

Indices for the Robustness of Causal Inferences for the Counterfactual

Kenneth A. Frank
Michigan State University

Paper presented at the Spring meetings of the ASA Methodology section, Ann Arbor, MI, April 2004

Direct correspondence to
Kenneth A. Frank
Visiting Professor
Population Research Center
The University of Texas at Austin
1 University Station G1800
Austin Texas, 78712-0544

e-mail kenfrank@msu.edu
phone: 512-419-0216
Fax: 512-471-4886

Author's footnote:

Kenneth Frank is Associate professor of Counseling, Educational Psychology and Special Education, and of Fisheries and Wildlife. He can be contacted at room 460 Erickson Hall, Michigan State University, East Lansing, MI 48824-1034 (e-mail: kenfrank@msu.edu). The author acknowledges the support of grants from the National Institute of Child Health and Human Development (R01 HD40428-02, PI: Chandra Muller) and the National Science Foundation (REC-0126167, Co-PI: Chandra Muller and Pedro Reyes) to the Population Research Center, University of Texas at Austin. Opinions reflect those of the author and not necessarily those of the granting agencies. The author is indebted to Sam Field, Raymond Mapuranga, Kyung-Seok Min, Stephen Morgan, Wei Pan and Howard Wainer for comments on earlier drafts and to Yea Lin Lin for conducting analyses on NELS data to confirm results from Morgan.
April 14, 2004

Abstract

The fundamental assumptions for causal inference are typically not sustained by the research designs and analyses used by social scientists. Thus robustness indices are generated to quantify how much the assumptions of causal inference, as defined by the counterfactual, must be violated to alter a statistical inference. The indices are defined in terms of an intuitive metric: the proportion of an observed treatment effect that would have to be attributed to violations of assumptions to alter an inference. The indices are also related to the quantities defining confidence intervals and effect sizes, linking to alternative thresholds for making causal inferences. Examples pertain to inferences regarding the effects of Catholic schools and class size on academic achievement. The discussion assesses the value of the counterfactual paradigm for generating robustness indices and the conclusion expands upon a metaphor linking statistical and causal inference.

Key Words: Causal Inference, Robustness, Counterfactual, Sensitivity Analysis.

“Assumptions are the strands that join the field of statistics to scientific disciplines” (Little and Rubin 2000, page 123).

Introduction

The fundamental assumptions for causal inference are typically not sustained by the research designs and analyses used by social scientists. Thus I generate robustness indices to quantify how much the assumptions of causal inference, as defined by the counterfactual, must be violated to alter a causal inference from a statistical analysis. Ultimately these indices provide a precise language in an intuitive metric for debate regarding the validity of causal inferences.

Robustness indices, as one form of sensitivity analysis, derive their meaning relative to specific inferences. Therefore, I motivate this paper through examples regarding inferences of the effects of Catholic schools and small classes on academic achievement. Interestingly, both findings link back to James Coleman’s analyses of school effects. First, Coleman and colleagues found that students who attended Catholic schools had higher, and statistically significant, levels of mathematics achievement in twelfth grade than students who attended public schools (Coleman and Hoffer 1987; Coleman, Hoffer and Kilgore 1982; Hoffer 1985). These findings sparked a line of research regarding the effect of Catholic schools that persists (Alexander and Pallas 1983, 1985; Goldberger and Cain 1982; Greene, Peterson and Du 1998; Howell and Peterson 2000; McEwan 2000; Morgan 1983; Morgan 2001).

Second, Coleman et al. (1966) focused on resource inputs (e.g., per pupil expenditures, teacher training, etc) and outputs (e.g., achievement), attending less to teaching style, curricular materials, and the like. Thus Coleman et al.’s analysis of equality and educational opportunity gave rise to studies of the production function of education. In particular, one strand of research on this

production function now focuses on the effects of per-pupil ratios and class size (e.g., Finn and Achilles 1990; Hanushek 1999).

Important causal inferences have been made from each line of research. From the finding that students in Catholic schools perform better than those in public schools some made the causal inference that Catholic schools educate students better than public schools. This causal inference was then used to support the policy that students should be given vouchers to attend any school: public, private, or religious (Chubb and Moe 1990; Coleman, Hoffer and Kilgore 1982). Similarly, with respect to class size, Finn and Achilles argued: “This research leaves no doubt that small classes have an advantage over larger classes in reading and mathematics in the early primary grades” (page 573). Correspondingly, the United States government and some states have allocated resources to reduce class size (e.g., California Board of Education 2001; US department of Education 2001).

