certain biases that would invalidate one analysis have no effect on another analysis, concurrence would weaken some claims that biases produced the ostensible treatment effects. To the extent that each analysis, each factor, is insensitive to the type of bias that could invalidate that factor, there is further weakening of claims that bias accounts for ostensible effects. Conversely, the absence of concurrence and sensitivity to small biases are warnings that bias could readily explain ostensible effects.

Using data from the Wisconsin Longitudinal Study, we will examine income from wages and salary in 1974 for 4450 male students who completed high school in Wisconsin in 1957. Table 1 depicts the structure of the three factors, essentially (i)-(iii) above. The 4450 students divide into 1501 students from urban Wisconsin and 2949 from rural Wisconsin, and presumably because Catholic schools are more accessible in urban areas, 22% of urban students attended Catholic school, while only 6% of rural students attended Catholic school. So the first analysis uses urban/rural as an instrument for Catholic education. The second analysis compares children in urban areas to other children in urban areas, and children in rural areas to other children in rural areas, so the second analysis views urban/rural as a covariate, not an instrument. In urban areas, roughly half the students were Catholic, and 44% of Catholics attended Catholic high schools, while none of the non-Catholics attended Catholic high schools. In rural areas, Catholic students were in the minority, and 17% of Catholics attended Catholic high schools, while only 2 of the non-Catholics, or less than half a percent, attended Catholic high schools. The second analysis views Catholic religion as an instrument, and urban/rural as a covariate. The third analysis has no instrument: it directly compares students attending Catholic and public schools, viewing both urban/rural and Catholic/non-Catholic as covariates.

3 Notation and background: several instruments plus a direct comparison

3.1 Notation: strata, covariates, outcomes, treatment and instruments

There are I strata, $i = 1, \ldots, I$, with n_i individuals ij in stratum i, $j = 1, \ldots, n_i$, and $N = \sum_{i=1}^{I} n_i$ individuals in total. There are K binary, 1 or 0, indicators, Z_{ijk}, $k = 1, \ldots, K$. The first $K - 1$ indicators are possible instruments for indicator K which is the active treatment. In §2: (i) $Z_{ij1} = 1$ for urban residence, $Z_{ij1} = 0$ for rural residence; (ii) Z_{ij2} distinguishes Catholics, $Z_{ij2} = 1$, from others, $Z_{ij2} = 0$; (iii) Z_{ij3} distinguishes attending a Catholic high school, $Z_{ij3} = 1$, from attending a public school, $Z_{ij3} = 0$.

Individual ij has an observed covariate $\mathbf{x}_{ij}$ controlled by stratification, so $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ for $1 \leq j < j' \leq n_i$. There is concern about an unobserved covariate u_{ijk}, $k = 1, \ldots, K$, not controlled by stratifying on $\mathbf{x}_{ij}$. The notation permits a different unobserved u_{ijk} for each Z_{ijk}, but there is no requirement that they be distinct; that is, the situation with $u_{ij1} = \cdots = u_{ijK} = u_{ij}$, say, is simply a special case.

If the exclusion restriction of Angrist et al. (1996) held for all of the first $K - 1$ indicators, then individual ij would exhibit response r_{Tij} if $Z_{ijK} = 1$ or response r_{Cij} if $Z_{ijK} = 0$. In fact, our K analyses do not assume that the exclusion restriction holds for all $K-1$ potential instruments, but rather assume much less. The analysis that uses Z_{ijk} as if it were an instrument assumes that the exclusion restriction holds for Z_{ijk},…,Z_{ijK-1} when we compare individuals who are the same in terms of Z_{ij1}, …, Z_{ijk-1}, so the exclusion restriction may not hold for Z_{ij1}, …, Z_{ijk-1}. The direct comparison of treated and control individuals, $Z_{ijK} = 1$ versus $Z_{ijK} = 0$, does not assume any exclusion restriction, simply adjusting for Z_{ij1}, …, Z_{ijK-1}.

There are K partial assignment vectors, $\mathbf{A}_{ijk} = (Z_{ij1}, \ldots, Z_{ijk})$ for $k = 1, \ldots, K$, and a matrix $\mathbf{A}_k$ whose N rows are the $\mathbf{A}_{ijk}$, so $\mathbf{A}_k$ records assignments up to step k for all

N individuals. It is notationally convenient to define $\mathbf{A}_{ij0} = \emptyset$, so that $\mathbf{A}_{ij,k-1}$ is well defined for $k = 1$, but conditioning on $\mathbf{A}_{ij0}$ means that no part of $\mathbf{A}_{ij} = (Z_{ij1}, \ldots, Z_{ijK})$ is actually being conditioned upon. Let $\mathcal{A}_k$ be the set containing the 2^k vectors of dimension k with 1 or 0 coordinates. The vector, $\mathbf{A}_{ijk} = (Z_{ij1}, \ldots, Z_{ijk})$, can take on 2^k possible values $\mathbf{a} \in \mathcal{A}_k$. The entire study amalgamates K partial studies, where study k fixes the 2^{k-1} values $\mathbf{a} \in \mathcal{A}_{k-1}$ of $\mathbf{A}_{ij,k-1}$, studies the effects of variations in Z_{ijk}, and lets $(Z_{ij,k+1}, \ldots, Z_{ijK})$ fluctuate as it will. In Table 1, there are $K = 3$ partial studies. At assignment k, individual ij has 2^k potential outcomes, $r_{ij\mathbf{a}}$ with $\mathbf{a} \in \mathcal{A}_k$, so each step is an instance of the Neyman (1923) - Rubin (1974) notation for causal effects. However, at step k, we are interested in comparing $r_{ij\mathbf{a}}$ and $r_{ij\mathbf{a}'}$ for each pair $(\mathbf{a}, \mathbf{a}')$ with $\mathbf{a}$, $\mathbf{a}' \in \mathcal{A}_k$ such that $\mathbf{a} = (a_1, \ldots, a_{k-1}, 1)$ and $\mathbf{a}' = (a_1, \ldots, a_{k-1}, 0)$, so $\mathbf{a}$ and $\mathbf{a}'$ differ only in the last, kth, coordinate; moreover, the kth partial study is focused on this comparison. At assignment k, Fisher's hypothesis H_k of no effect of assignment k asserts that $r_{ij\mathbf{a}} = r_{ij\mathbf{a}'}$ for each pair $(\mathbf{a}, \mathbf{a}')$ with $\mathbf{a}$, $\mathbf{a}' \in \mathcal{A}_k$ such that $\mathbf{a} = (a_1, \ldots, a_{k-1}, 1)$ and $\mathbf{a}' = (a_1, \ldots, a_{k-1}, 0)$. In total, individual ij has $\sum_{k=1}^{K} 2^k = 2^{K+1} - 2$ potential outcomes $r_{ij\mathbf{a}}$ for $\mathbf{A}_{ijk} = a \in \mathcal{A}_k$, for $k = 1, \ldots, K$, which we collect in a vector $\mathbf{r}_{ij}$ of dimension $2^{K+1} - 2$. Later, in §3.3, when we impose a "partial exclusion condition" on $\mathbf{r}_{ij}$, the potential complexity of $\mathbf{r}_{ij}$ will be greatly restricted. In effect, the partial exclusion restriction will say that for people who are the same in terms of $(Z_{ij1}, \ldots, Z_{ij,k-1})$, the vector $(Z_{ijk}, \ldots, Z_{ij,K-1})$ affects $\mathbf{r}_{ij}$ only indirectly by altering Z_{ijK}; see, again, Table 1. Ultimately, we observe only one coordinate of $\mathbf{r}_{ij}$, namely $R_{ij} = r_{ij\mathbf{a}}$ for $\mathbf{A}_{ijK} = \mathbf{a}$ for $\mathbf{a} \in \mathcal{A}_K$, so much of $\mathbf{r}_{ij}$ is inaccessible to us. Write $\mathbf{R} = (R_{11}, \ldots, R_{I,n_I})^T$ for the N-dimensional vector of observed responses. Concerning notation, note that a vector $\mathbf{a} \in \mathcal{A}_k$ has dimension k, so the notation $r_{ij\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}_k$, is well defined without mentioning k.