Aside from policy implications, the two lines of research emanating from the Coleman report have been central to methodological and philosophical debates regarding causal inferences in the social sciences. The causal inferences regarding Catholic schools have been primarily based on observational studies (e.g., High School and Beyond – HS&B, National Educational Longitudinal Study of 1988–NELS88) featuring Catholic schools as the “treatment” to which students are assigned, but not at random (Cochran 1965; Rosenbaum 2002)¹. The strength of the observational studies of Catholic schools is their external validity deriving from nationally representative samples

¹There is some ambiguity regarding the characterization of HS&B and NELS88 as observational studies of the effect of Catholic schools because the treatment, Catholic schools, may not be uniform. For example, at the very least Catholic schools differ by region, student composition, leadership, faculty, and the like which could generate variation in the educational “treatment” and thus the effect. On the other hand, Catholic schools are typified by a few key features, such as placing a large proportion of students in the academic track, that distinguishes them from most public schools (Bryk et al 1992; Lee and Bryk 1988). Therefore analytically I will consider attendance at Catholic school as a single treatment, as have previous researchers in this area (Coleman and Hoffer 1987; Coleman, Hoffer and Kilgore 1982; Hoffer 1985; Morgan 2001).

(when sample weights are applied). Thus any findings valid across the sample can be generalized to the corresponding national cohort.

Of course the weakness of these observational studies is in their limited internal validity because of the non-random assignment to any given experience or treatment that is the focus of study (e.g., Alexander and Pallas 1983, 1985; Rosenbaum 2002; Shadish et al 2002). As Alexander and Pallas (1983) put it: “Probably the single greatest burden of school effects research is to distinguish convincingly between outcome differences that reflect simply differences in the kinds of students who attend various schools from differences that are attributable to something about the schools themselves” (page 170). Indeed, this is a fundamental problem for any sociological research that attempts to estimate effects of social organizations into which people select themselves.

The strengths and weaknesses of the Finn and Achilles’ class size experiments complement those of observational studies. In particular, the class size experiments have strong internal validity because classrooms were randomly assigned to be of differing sizes. But the Tennessee classrooms in Finn and Achilles’ study were not a random sample from Tennessee or the nation, and thus results may not generalize beyond the sample – external validity is weak. In particular, Hanushek (1999) argues that the effects of class size varied by grade and were somewhat unique to Tennessee, and therefore not generalizable to policies to reduce class size in other contexts, such as California (see Hanushek, 1999, page 143).

The traditional critiques of the observational study as lacking internal validity and randomized experiments as lacking external validity have been vigorously defended for the inferences regarding Catholic schools and small classes. Studies of the effects of Catholic schools have leveraged longitudinal designs and extensive survey questions to control for pre-measures of achievement and a rich set of covariates (e.g., Coleman et al. 1982; Morgan 2001). Further, some studies (e.g, Noell 1982) have used instrument variables (Angrist, Imbens and Rubin 1996; Heckman and Robb 1985)

and propensity score matching (Morgan 2001) to control for possible selection bias due to differences between those who attended Catholic and public schools (Rosenbaum 2002; Rosenbaum and Rubin 1983). Anticipating challenges regarding external validity, Finn and Achilles support their inference by reporting that the schools studied were very similar to others in Tennessee in terms of teacher-pupil ratios and percentages of teachers with higher degrees (pages 559-560).

The contentious debates regarding the effects of Catholic schools and small classes reflect a fundamental limitation to full randomization of selection of a sample *and* assignment to treatment conditions in the social sciences. If a researcher randomly assigns a treatment to a randomly selected subject then the researcher has exerted undue control in randomly determining the fate of the subject. Thus social scientists must typically compromise the implementation of full randomization either in selection of a sample or in assignment to treatment conditions. That is, social scientists typically must choose between observing random samples or randomly assigning volunteer subjects to conditions. The limitations to full randomization then lead to restrictions on internal or external validity; it is a rare day when a social scientist can make a causal inference, claiming internal *and* external validity, with certainty.

Because of limitations in study designs in the social sciences, some have suggested that social scientists should not make causal inferences (Abbott 1998; Sobel 1998). But in his introduction of the counterfactual, Rubin (1978, page 688) anticipated such criticisms with respect to non-randomized experiments:

Recent psychological and educational literature has included extensive criticism of the use of nonrandomized studies [e.g., Campbell and Erlebacher 1970]. The implication in much of this literature is that only properly randomized experiments can lead to useful estimates of causal effects. If taken as applying to all fields of study, this position is untenable. Since extensive use of randomized experiments is limited to the last century², and in fact is *not* used

²Essentially since Fisher (1925 [1973]).

much in scientific investigation today³, one is led to the conclusion that most scientific ‘truths’ have been established without using randomized experiments. In addition, most of us successfully determine causal effects of many of our everyday actions, even interpersonal behaviors, without the benefit of randomization.

Similar to Rubin, but with respect to external validity, Campbell long argued for the possibility of making causal inferences in the absence of homogeneous effects or a perfectly random sample (e.g., Campbell and Stanley 1963; Shadish Cook and Campbell, pages 83-86, pages 471-472). For example, most recently, Shadish, Cook and Campbell state “Inferences from completed studies to as-yet-unstudied applications are necessary to both science and society” (page 84)⁴.

As discussed by Rubin and Campbell, causal inference does not depend exclusively on full randomization. This is consistent with the classic definition of inference (from the Oxford English Dictionary): “reasoning from something known *or assumed* to something else which follows from it” (emphasis added). Thus assumptions are part of inference; *if* assumptions hold, and given the data, one can make an inference.

Nonetheless, recognizing that inference depends on assumptions, I propose to quantify the robustness of causal inferences in terms of potential violations of assumptions. In particular, I will establish thresholds of robustness in terms of the extent to which the assumptions of causal inference would have to be violated to alter a statistical inference.

I recognize the controversy regarding the use of statistical significance as a criterion for causal inference (e.g., Hunter, 1997; Wilkinson et al 1999). In making a causal inference, one should consider effect sizes, confidence intervals, causal mechanisms, and the nuances of implementation.

³ For example, in Davies (1954), a well-known textbook on experimental design in industrial work, randomization is not emphasized.

⁴In this sense there is agreement with Cronbach (e.g., Cronbach et al. 1980).

But it would also be unusual for an empirical relationship that was *not* statistically significant to be relied upon as a basis of policy change. This is consistent with the recommendation of Wainer and Robinson (2003) to use a “two-step procedure where first the likelihood of an effect (small p value) is established before discussing how impressive it is” (page 25). Therefore statistical significance is treated as an essential condition for causal inference, and thus it is reasonable to define thresholds for the violation of assumptions in terms of statistical inference.

Recognizing that statistical significance is not the only criterion for making a causal inference, I will link the indices to the quantities defining a confidence interval and then I will derive a complementary index based on an effect-size threshold. Therefore the approach here is general enough to apply to multiple thresholds for making causal inferences from quantitative data.

No matter which threshold is used, when robustness indices are employed researchers explicitly recognize that inferences are based on assumptions that are likely violated to some degree. The question then is how extreme the violations would have to be to alter an inference, where alter implies a quantity changes from being greater than a threshold to less than the threshold.

Robustness indices and sensitivity analysis

The approach in this paper can best be considered an extension of sensitivity analysis (cf. Robins, Rotnisky and Scharfstein 2000; Rosenbaum 1986; Scharfstein 2002). With sensitivity analysis one would represent a distribution of possible estimates given a broad set of alternative conditions, helping researchers nuance interpretations of the relationship between treatment and outcome. As in sensitivity analysis, I will consider how unknown quantities could affect estimation. But here I focus on the extent to which the assumptions of inference must be violated to reduce an

estimated quantity below a decision-making threshold. Thus the indices are more a property of the original inference than an exploration of multiple possible estimates under various conditions.

There have been important recent developments in the indexing of the robustness of inferences to violations of the assumptions of inference. As one example, Frank (2000) defined the impact of a confounding variable on a regression estimate and its standard error in terms of the product of two hypothetical correlations. In particular, Frank's impact = $r_{vy} \times r_{vx}$, where r_{vy} is the hypothetical correlation between a confounding variable, v , and the outcome, y , and r_{vx} is the hypothetical correlation between v and the predictor of interest, x . Frank then indexed the robustness of an inference in terms of the impact of an unmeasured confound necessary to alter a statistical inference regarding a regression coefficient.