Write $\mathcal{F} = \{\mathbf{r}_{ij}, \mathbf{x}_{ij}, u_{ijk},\ i = 1, \ldots, I,\ j = 1, \ldots, n_i,\ k = 1, \ldots, K\}$. Conditionally given

$\mathcal{F}$, distinct individuals, say ij and $i'j'$, are assumed to have independent values of the K-dimensional assignment vector, $\mathbf{A}_{ij} = (Z_{ij1}, \ldots, Z_{ijK})$ and $\mathbf{A}_{i'j'}$. Write $\mathbf{Z}_k = (Z_{11k}, \ldots, Z_{I,n_I,k})^T$ for the N-dimensional vector of assignments at step k, for $k = 1, \ldots, K$.

3.2 Treatment assignment in K steps

Consider the model for treatment assignment

$$\Pr\left(Z_{ijk} = 1 \,|\, \mathcal{F},\, \mathbf{A}_{ij,k-1}\right) = \frac{\exp\left\{\kappa_k\left(\mathbf{x}_{ij}, \mathbf{A}_{ij,k-1}\right) + \gamma_k\, u_{ijk}\right\}}{1 + \exp\left\{\kappa_k\left(\mathbf{x}_{ij}, \mathbf{A}_{ij,k-1}\right) + \gamma_k\, u_{ijk}\right\}}, \text{ with } 0 \le u_{ijk} \le 1, \quad (1)$$

where $\kappa_k(\cdot)$ is an unknown function and $\gamma_k \ge 0$ is an unknown sensitivity parameter. This model says that Z_{ijk} depends in an entirely arbitrary way on the observable $(\mathbf{x}_{ij}, \mathbf{A}_{ij,k-1})$, but otherwise depends upon $\mathcal{F}$ only through u_{ijk}. Model (1) says that two individuals, j and j', with the same $(\mathbf{x}_{ij}, \mathbf{A}_{ij,k-1})$ differ in their conditional odds of treatment at step k by at most $\Gamma_k = \exp(\gamma_k) \ge 1$; that is, if $(\mathbf{x}_{ij}, \mathbf{A}_{ij,k-1}) = \left(\mathbf{x}_{ij'}, \mathbf{A}_{ij',k-1}\right)$ then

$$\frac{1}{\Gamma_k} \le \frac{\Pr\left(Z_{ijk} = 1 \,|\, \mathcal{F},\, \mathbf{A}_{ij,k-1}\right) \Pr\left(Z_{ij'k} = 0 \,\middle|\, \mathcal{F},\, \mathbf{A}_{ij',k-1}\right)}{\Pr\left(Z_{ij'k} = 1 \,\middle|\, \mathcal{F},\, \mathbf{A}_{ij',k-1}\right) \Pr\left(Z_{ijk} = 0 \,|\, \mathcal{F},\, \mathbf{A}_{ij,k-1}\right)} \le \Gamma_k.$$

If $\gamma_k = 0$, then there is no bias in assignment at step k, in the sense that everyone with the same observed $(\mathbf{x}_{ij}, \mathbf{A}_{ij,k-1})$ has the same probability of $Z_{ijk} = 1$, so assignment is ignorable at step k. Write $\mathbf{u}_k = (u_{11k}, \ldots, u_{I,n_I,k})^T$ and $\mathcal{U} = [0,1]^N$ for the N-dimensional unit cube. If $\mathcal{S}$ is a finite set, write $|\mathcal{S}|$ for the number of elements of $\mathcal{S}$.

If $\mathbf{a} \in \mathcal{A}_{k-1}$ is a $(k-1)$-dimensional vector of 0s and 1s, let $\mathcal{T}_{i,\mathbf{a}} \subseteq \{1, \ldots, n_i\}$ be the subset of individuals in stratum i with $\mathbf{A}_{ij,k-1} = \mathbf{a}$, let $n_{i,\mathbf{a}} = |\mathcal{T}_{i,\mathbf{a}}|$ and $m_{i,\mathbf{a}} = \sum_{j \in \mathcal{T}_{i,\mathbf{a}}} Z_{ijk}$. Write $\mathbf{m}_{ik}$ for the vector of dimension 2^{k-1} whose coordinates are the $m_{i,\mathbf{a}}$ for the 2^{k-1} possible values of $\mathbf{a} \in \mathcal{A}_{k-1}$. Again, it is convenient to use the same notation for $k = 1$, where $k - 1 = 0$, $\mathbf{a} = \emptyset$, $\mathcal{T}_{i,\mathbf{a}} = \{1, \ldots, n_i\}$, $n_{i,\mathbf{a}} = n_i$, $m_{i,\mathbf{a}} = \sum_{j=1}^{n_i} Z_{ij1}$. For $\mathbf{a} \in \mathcal{A}_{k-1}$,

let $Z_{i,\mathbf{a},\ell}$ be the value of $Z_{i\ell k}$ for individual $\ell \in \mathcal{T}_{i,\mathbf{a}}$, let $\mathbf{Z}_{i,\mathbf{a}}$ be the corresponding column vector of dimension $n_{i,\mathbf{a}}$ with $m_{i,\mathbf{a}}$ ones and $n_{i,\mathbf{a}} - m_{i,\mathbf{a}}$ zeros, and let $\mathbf{u}_{i,\mathbf{a}}$ be the column vector of dimension $n_{i,\mathbf{a}}$ with the u_{ijk} for individuals ij with $j \in \mathcal{T}_{i,\mathbf{a}}$. Let $\mathcal{Z}_{i,\mathbf{a}}$ be the set containing $\binom{n_{i,\mathbf{a}}}{m_{i,\mathbf{a}}}$ vectors such that $\mathbf{z} \in \mathcal{Z}_{i,\mathbf{a}}$ if $\mathbf{z}$ is of dimension $n_{i,\mathbf{a}}$ with $m_{i,\mathbf{a}}$ ones and $n_{i,\mathbf{a}} - m_{i,\mathbf{a}}$ zeros. All of these quantities and sets — $n_{i,\mathbf{a}}$, $\mathcal{T}_{i,\mathbf{a}}$, etc. — are random variables because $\mathbf{A}_{ij} = (Z_{ij1}, \ldots, Z_{ijK})$ is a random variable, but they are functions of $\mathbf{A}_{k-1}$ and $m_{i,\mathbf{a}}$, so they are fixed by conditioning on $\mathbf{A}_{k-1}$ and $\mathbf{m}_{ik}$. Then from (1),

$$\Pr\left(\mathbf{Z}_{i,\mathbf{a}} = \mathbf{z} \,|\, \mathcal{F},\, \mathbf{A}_{k-1}, \mathbf{m}_{ik}\right) = \frac{\exp\left(\gamma_k\, \mathbf{z}^T \mathbf{u}_{i,\mathbf{a}}\right)}{\sum_{\mathbf{b}\in\mathcal{Z}_{i,\mathbf{a}}} \exp\left(\gamma_k\, \mathbf{b}^T \mathbf{u}_{i,\mathbf{a}}\right)} \text{ for } \mathbf{z} \in \mathcal{Z}_{i,\mathbf{a}},\, \mathbf{u}_k \in \mathcal{U}, \qquad (2)$$

because $\kappa_k\left(\mathbf{x}_{ij}, \mathbf{A}_{ij,k-1}\right)$, though unknown, takes the same value for all individuals ij with $j \in \mathcal{T}_{i,\mathbf{a}}$. Moreover, the 2^{k-1} distinct $\mathbf{Z}_{i,\mathbf{a}}$ for the 2^{k-1} values of $\mathbf{a}$ are conditionally independent of each other given $\mathcal{F}$, $\mathbf{A}_{k-1}, \mathbf{m}_{ik}$. For fixed k, as $\mathbf{a} \in \mathcal{A}_{k-1}$ varies over its 2^{k-1} possible values, model (2) is a conventional model for stratified, treatment/control sensitivity analyses with $I \times 2^{k-1}$ strata; see Rosenbaum (2002, §4), Rosenbaum and Small (2017) and Rosenbaum (2018). If $\gamma_k = 0$, then (2) becomes random assignment, $\Pr\left(\mathbf{Z}_{i,\mathbf{a}} = \mathbf{z} \,|\, \mathcal{F},\, \mathbf{A}_{k-1}, \mathbf{m}_{ik}\right) = |\mathcal{Z}_{i,\mathbf{a}}|^{-1}$ for each $\mathbf{z} \in \mathcal{Z}_{i,\mathbf{a}}$, so model (2) permits one of the K steps to be free of bias from u_{ijk} while the other $K-1$ steps are biased.