As a second example, Rosenbaum (2002, chapter 4), developed an index of the robustness of inferences to possible selection bias. Rosenbaum defines Γ in terms of the probabilities ($\pi_{[j,k]}$) that two subjects j and k who have the same values on covariates receive the treatment:

$$\frac{1}{\Gamma} \leq \frac{\pi_{[j]}(1-\pi_{[k]})}{\pi_{[k]}(1-\pi_{[j]})} \leq \Gamma \quad (1)$$

Thus if $\pi_{[j]} = \pi_{[k]}$ the units have equal probability of being assigned to the treatment and $\Gamma=1$. More importantly, the larger the value of Γ (possibly due to some unobserved covariate), the greater the selection bias. Then Rosenbaum expresses the robustness of a statistical inference in terms of the size of Γ necessary to alter the inference.

While drawing on the spirit of these recent efforts the approach in this paper departs in important ways because it is based on the counterfactual framework. First, Frank's (2000) index is defined by the general linear model without consideration for the matching of pairs for possible

outcomes on the same unit. Second, Rosenbaum's index does not pertain directly to the relationship between the selection process and expected values on the possible outcomes, a key feature of the counterfactual.

Perhaps most importantly, both Frank's and Rosenbaum's indices essentially address only concerns about internal validity. Frank focuses on unmeasured confounding variables (with Frank and Min, under review, developing a second index for external validity), whereas Rosenbaum reduces additivity concerns by using non-parametric approaches that can account, for example, for dilation effects of the treatment (see Rosenbaum chapter 5). In contrast, I will develop a single expression of robustness for additivity and independence by exploiting the counterfactual paradigm.

In the next section, I describe the assumptions of inference in terms of the counterfactual. I then develop robustness indices in terms of the proportion of an estimated treatment effect that would have to be attributed to violations of assumptions to alter an inference. In the fourth section I establish a relationship between the robustness indices and the quantities of a confidence interval, and then explore alternatives to statistical inference for defining a robustness index. In the fifth section I apply the indices to causal inferences regarding the effects of Catholic schools and small classes on academic achievement. In the discussion I assess the value of the counterfactual paradigm for generating a robustness index. In the conclusion, I extend a metaphor of bridges that span between statistical analyses and causal inferences.

The Counterfactual and Assumptions of Inference

Holland (1986) describes the overall strong assumption for causal inference that all units are homogeneous. If units are homogeneous, then any observed differences between those units that received the treatment and those that received the control can be attributed to a treatment effect. And if a population of units is homogeneous, then those in any sample represent those of the whole population, whether or not the sample was randomly drawn. But while often employed by laboratory scientists, the homogeneity assumption is less valid for social scientists who work with units that vary from one to another and over time.

Homogeneity can be approximated by alternative assumptions that can be expressed in terms of the counterfactual for non-laboratory settings (the counterfactual framework has been reviewed and presented extensively in the statistical literature – Dawid 2000; Heckman 1997; Rubin 1974; Sobel 1995, 1998; the presentation here follows most closely that of Holland 1986, pages 946-947). To begin, the counterfactual is defined by possible outcomes: y_i^t , the value on the dependent variable that would be observed if unit i were exposed to the treatment; and y_i^c , the value on the dependent variable that would be observed if unit i were exposed to the control. The effect of the treatment on a unit is then specified as $y_i^t - y_i^c$. The fundamental problem of causal inference, as stated by Holland (1986) is that “it is impossible to *observe* the value of y_i^t and y_i^c on the same unit, and therefore it is impossible to *observe* the effect of t [the treatment] on i ” (page 947). Thus the paradigm for making the inference is counterfactual.

Though the counterfactual appears to establish impossible conditions for inference, it can better be thought of as a framework for elucidating the assumptions necessary for inference. First, a key assumption is that those units receiving the treatment were comparable to those receiving the control. Within the counterfactual framework, define $A = t$ if the unit was exposed to the treatment,

and $A = c$ if the unit was exposed to the control. Thus A determines which possible outcome, y_i^T or y_i^C , is observed. The assumption that those units receiving the treatment were similar to those receiving the control is satisfied when $E(y^T) = E(y^T | A=t)$ and $E(y^C) = E(y^C | A=c)$. This is called the independence assumption because it is satisfied when $E(y^T)$ and $E(y^C)$ are independent of which units were assigned to treatment and control. Thus if independence holds then threats to internal validity based on uncontrolled confounding variables are eliminated, and the treatment effect, $E(y^T|A=t) - E(y^C | A=c)$, can be directly estimated based on the observed difference on y between those who received the treatment and those who received the control.