3.3 The partial exclusion restriction

As there is typically uncertainty and contention about whether the exclusion restriction actually holds for possible instruments, we introduce a partial exclusion restriction. In effect, this condition says that some of the Z_{ijk} satisfy an exclusion restriction, others do not, consistent with some Z_{ijk} being instruments, while other Z_{ijk} require adjustments, similar to the adjustments for covariates.

If $(Z_{ij1}, \ldots, Z_{ijK})$ were K two-level treatments in a 2^K factorial experiment, then each

individual would have 2^K potential outcomes depending upon the 2^K ways that the K treatments $(Z_{ij1}, \ldots, Z_{ijK})$ might be set. In contrast, the assumption that $(Z_{ij1}, \ldots, Z_{ijK-1})$ are $K-1$ valid instruments for an active treatment Z_{ijK} entails, among other things, an exclusion restriction which says there are only two potential outcomes, r_{Tij} if $Z_{ijK} = 1$ or r_{Cij} if $Z_{ijK} = 0$. In words, the exclusion restriction says $(Z_{ij1}, \ldots, Z_{ijK-1})$ may push an individual ij towards treatment, $Z_{ijK} = 1$, or towards control, $Z_{ijK} = 0$, but it is only the active treatment, Z_{ijK}, that affects outcomes. In §2, the exclusion restriction says that being Catholic or being in an urban area affects your educational outcomes only indirectly to the extent to which it shifts you from a public to a Catholic high school, from $Z_{ijK} = 0$ to $Z_{ijK} = 1$. This may or may not be true. It is common to criticize conclusions based on a purported instrument by claiming that the exclusion restriction does not hold, for instance that Catholics should be compared to other Catholics because being Catholic is directly relevant to educational outcomes quite apart from attending Catholic school. To address such concerns, Definition 1 entertains the possibility that the exclusion restriction holds for parts of $(Z_{ij1}, \ldots, Z_{ijK})$ but not all of it.

Definition 1 *Let $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$, and let k be the smallest element in $\mathcal{K}$. The* ***partial exclusion restriction*** *holds for $\mathcal{K}$ if, with $\mathbf{A}_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})$ fixed by conditioning, each individual ij has two potential outcomes depending upon the value of Z_{ijK}, namely r_{Tij} if $Z_{ijK} = 1$ or r_{Cij} if $Z_{ijK} = 0$.*

A partial exclusion restriction places a restriction on $\mathbf{r}_{ij}$, saying that $r_{ij\mathbf{a}}$ for $\mathbf{a} = (a_1, \ldots, a_K)$ may vary with $(a_1, \ldots, a_{k-1})$ and a_K, but not with $(a_k, \ldots, a_{K-1})$. More specifically, this restriction says: if $\mathbf{a}$, $\mathbf{a}' \in \mathcal{A}_K$ with $\mathbf{a} = (a_1, \ldots, a_{k-1}, a_k, \ldots, a_{K-1}, a_K)$ and $\left(a_1, \ldots, a_{k-1}, a'_k, \ldots, a'_{K-1}, a_K\right)$, then $r_{ij\mathbf{a}} = r_{ij\mathbf{a}'}$, and we write $r_{ij\mathbf{a}} = r_{Cij}$ if $a_K = 0$ or $r_{ij\mathbf{a}} = r_{Tij}$ if $a_K = 1$ in an analysis that fixes $\mathbf{A}_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})$ by conditioning; however, $r_{ij\mathbf{a}}$ may vary with $(a_1, \ldots, a_{k-1})$, so this notation is meaningful only

with $\mathbf{A}_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})$ fixed, as it would be fixed if it were a covariate rather than an instrument. If a partial exclusion restriction holds for $\mathcal{K}$ and $\mathcal{K}' \subset \mathcal{K}$, then a partial exclusion restriction holds for $\mathcal{K}'$. If the partial exclusion restriction holds for $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$, then Fisher's hypothesis H_k of no effect at assignment k is the same null hypothesis for each $k \in \mathcal{K}$, namely $H_0 : r_{Tij} = r_{Cij}$, $\forall i, j$.

To clarify Definition 1, consider a few special cases. If $\mathcal{K} = \{1, 2, \ldots, K\}$, then partial exclusion is no different from the usual exclusion restriction for the $K - 1$ instruments jointly. If $\mathcal{K} = \{K\}$, then partial exclusion is simply the Neyman-Rubin notation for causal effects, with $\mathbf{A}_{ij,K-1} = (Z_{ij1}, \ldots, Z_{ij,K-1})$ fixed as covariates rather than instruments, that is, with $I \times 2^{K-1}$ strata defined by $(\mathbf{x}_{ij}, \mathbf{A}_{ij,K-1})$. In §2, partial exclusion for $\mathcal{K} = \{2, 3\}$ would be the usual exclusion restriction for 'being Catholic,' Z_{ij2}, if 'being urban or rural,' Z_{ij1}, were controlled as a covariate, that is, with $I \times 2$ strata. In §2, passing from $\mathcal{K} = \{1, 2, 3\}$ to $\mathcal{K} = \{2, 3\}$ entails two changes: first, 'being urban or rural,' Z_{ij1}, is no longer assumed to satisfy the exclusion restriction; second, 'being Catholic,' Z_{ij2}, is assumed to satisfy the exclusion restriction only after adjustment for 'being urban or rural.'

Definition 1 mentions a set $\mathcal{K}$ but makes use only of the smallest $k \in \mathcal{K}$: whether the partial exclusion restriction holds for the set $\mathcal{K}$ depends only on its smallest element. This will be convenient later, in particular in Definitions 2 and 3. There is more to a valid instrument than the exclusion restriction; the instrument must be randomized in a certain sense. An analysis might omit a potential instrument, $Z_{ij\ell}$, because of concern that $Z_{ij\ell}$ is not randomized. The analysis for $\mathcal{K} = \{1, 2, 3\}$ and $\mathcal{K}' = \{1, 3\}$ will entail the same partial exclusion restriction because the minimal element is $k = 1$ in both $\mathcal{K}$ and $\mathcal{K}'$, but the analysis for $\mathcal{K}' = \{1, 3\}$ will not use Z_{ij2}, so it will not require that Z_{ij2} be randomized. If the analyses for $\mathcal{K}$ and $\mathcal{K}'$ concur, then we are less worried that a doubtful assumption about Z_{ij2} is critical to the study's conclusions.

Suppose that the partial exclusion restriction holds for $\mathcal{K}$, and let k be the smallest element in $\mathcal{K}$. Then, by definition, an analysis that fixes $\mathbf{A}_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})$ by conditioning, by stratifying on $\mathbf{A}_{ij,k-1}$, has two potential outcomes, (r_{Tij}, r_{Cij}), for individual ij depending upon the value of Z_{ijK}. In this case, we may entertain the null hypothesis of a shift effect, $H_k^\beta : r_{Tij} = r_{Cij} + \beta$, so that $R_{ij} - \beta Z_{ijK} = r_{Cij}$ satisfies the null hypothesis of no effect. In the conventional way, we may invert a test of no effect to obtain a $1 - \alpha$ confidence interval for β, testing every possible β and retaining for interval the values of β not rejected at level α. If the partial exclusion restriction holds for $\mathcal{K}$, then the hypothesis $H_\ell^\beta : r_{Tij} = r_{Cij} + \beta$ is the same hypothesis for every $\ell \geq k$ and in particular for every $\ell \in \mathcal{K}$. As a consequence, we can ask whether several analyses concur in their assessment of the evidence against a specific value of β; that is, we are not restricted to asking about whether analyses concur in testing no effect.

3.4 Test statistics and sensitivity analyses

For each k, there is a stratified comparison of individuals with $Z_{ijk} = 1$ or $Z_{ijk} = 0$ within $I \times 2^{k-1}$ strata defined by the I original strata based on $\mathbf{x}_{ij}$, together with the 2^{k-1} strata defined by $\mathbf{A}_{ij,k-1} = (Z_{ij1}, \ldots, Z_{ij,k-1})$. A statistic testing H_0, say $T_k = t_k(\mathbf{Z}_k, \mathbf{R})$, at step k is a function of the observed responses, $\mathbf{R}$, and the treatment assignments, $\mathbf{Z}_k$, at step k. In principle, T_k may depend also on the $\mathbf{x}_{ij}$, the $\mathbf{A}_{ij,k-1}$, and the $\mathbf{m}_{ik}$, but the notation does not indicate this explicitly. In the current paper, T_k is van Elteren (1960)'s weighted combination of Wilcoxon rank sum statistics, but the Hodges-Lehmann aligned rank test or tests based on M-statistics are alternatives.

Consider the K steps one at a time, delaying consideration of their interdependence to §4. At step k, assume the partial exclusion restriction holds for $\mathcal{K} \subseteq \{k, k+1, \ldots, K\}$ with $k \in \mathcal{K}$. Then the sensitivity analysis at each step k has a conventional form, and

can be analyzed in a conventional way, as in Rosenbaum and Small (2017) and Rosenbaum (2018). If (2) were true at step k, and if $\gamma_k = 0$, then Fisher's hypothesis H_0 of no effect would be tested by comparing $T_k = t_k(\mathbf{Z}_k, \mathbf{R})$ to its stratified randomization distribution. For $\gamma_k = \log(\Gamma_k) > 0$, there is a P-value testing H_0 at the true value of $\mathbf{u}_k$ obtained, from elementary principles, by multiplying $I \times 2^{k-1}$ expressions of the form (2) over the $I \times 2^{k-1}$ strata to obtain the probability of a single possible value $\mathbf{z}_k$ of $\mathbf{Z}_k$, then summing such terms over all $\mathbf{z}_k$ such that $t_k(\mathbf{z}_k, \mathbf{R}) \geq t_k(\mathbf{Z}_k, \mathbf{R})$. This true P-value is not available to us because $\mathbf{u}_k$ is not observed, so we find the maximum such P-value over $\mathbf{u}_k \in \mathcal{U}$, say $\overline{P}_{k,\Gamma_k}$. To make the computations practical, a large sample approximation is used in place of the exact distribution in (2). If $\gamma_k = 0$, this maximum P-value is the randomization P-value, but as $\gamma_k \to \infty$ the bound $\overline{P}_{k,\Gamma_k} \to 1$, reflecting the familiar fact that an association, no matter how strong, does not logically entail causation — sufficiently large biases can explain away an association. The practical question is quantitative, not logical: How much bias, measured by $\Gamma_k = \exp(\gamma_k)$, would need to be present to render H_0 plausible?

4 Evidence factors: Combining the K steps

4.1 Valid or biased assignment

Perhaps some of the K comparisons are valid and others are not.

Definition 2 *Let $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$. The instrumental and direct comparisons in $\mathcal{K}$ are valid if: (i) partial exclusion holds for $\mathcal{K}$, and (ii) treatment assignment is governed by (2) with $\gamma_k = 0$ for each $k \in \mathcal{K}$.*

If the instrumental and direct comparisons in $\mathcal{K}$ were valid, then we could perform $|\mathcal{K}|$ separate valid tests of H_0 using stratified randomization inference, one for each $k \in \mathcal{K}$. For instance, in §2, if the instrumental comparisons in $\mathcal{K} = \{1, 2\}$ were valid, then: (i)

using "urban/rural", Z_{ij1}, alone as an instrument would yield a valid randomization test of H_0, (ii) using "being Catholic", Z_{ij2}, as an instrument within $2 \times I$ strata that controlled for "urban/rural", Z_{ij1}, would yield a valid randomization test of H_0, but (iii) the direct comparison, Z_{ij3}, of students in Catholic and public school, adjusting for Z_{ij1} and Z_{ij2}, may be biased by u_{ij3}.

In the absence of actual randomization, we cannot be sure a comparison is valid. Definition 3 refers to a measured degree of bias in some comparisons, with the possibility that other comparisons are severely biased.

Definition 3 *Let $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$. The comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$ if: (i) partial exclusion holds for $\mathcal{K}$, and (ii) for each $k \in \mathcal{K}$, treatment assignment is governed by (2) with $\gamma_k = \log(\Gamma_k)$ for some unknown $\mathbf{u}_k \in \mathcal{U}$.*

If the instrumental and direct comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$, then we could perform $|\mathcal{K}|$ separate stratified sensitivity analyses for $|\mathcal{K}|$ tests of H_0, with one upper bound $\overline{P}_{k,\Gamma_k}$ on the P-value for test $k \in \mathcal{K}$. This bound says: if H_0 is true and if the bias in treatment assignment in comparison $k \in \mathcal{K}$ is at most $\exp(\gamma_k) = \Gamma_k$, then the chance that $\overline{P}_{k,\Gamma_k} \leq \alpha$ is at most α. Moreover, each bound $\overline{P}_{k,\Gamma_k}$ is sharp, being attained for some $\mathbf{u}_k \in \mathcal{U}$ for $\gamma_k = \log(\Gamma_k)$. Definition 2 is Definition 3 with $\Gamma_k = 1$.

4.2 Evidence factors

If the comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$, then we may obtain $|\mathcal{K}|$ upper bounds $\overline{P}_{k,\Gamma_k}$ on valid P-values testing H_0, where these $|\mathcal{K}|$ tests make different assumptions about which instruments and comparisons are valid or biased to a limited degree. How are these $|\mathcal{K}|$ analyses related? Are they strongly dependent, merely repeating the same evidence in different forms? Is it nearly a foregone conclusion that the $|\mathcal{K}|$ comparisons will concur? Or are $|\mathcal{K}|$ comparisons nearly statistically independent, so that each comparison

provides new evidence? Proposition 4 shows that the $|\mathcal{K}|$ random variables $\overline{P}_{k,\Gamma_k}$ may be treated as if they were statistically independent P-values under H_0 if the comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$.

Proposition 4 *If H_0 is true and the comparisons in $\mathcal{K}$ are biased by at most $\Gamma_k \geq 1$, $k \in \mathcal{K}$, then the $|\mathcal{K}|$ bounds $\overline{P}_{k,\Gamma_k}$ on P-values testing H_0 are stochastically larger than the uniform distribution of the $|\mathcal{K}|$-dimensional unit cube.*

Proof. The proof uses Lemma 4 in Rosenbaum (2011) and runs parallel to the proof of Proposition 3 of that paper. The stratified structure with instruments in (1) is different from the structure in Rosenbaum (2011), but these differences do not affect the proof. ■

Being larger than the uniform distribution on the cube is not the same as being independent, but it suffices for hypothesis testing. Various methods combine $|\mathcal{K}|$ independent P-values, resulting in a single P-value. The combined statistic is a monotone function of the component P-values. Lemma 1 of Rosenbaum (2011) shows that such a combination yields a valid combined P-value when the components are stochastically larger than the uniform. Zaykin et al. (2002) combined independent P-values using the product of those P-values smaller than some truncation point, $\varkappa$, resulting in Fisher's method for $\varkappa = 1$. Hsu et al. (2013) show that the truncated product with $\varkappa = 0.1$ or $\varkappa = 0.2$ often has higher power than Fisher's method when applied to P-value bounds $\overline{P}_{k,\Gamma_k}$ from a sensitivity analysis, because the individual bounds are often larger than uniform on $[0, 1]$.

5 Evaluating the performance of the proposed analysis

5.1 A model for evaluating performance

The method in §4 considers assumptions that might identify causal effects, but it makes few other assumptions. To evaluate the performance of that method in comparison to

other methods, such as two-stage least squares, we consider a specific model for response R in terms of possibly invalid binary instruments Z_1 and Z_2 and treatment Z_3:

$$R = \alpha + \lambda_1 Z_1 + \lambda_2 Z_2 + \beta Z_3 + \epsilon \tag{3}$$

$$\zeta = \nu + \psi_1 Z_1 + \psi_2 Z_2 + \eta \;\text{ with } \mathrm{E}(\epsilon, \eta \,|\, Z_1, Z_2) = (0, 0)\,, \tag{4}$$

$$\Pr(\, Z_3 = 1 \,|\, Z_1, Z_2, \eta) = \max\{0,\, \min(1,\, \zeta)\} \tag{5}$$

where the bivariate (ϵ, η) are independent and identically distributed given Z_1, Z_2 with finite variances; see Small (2007) for a related model. We follow A. P. Dawid and write $A \perp\!\!\!\perp B \,\big|\, C$ for A is conditionally independent of B given C.

If ϵ and η are unrelated, then we do not need instruments. More precisely, if $\epsilon \perp\!\!\!\perp \eta \,\big|\, (Z_1, Z_2)$, then $\epsilon \perp\!\!\!\perp Z_3 \,\big|\, (Z_1, Z_2)$, and we can draw inferences about β using (3) alone, adjusting for (Z_1, Z_2), comparing the treated $Z_3 = 1$ and control $Z_3 = 0$ groups directly. Inference about β could be based on a least squares regression in (3), ignoring (4) and (5), or it could be based on the direct comparison of treated and control groups stratified for Z_1 and Z_2; that is, (iii) with $\gamma_3 = 0$ in §2 or step $k = 3$ in §3.

If ϵ and η were dependent given (Z_1, Z_2), but $\lambda_1 = \lambda_2 = 0$ with $\psi_1 \neq 0$ and $\psi_2 \neq 0$, then Z_1, Z_2 would be instruments for Z_3, so inference about β could be based on two-stage least squares. Factors $k = 1$ and $k = 2$ in §3 would each provide valid inferences about β with $\gamma_1 = 0$ and $\gamma_2 = 0$. If ϵ and η were dependent but either $\lambda_1 \neq 0$ or $\lambda_2 \neq 0$, then both least squares and two-stage least squares would not yield valid inferences for β.

If ϵ and η were dependent given (Z_1, Z_2), but $\lambda_2 = 0$ with $\psi_2 \neq 0$, then Z_2 would be a valid instrument for Z_3 after adjustment for Z_1. For instance, after stratifying for Z_1, factor $k = 2$ in §3 would provide valid inferences about β with $\gamma_2 = 0$. However, even factor $k = 2$ would be invalid if ϵ and η were dependent but $\lambda_2 \neq 0$.

5.2 Details of the model for numerical results

For numerical results, we specified the distributions in (3)-(5). Parameters were either fixed or variable, with some fixed parameters chosen to resemble actual distributions in the example in §2. In particular, we set $\Pr(Z_1 = 1) = 0.33$, $\Pr(Z_2 = 1) = 0.40$ as fixed parameters, but we measured dependence between Z_1 and Z_2 by a variable parameter $\delta = \Pr(Z_2 = 1|Z_1 = 1) - \Pr(Z_2 = 1|Z_1 = 0)$, with either $\delta = 0$ for independence or $\delta = 0.14$ for dependence. Given (Z_1, Z_2), the two errors (ϵ, η) were bivariate Normal with zero expectations, variable correlation ρ, and fixed variances 1 and 0.06. In the simulation, to resemble §2, we set $\psi_1 = 0.20$ and $\psi_1 = 0.25$, but in later calculations we varied these parameters to include weak instruments. We set $\nu = 0$, varying λ_1 and λ_2. In the simulation, the sample size is $N = 4450$, with $I = 178$ strata, $i = 1, \ldots, I$, of size $n_i = 25$, as in the example, whereas calculations of design sensitivity let $I \to \infty$ with $n_i = 25$. An on-line supplement expands the scope of the simulation.

5.3 Simulation of the probability of finding an effect when there is none

The simulation concerns the validity of various tests of a true hypothesis that $H_0 : \beta = \beta_0$. The simulation evaluates the size of a test that aspires to have level 0.05, so the test fails if it rejects a true null hypothesis at a rate above 0.05. Theory tells us whether a test is valid (V), with size at most equal to level, or biased (B), with size sometimes above level, and the judgement of theory is indicated by a B or V in Table 2. The simulated sizes agree with theory, but they add a quantitative assessment. Table 2 compares four methods: two-stage least squares using both (Z_1, Z_2), and the three evidence factors that use Z_j adjusting for $Z_{j-1}, \ldots, Z_1$.

Table 2 shows the results. Each situation was replicated 10,000 times, so that a proportion has a standard error of at most $\sqrt{0.25/10000} = 0.005$. Table 2 has sixteen

sampling situations. In case 5, for example, all three tests of H_0 based on instruments have the correct level, but the direct comparison is all but certain to falsely reject H_0.

All four tests are valid only in cases 1 and 9, in which both instruments and the direct comparison are valid. In total, two-stage least squares is valid in only four cases, 1, 5, 9, and 13, in which both instruments are valid.

Suppose that an investigator rejected H_0 only if all three evidence factors concur in rejecting H_0. Used in this way, the evidence factor analysis fails to provide a warning about biased comparisons only in cases 8, 15 and 16; it is valid in 13/16 cases. More formally, suppose at least one of the three factors is valid. Under this supposition, following Berger's (1982) reasoning about intersection-union tests, if we reject H_0 only when all three factors reject H_0, then we would falsely reject H_0 with probability at most 0.05.

A weaker standard in §6.4 uses the idea of partial conjunction from Benjamini and Heller (2008), saying that the evidence factors partially concur if at least two of them reject H_0. This weaker standard would fail, for instance, in case 4 where two factors are likely to falsely reject H_0; however, it provides protection whenever there is a single B in columns (i)-(iii). It provides protection in cases 2 and 3 with one invalid instrument, and in case 5 where both instruments are valid but the direct comparison is not.

To require evidence factors to concur is to require agreement among several nearly independent analyses that are valid under different assumptions. Table 2 shows this approach is not infallible, but it does offer substantially more protection than opting for any single analysis, say two-stage least squares or the direct treatment-control comparison.

5.4 Design sensitivity

Section 5.3 examined the protection afforded by evidence factors against falsely rejecting a true H_0. We now consider testing a false null hypothesis, one that we hope to reject

for valid reasons, that is, in the subset of valid analyses. Specifically, we test $H_0 : \beta = 0$ when in fact $\beta = 0.5$. So, we delete the biased analyses that may reject because of bias, and consider our prospects for rejecting H_0 in valid analyses.

Fix a sampling situation and let the sample size increase, $I \rightarrow \infty$. For each fixed Γ, the power of a sensitivity analysis tends either to 1 or to 0 as $I \rightarrow \infty$, depending upon the value of Γ. The transition point, $\widetilde{\Gamma}$, is called the design sensitivity: the power tends to 1 if $\Gamma < \widetilde{\Gamma}$ or to 0 if $\Gamma > \widetilde{\Gamma}$. In that sense, $\widetilde{\Gamma}$ is the limiting sensitivity to unmeasured bias when sampling variability has been eliminated by a sufficient increase in sample size.

Table 3 shows design sensitivities. In Table 3, instruments are sometimes weak, sometimes strong, and sometimes one is weak when the other is strong. It is known from theory that an analysis that uses a weak instrument is invariably sensitive to small biases, its design sensitivity $\widetilde{\Gamma}$ being barely larger than 1; see Small and Rosenbaum (2008).

Table 3 reminds us of a couple of basic quantitative facts. First, when there is a substantial treatment effect and no unmeasured bias, a direct comparison of treated and control groups may be insensitive to quite large biases. In a matched pair, a bias of $\Gamma = 2.5$ could be produced by an unobserved covariate u that increases the odds of treatment by a factor of 4 and increases the odds of a positive pair difference in outcomes by a factor of 6; see Rosenbaum and Silber (2009) and Rosenbaum (2017b, Table 9.1). Even if there is reason to worry that a direct comparison might be slightly biased, we may discover that it would have to be very biased to change the study's qualitative conclusion.

In contrast, the instrumental variable analyses in Table 3 are all sensitive to smaller biases. A bias of $\Gamma = 1.25$ is not trivially small: in a matched pair, it could be produced by an unobserved covariate that doubled the odds of treatment and doubled the odds of a positive pair difference in responses. In Table 3, design sensitivities of $\widetilde{\Gamma} \approx 1.2$ occur with strong instruments or $\widetilde{\Gamma} \approx 1.06$ with weak instruments. A weak instrument has to

be almost flawless to be convincing.

6 Effects of Catholic versus public high schools

6.1 Adjustments for observed covariates

In our examination of income from wages and salary for men in 1974, a preliminary step is to adjust for observed covariates. A simple analysis uses 178 strata defined by covariates, and a second analysis combines these 178 strata with a robust covariance adjustment. Following Kim (2011), we adjust for an IQ score prior to high school, father's and mother's education, parent's income, father's occupation score and occupational prestige score. Missing values in covariates were handled using the tactic in Rosenbaum and Rubin (1984, Appendix) in which treated subjects are compared to controls with a similar pattern of missing data.

Strata were built using the `blockingChallenge` package in `R`, where details may be found. The method samples 178 students, uses optimal matching to match 24 other students to each of the initial 178 students, making 178 blocks of size 25. The matching minimizes a robust covariate distance. The student in each block most distant from the remaining 24 is separated, and optimal matching is used again to pair these 178 individual students with 178 blocks of size 24. This process is repeated until no further changes are produced. The process was done 250 times, and we used the best stratification, that is, the one with the smallest total within-block distance. The name `blockingChallenge` invites efforts to build a better algorithms for minimum distance stratification.

For each covariate, there is an F-ratio in a one-way anova defined by the 178 strata, comparing variation between strata relative to within strata. There is substantial variation in the covariates between strata: $F = 225.6$ for IQ, $F = 61.2$ for the log of parental income, $F = 80.1$ and $F = 113.4$ for education of mother and father, respectively, $F = 44.3$ and $F = 41.4$ for occupation score and occupational prestige score, respectively. Although the

covariates do vary inside strata, they vary much less than a random sample.

The covariance adjustment used the method in Rosenbaum (2002b). The outcome, income from wages and salary in 1974, was regressed using M-estimation on indicators for the 178 strata plus the covariates themselves, and the residuals became the outcome to be analyzed by the method in §4. Importantly, this regression used the strata and covariates, but not the treatment variables in Table 1. Use of this form of covariance adjustment with an instrument is discussed and illustrated in Rosenbaum (2002b).

6.2 Naive analysis: each comparison is either flawless or useless

Table 4 performs the analyses from §4 on the wage data, testing three hypotheses, namely that Catholic schooling does not increase wages, $H_0 : \beta \leq 0$, that it increases wages by at most \$500, $H_0 : \beta \leq 500$, and that it increases wages by at most \$1000, $H_0 : \beta \leq 1000$. As the median annual wage in 1974 for these men was \$14000, an increase of \$500 is about 3.6%. Two analyses are performed for $H_0 : \beta \leq 0$, namely a stratified analysis, and a stratified analysis on residuals from covariance adjustment. Because there is no reason to prefer the merely stratified analysis, the latter analysis is presented in greater detail.

The current section assumes that each of the three comparisons is essentially a randomized experiment, once adjustments have been made for observed covariates. This is the situation with $\Gamma = 1$ in Table 4. The case of $\Gamma > 1$ is discussed in §6.3.

With or without covariance adjustment, each of the three comparisons rejects the null hypothesis of no effect of Catholic schools on wages, so the three evidence factors concur. These three factors depend upon very different assumptions, and they would be nearly statistically independent were the null hypothesis true, so it is news that the three analyses concur. When the three analyses are pooled using the truncated product of P-values, with the default truncation of 0.2, the resulting P-value is extremely small.

When testing the hypothesis that the effect is at most \$500, the situation is quite different. Two analyses reject \$500 as too small, but the remaining factor does not concur. The pooled analysis is significant because of the urban/rural comparison; remove that, and the pooled P-value from the two remaining factors is 0.103. A similar pattern is seen when testing that the effect is at most \$1000.

6.3 Sensitivity analysis: allowing for small or moderate imperfections in each analysis

How might small unmeasured biases alter the analyses in §6.2? Table 4 considers biases of $\Gamma = 1.1$, 1.2, and 1.25. The parameter Γ has the same meaning in matched pairs and in strata, but it is easiest to understand the paired case; see Rosenbaum and Silber (2009) and Rosenbaum (2017b, Table 9.1) for detailed discussion of the issues discussed in this paragraph. If we paired people based on covariates and flipped a fair coin to assign one person in the pair to live in an urban area, or to be Catholic, or to attend Catholic school, then each person in the pair would have probability 1/2 of each of these assignments. If $\Gamma = 1.25$, then the same probability is somewhere in the interval [0.444, 0.556] rather than 1/2. This way of describing Γ focuses upon the impact of the unobserved covariate on treatment assignment, but Γ may be understood, equivalently, in terms of an unobserved covariate affecting both treatment assignment and outcome. Specifically, Γ may be given a two-parameter interpretation, $\Gamma = (\Lambda\Delta + 1)\,/\,(\Lambda + \Delta)$, called an amplification, where Δ controls the association between the unobserved covariate and the outcome, while Λ controls the association between the unobserved covariate and the treatment. As $\Gamma = 1.25 = (2 \times 2 + 1)\,/\,(2 + 2) = (\Lambda\Delta + 1)\,/\,(\Lambda + \Delta)$, in a matched pair, a bias of $\Gamma = 1.25$ is the same as an unobserved covariate that doubles the odds of treatment, $\Lambda = 2$, and doubles the odds, $\Delta = 2$, of a positive pair difference in wages, so a bias of $\Gamma = 1.25$ is neither extremely large nor trivially small. A bias of $\Gamma = 1.05$ is small, and would be hard

to rule out based on a priori considerations in most observational studies.

In Table 4, the pooled test of no effect of Catholic school using stratification and covariance is insensitive to a bias $\Gamma = 1.2$; however, this is entirely due to the contribution of the urban/rural instrument. Without the urban/rural instrument, the pooled test of no effect using the other two factors is sensitive at $\Gamma = 1.2$.

So, the analysis depends rather heavily on the validity of urban/rural as an instrument. The instrumental variable analysis notes higher wages for students from urban areas, and attributes that difference in wages to a higher frequency of Catholic schooling in urban areas. That attribution is suspect here. Among non-Catholics attending public school, median wages were higher in urban areas, a median of \$15000 in urban areas versus \$13000 in rural areas. Among Catholics attending public schools, median wages were higher in urban areas, a median of \$14000 in urban areas versus \$13400 in rural areas. It is a concern that the analysis depends so heavily on the urban/rural instrument, as it is plausible that wages are higher for students from urban areas for reasons other than Catholic schooling.

6.4 Partial conjunction

As just noted, the combined analyses in Table 4 lean heavily on the validity of urban/rural as an instrument. Is it possible to quantify the degree to which a combined analysis depends upon one of its components? How large can Γ be while still securing concurrence in rejecting H_0 by at least two components?

Partial conjunction hypotheses ask for concurrence among at least $\overline{K}$ of K sources of evidence, $1 < \overline{K} < K$, without specifying in advance which $\overline{K}$ sources will concur. In Table 4, $K = 3$ so the only possible value of $\overline{K}$ is 2. The partial conjunction null hypothesis asserts that at most $\overline{K}-1$ null hypotheses are false, so rejection of that hypothesis entails at least $\overline{K}$ null hypotheses are false. In Table 4, we seek strong evidence that at least $\overline{K} = 2$

factors concur in rejecting H_0. Benjamini and Heller (2008) and Wang and Owen (2019) propose methods for partial conjunction hypotheses. Applying their results to Proposition 4, we may reject at level α the $\overline{K}$ partial conjunction hypothesis in the presence of biases of at most Γ_k, $k = 1, \ldots, K$, if the P-value determined by the truncated product is $\leq \alpha$ when computed from $K - \overline{K} + 1 = 3 - 2 + 1 = 2$ largest $\overline{P}_{k,\Gamma_k}$. In Table 4, the smallest $\overline{K} - 1 = 1$ smallest $\overline{P}_{k,\Gamma_k}$ is always from $k = 1$ for the urban/rural comparison; however, this method acknowledges that we did not know that prior to examining the data.

Consider partial conjunction testing of $H_0 : \beta = 0$ using both stratification and covariance adjustment in Table 4. In randomization tests, $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1$, applying the truncated product to $\overline{P}_{2,\Gamma_2} = 0.0065$ and $\overline{P}_{3,\Gamma_3} = 0.0149$ yields a P-value of 0.00084, so at least two factors concur in rejecting H_0. At $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1.1$, the two bounds, $\overline{P}_{2,\Gamma_2} = 0.1115$ and $\overline{P}_{3,\Gamma_3} = 0.0667$, combine to yield a P-value of 0.0319; however, at $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1.2$, the combined P-value is 0.34. In short, at least two factors concur in rejecting H_0 if the unmeasured bias is quite small, $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1.1$, but for larger biases, rejection depends entirely on the validity of the urban/rural instrument.

7 Discussion

Conventional analyses with two or more instruments, such as two-stage least squares, assume all instruments are jointly flawless, and ignore the direct comparison of treated and control groups. In contrast, our proposed analysis assumes less and reveals more. With $K-1$ instruments, we produce K essentially independent comparisons that each make very different assumptions for the validity of different comparisons, successively changing the role of each instrument from instrument to covariate. To a considerable extent, neither identical assumptions nor chance can explain concurrence among these K analyses, whereas an actual causal effect could explain concurrence. Each analysis is subjected to a sensitivity

analysis that quantitatively evaluates the gradual failure of the assumptions upon which that one analysis depends. A partial conjunction analysis asks about the evidence that remains when the quantitatively most compelling analyses are set aside.

In their formulation, with a binary instrument and a binary treatment, Angrist, Imbens and Rubin (1996) show that a valid instrument yields a consistent estimate of the average effect of a treatment on compliers, that is, on the subpopulation that accepts the treatment if and only if the instrument encourages them to do so. Deaton (2009) argued that an effect on the subpopulation of compliers to a particular instrument is unlikely to be of interest unless similar effects, or understandably different effects, are produced in other subpopulations. Imbens (2010) countered that it is best to be candid about what instruments estimate, even if this candid description is less than we might prefer. Our sense is that a comparison of several analyses with different instruments, as in §6, speaks to Deaton's concerns while meeting Imbens' standard for candor. If an evidence factor analysis produces similar inferences about effects using very different instruments, then this is not incompatible with an effect that does not vary greatly with the subpopulation of compliers moved by a particular instrument. Conversely, if inferences about effect vary from one factor to the next, then this raises concerns either about the validity of the instrument or about the impact of the changing subpopulation of compliers.

Because instruments, particularly weak instruments, are greatly unsettled by the slightest flaw, theory suggests that the least sensitive finding — more precisely, the finding with the largest design sensitivity $\widetilde{\Gamma}$ — is expected to come from the direct comparison of treated and control groups when, indeed, there is a causal effect without bias. This suggests, the direct comparison should be one of the K factors considered and displayed. True, the direct comparison may be the most biased comparison, so we might tolerate sensitivity to smaller biases in an instrument than in a direct comparison.

In the study of the effects of Catholic high schools, only one of the three evidence factors points to an effect of Catholic schools that is insensitive to moderately small biases, namely the factor that attributes higher incomes in urban areas to greater attendance of Catholic schools in those areas. The other two evidence factors do not concur, neither the comparison of Catholics and non-Catholics, nor the comparison of students attending Catholic schools and students attending traditional public schools. As it is not implausible that the urban/rural distinction acts on income in multiple ways, not just in virtue of attendance in Catholic schools, in violation of the exclusion restriction, doubts are raised about the strength of the evidence that Catholic schools are more effective than traditional public schools, a conclusion compatible with the concerns raised by Altonji et al. (2005).

References

Altonji, J.G., Elder, T.E., Taber, C.R. (2005), "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 113, 151-184.

Angrist, J.D., Imbens, G.W., Rubin, D.B. (1996), "Identification of causal effects using instrumental variables (with Discussion)," *Journal of the American Statistical Association*, 91, 444-455.

Baiocchi, M., Cheng, J., Small, D.S. (2014), "Instrumental variable methods for causal inference," *Statistics in Medicine*, 33, 2297-2340.

Benjamini, Y, Heller, R. (2008), "Screening for partial conjunction hypotheses," *Biometrics*, 64, 1215-1222.

Berger, R.L. (1982), "Multiparameter hypothesis testing and acceptance sampling," *Technometrics*, 24, 295-300.

Brookhart, M.A., Rassen, J.A, Schneeweiss, S. (2010), "Instrumental variable methods in

comparative safety and effectiveness research," *Pharmacoepidemiology and Drug Safety*, 19, 537-554.

Card, D., Dooley, M.D, Payne, A.A. (2010), "School competition and efficiency with publicly funded Catholic achools," *Applied Economics*, 2, 150-176.

Coleman, J., Hoffer, T., Kilgore, S. (1982), "Cognitive outcomes in public and private schools," *Sociology of Education*, 55 65-76.

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

Deaton, A. (2010), "Instruments, randomization, and learning about development," *Journal of Economic Literature*, 48, 424-455.

Garcia, D.R. (2018), *School Choice*, Cambridge, MA: MIT Press.

Gilbert, P., Bosch, R., and Hudgens, M. (2003), "Sensitivity analysis for the assessment of the causal vaccine effects on viral load in HIV vaccine trials," *Biometrics*, 59, 531-41.

Goldberger, A.S., Cain, G.G. (1982), "The causal analysis of cognitive outcomes in the Coleman, Hoffer and Kilgore report," *Sociology of Education*, 55, 103-122.

Hogan, J.W., Lancaster, T. (2004), "Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies," *Statistical Methods in Medical Research*, 13, 17-48.

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

Hoxby, C.M. (2000), "Does competition among public schools benefit students and taxpayers?" *American Economic Review*, 90, 1209-1238.

Hsu, J.Y., Small, D.S., Rosenbaum, P.R. (2013), "Effect modification and design sensitivity

in observational studies," *Journal of the American Statistical Association*, 108, 135-148.

Imbens, G.W. (2010), "Better LATE than nothing," *Journal of Economic Literature*, 48, 399-423.

Imbens, G.W., Rosenbaum, P.R. (2005), "Robust, accurate confidence intervals with a weak instrument," *Journal of the Royal Statistical Society* A, 168, 109-126.

Kang, H., Zhang, A., Cai, T.T., Small, D.S. (2016), "Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization," *Journal of the American Statistical Association*, 111, 132-144.

Kang, H. (2016), "Matched instrumental variables: A possible solution to severe confounding in matched observational studies," *Epidemiology*, 27, 633-636.

Keele, L., Morgan, J.W.. (2016), "How strong is strong enough? Strengthening instruments through matching and weak instrument tests," *Annals of Applied Statistics*, 10, 1086-1106.

Kim, Y.J. (2011), "Catholic schools or school quality? The effects of Catholic schools on labor market outcomes," *Economics of Education Review*, 30, 546-558.

Larcker, D.F., Rusticus, T.O. (2010), "On the use of instrumental variables in accounting research," *Journal of Accounting and Economics*, 49, 186-205.

Li, Y., Lee, Y., Wolfe, R.A., Morgenstern, H., Zhang, J., Port, F.K., Robinson, B.M. (2015), "On a preference-based instrumental variable approach in reducing unmeasured confounding-by-indication," *Statistics in Medicine*, 34, 1150-1168.

Lu, B., Marcus, S. (2012), "Evaluating long-term effects of a psychiatric treatment using instrumental variable and matching approaches," *Health Services and Outcomes Research Methodology*, 12, 288-301.

Neal, D. (1997), "The effects of Catholic secondary schooling on educational achievement," *Journal of Labor Economics*, 15, 98-123.

Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Statistical Science*, 5, 463-480.

Rosenbaum, P.R. (2002a), *Observational Studies* (2nd edition), New York: Springer.

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

Rosenbaum, P.R. (2010), "Evidence factors in observational studies," *Biometrika*, 97, 333-345.

Rosenbaum, P.R. (2011), "Some approximate evidence factors in observational studies," *Journal of the American Statistical Association*, 106(493), 285-295.

Rosenbaum, P.R. (2017a), "The general structure of evidence factors in observational studies," *Statistical Science*, 32, 514-530.

Rosenbaum, P.R. (2017b), *Observation and Experiment*, Harvard University Press.

Rosenbaum, P.R. (2018), "Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels," *Annals of Applied Statistics*, 12, 2312-2334. (`R` package `senstrat`)

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

Rosenbaum, P.R., Small, D.S. (2017), "An adaptive Mantel–Haenszel test for sensitivity analysis in observational studies," *Biometrics*, 73(2), 422-430. (`R` package `sensitivity2x2xk`)

Rubin, D.B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688-701.

Rudolph, K.E., Stuart, E.A. (2017), "Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods," *American Journal of Epidemiology*, 187, 604-613.

Schwartz, S., Li, F., Reiter, J.P. (2012), "Sensitivity analysis for unmeasured confounding

in principal stratification settings with binary variables," *Statistics in Medicine*, 31, 949-962.

Small, D.S. (2007), "Sensitivity analysis for instrumental variables regression with overidentifying restrictions," *Journal of the American Statistical Association*, 102, 1049-1058.

Small, D.S., Rosenbaum, P.R. (2008), "War and wages: the strength of instrumental variables and their sensitivity to unobserved biases," *Journal of the American Statistical Association*, 103, 924-933.

Wang, J., Owen, A.B. (2019), "Admissibility in partial conjunction testing," *Journal of the American Statistical Association*, 114, 158-168.

van Elteren, P.H. (1960), "On the contribution of independent two sample tests of Wilcoxon," *Bulletin of the International Statistical Institute*, 12, 351-361.

Yu, B.B., Gastwirth, J.L. (2005), "Sensitivity analysis for trend tests," *Biostatistics*, 6, 201-209.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., & Weir, B.S. (2002), "Truncated product method for combining P-values," *Genetic Epidemiology*, 22(2), 170-185.

Zhang, K., Small, D.S., Lorch, S., Srinivas, S., Rosenbaum, P.R. (2011), "Using split samples and evidence factors in an observational study of neonatal outcomes," *Journal of the American Statistical Association*, 106, 511-524.

Zubizarreta, J.R., Neuman, M., Silber, J., Rosenbaum, P.R. (2012), "Contrasting evidence within and between institutions that provide treatment in an observational study of forms of anesthesia," *Journal of the American Statistical Association*, 107, 901-915.

Table 1: Counts and percents attending Catholic school for two potential instruments and a direct comparison of Catholic and public schools.

Group			Count			% Attending Catholic School		
Urban	Religion	School	Urban	Religion	School	Urban	Religion	School
Urban	Catholic	Catholic	1501	741	327	22	44	100
		Public			414			0
	Other	Catholic		760	0		0	0/0
		Public			760			0
Rural	Catholic	Catholic	2949	1045	177	6	17	100
		Public			868			0
	Other	Catholic		1904	2		0	100
		Public			1902			0
		Total	4450	4450	4450			

Table 2: Simulated probability of falsely rejecting, at the 0.05 level, the true null hypothesis $H_0 : \beta = \beta_0$ using two-stage least squares (TSLS), (i) using the first binary instrument alone, (ii) using the second binary instrument stratifying for the first instrument, (iii) using a direct comparison of treated and control groups stratifying for both instruments.

	Parameters				Valid (V) or asymptotically biased (B)				Probability of rejecting $H_0 : \beta = \beta_0$			
						Our method				Our method		
Case	λ_1	λ_2	ρ	δ	TSLS	(i)	(ii)	(iii)	TSLS	(i)	(ii)	(iii)
1	0	0	0	0	V	V	V	V	0.05	0.05	0.05	0.05
2	0.10	0	0	0	B	B	V	V	0.48	0.91	0.05	0.05
3	0	0.10	0	0	B	V	B	V	0.73	0.05	0.90	0.05
4	0.10	0.10	0	0	B	B	B	V	1.00	0.91	0.90	0.05
5	0	0	0.82	0	V	V	V	B	0.05	0.04	0.05	1.00
6	0.10	0	0.82	0	B	B	V	B	0.52	0.91	0.05	1.00
7	0	0.10	0.82	0	B	V	B	B	0.75	0.05	0.91	1.00
8	0.10	0.10	0.82	0	B	B	B	B	1.00	0.91	0.91	1.00
9	0	0	0	0.14	V	V	V	V	0.05	0.05	0.05	0.05
10	0.10	0	0	0.14	B	B	V	V	0.57	0.91	0.05	0.05
11	0	0.10	0	0.14	B	B	B	V	0.76	0.10	0.90	0.05
12	0.10	0.10	0	0.14	B	B	B	V	1.00	0.96	0.90	0.05
13	0	0	0.82	0.14	V	V	V	B	0.05	0.05	0.05	1.00
14	0.10	0	0.82	0.14	B	B	V	B	0.60	0.91	0.05	1.00
15	0	0.10	0.82	0.14	B	B	B	B	0.78	0.11	0.89	1.00
16	0.10	0.10	0.82	0.14	B	B	B	B	1.00	0.96	0.90	1.00

Table 3: Design sensitivities $\widetilde{\Gamma}$ for valid analyses. For biased analyses, a "B" is given in place of a design sensitivity.

Parameters				Two strong instruments $\psi_1 = 0.20, \psi_2 = 0.25$			Two weak instruments $\psi_1 = 0.09, \psi_2 = 0.09$		
λ_1	λ_2	ρ	δ	(i)	(ii)	(iii)	(i)	(ii)	(iii)
0	0	0	0	1.18	1.21	2.58	1.06	1.06	2.59
0.10	0	0	0	B	1.21	2.58	B	1.06	2.60
0	0.10	0	0	1.17	B	2.57	1.06	B	2.60
0.10	0.10	0	0	B	B	2.57	B	B	2.60
0	0	0.82	0	1.17	1.21	B	1.06	1.06	B
0.10	0	0.82	0	B	1.20	B	B	1.05	B
0	0.10	0.82	0	1.17	B	B	1.06	B	B
0.10	0.10	0.82	0	B	B	B	B	B	B
0	0	0	0.14	1.21	1.21	2.56	1.07	1.06	2.59
0.10	0	0	0.14	B	1.21	2.56	B	1.06	2.59
0	0.10	0	0.14	B	B	2.56	B	B	2.60
0.10	0.10	0	0.14	B	B	2.56	B	B	2.59
0	0	0.82	0.14	1.21	1.21	B	1.07	1.06	B
0.10	0	0.82	0.14	B	1.21	B	B	1.06	B
0	0.10	0.82	0.14	B	B	B	B	B	B
0.10	0.10	0.82	0.14	B	B	B	B	B	B

Parameters				Z_1 weak, Z_2 strong $\psi_1 = 0.09, \psi_2 = 0.25$			Z_1 strong, Z_2 weak $\psi_1 = 0.20, \psi_2 = .09$		
λ_1	λ_2	ρ	δ	(i)	(ii)	(iii)	(i)	(ii)	(iii)
0	0	0	0	1.07	1.21	2.58	1.16	1.06	2.58
0.10	0	0	0	B	1.20	2.58	B	1.06	2.57
0	0.10	0	0	1.07	B	2.58	1.16	B	2.58
0.10	0.10	0	0	B	B	2.59	B	B	2.59
0	0	0.82	0	1.07	1.19	B	1.15	1.06	B
0.10	0	0.82	0	B	1.19	B	B	1.06	B
0	0.10	0.82	0	1.07	B	B	1.16	B	B
0.10	0.10	0.82	0	B	B	B	B	B	B
0	0	0	0.14	1.10	1.20	2.58	1.17	1.06	2.57
0.10	0	0	0.14	B	1.20	2.58	B	1.06	2.58
0	0.10	0	0.14	B	B	2.58	B	B	2.57
0.10	0.10	0	0.14	B	B	2.57	B	B	2.57
0	0	0.82	0.14	1.10	1.20	B	1.17	1.06	B
0.10	0	0.82	0.14	B	1.19	B	B	1.06	B
0	0.10	0.82	0.14	B	B	B	B	B	B
0.10	0.10	0.82	0.14	B	B	B	B	B	B

Table 4: Three evidence factors and their combination using the truncated product, with and without covariance adjustment. The case $\Gamma = 1$ assumes comparisons are flawless, three stratified randomized experiments. The table shows one equivalent amplification of each $\Gamma > 1$. The table displays upper bounds on one-sided P-values testing the null hypothesis that Catholic schooling raises wages by at most β dollars in the presence of a bias of at most Γ. As the median annual wage was \$14000, a \$500 increase is about 3.6%.

Sensitivity	Equivalent	3 Independent Factors			
Parameter	Amplification	Urban/Rural	Religion	Direct	Combined
Γ	(Λ, Δ)	Stratifed analysis			
		$H_0 : \beta \leq \$0$			
1	(1, 1)	0.0000	0.0041	0.0082	0.0000
1.1	(1.4, 1.8)	0.0000	0.0835	0.0422	0.0000
1.2	(1.75, 2)	0.0004	0.4095	0.1331	0.0022
1.25	(2, 2)	0.0023	0.6225	0.2049	0.0330
Γ	(Λ, Δ)	Stratifed + covariance adjustment			
		$H_0 : \beta \leq \$0$			
1	(1, 1)	0.0000	0.0065	0.0149	0.0000
1.1	(1.4, 1.8)	0.0001	0.1115	0.0667	0.0001
1.2	(1.75, 2)	0.0048	0.4738	0.1876	0.0170
1.25	(2, 2)	0.0182	0.6827	0.2747	0.1211
		$H_0 : \beta \leq \$500$			
1	(1, 1)	0.0000	0.0394	0.2013	0.0000
1.1	(1.4, 1.8)	0.0005	0.3105	0.4345	0.0110
1.2	(1.75, 2)	0.0128	0.7451	0.6735	0.0982
		$H_0 : \beta \leq \$1000$			
1	(1, 1)	0.0000	0.1304	0.6975	0.0002
1.1	(1.4, 1.8)	0.0016	0.5550	0.8826	0.0258
1.2	(1.75, 2)	0.0303	0.9018	0.9643	0.1592