Complementing the assumption of independence, if the treatment has a constant effect across units then a causal inference can be made regarding the effect of the treatment on any unit that is included in, or represented by, the sample. Holland refers to this as the assumption of “additivity” – that the treatment has an additive, not interactive, effect on each unit. When additivity applies, units are homogeneous in their response to the treatment and the effect of the treatment equals $y_i^T - y_i^C$ for all units i . Correspondingly, assuming independence is satisfied, the overall treatment effect can again be estimated by $E(y^T|A=t) - E(y^C | A=c)$.

Technically, a treatment effect can be estimated for an individual and as an average for a population without assuming additivity. But I characterize additivity as a key assumption of causal inference because challenges to external validity are based on claims that the effect inferred for one sub-population would not apply to another – the treatment effect varies across sub-populations. Correspondingly, violation of additivity challenges any explanation in terms of a single causal mechanism (also known as construct validity – see Cronbach 1982 or Shadish et al pages 64-82) and thus challenges corresponding causal inferences.

Of course there are extant responses to concerns regarding violations to the assumptions of causal inference. Concerns about violations to the independence assumption can be addressed by

measuring and statistically controlling for covariates either through the general linear model or propensity scores and matching (Rosenbaum 2002; Shadish et al 2002). Similarly, concerns about violations to the additivity assumption can be addressed by estimating interaction effects of the treatment with covariates or propensity scores designating specific sub-populations. But no matter how many covariates one controls for, in the absence of randomization it is relatively easy for a critic to conjure a hidden source of violation to the assumption of independence. Similarly, no matter how many interactions one estimates, a critic can easily argue that there are still un-estimated variations in treatment effects indicating that the causal mechanism is not well understood.

Deriving Indices of Robustness from the Counterfactual

Rubin has exploited the counterfactual primarily as a basis for making inferences from matched treatment and control cases in the absence of random assignment (e.g., Rosenbaum and Rubin 1983). But by recognizing that most causal inferences can inevitably be challenged, in this paper I use the counterfactual to develop indices of the robustness of inferences to violations to the assumptions of inference.

Assume one intends to analyze quantitative data to determine whether a treatment has an effect on an outcome. One might begin by conducting a two sample t-test to assess whether the difference in the sample means, $\bar{y}_{\text{treatment}} - \bar{y}_{\text{control}}$, is statistically significant in the usual sense of rejecting the null hypothesis that the population means are equal with the assumption that the error terms are *iid* $N(0, \sigma^2)$ and a fixed probability level (e.g., .05).

To develop an index of robustness directly from the counterfactual framework, note that if one could observe the counterfactual data then one would make statistical inferences from a paired t-

test based on the difference between the outcome for the treatment and control for the same unit. Formally, let \bar{y} refer to a sample mean, s refer to a sample standard deviation and n to a sample size. Then let subscripts indicate whether the quantity refers to the observed sample (o), the unobserved (u) counterfactual sample, or a combined (u+o) sample. Superscripts continue to refer to whether the quantity applies to the treatment (t) or control (c). As examples, \bar{y}_o^t refers to the observed sample mean of y for the treatment group, s_{u+o}^c refers to the sample standard deviation for the combined control group, and n_u^t refers to the sample size for the unobserved cases for the treatment group. Using this notation, equation (2) represents the paired t-test for a combined sample including observed cases and the unobserved, hypothetical cases of the counterfactual:

$$t = \frac{\bar{y}_{u+o}^t - \bar{y}_{u+o}^c}{\sqrt{\frac{(s_{u+o}^t)^2}{n_{u+o}^t} + \frac{(s_{u+o}^c)^2}{n_{u+o}^c} - 2r_{y:t \cdot y:c} \frac{s_{u+o}^t}{\sqrt{n_{u+o}^t}} \frac{s_{u+o}^c}{\sqrt{n_{u+o}^c}}}} \quad (2)$$

Thus the difference in means for the combined sample appears in the numerator, and variances for the combined sample appear in the denominator along with combined sample sizes and $r_{y:t \cdot y:c}$ the correlation between pairs of possible outcomes for a common unit.

Because $r_{y:t \cdot y:c}$, as a correlation, can be directly interpreted by social scientists, specification of the match between potential outcomes in terms of $r_{y:t \cdot y:c}$ reflects the value of the counterfactual. When $r_{y:t \cdot y:c} \rightarrow 1$ pairs on a unit are well-matched, and thus the potential violation to the assumption of independence is reduced. Furthermore, as $r_{y:t \cdot y:c} \rightarrow 1$, the treatment has near uniform, and thus additive, effects. On the other hand, (2) reduces to an independent samples t-test when $r_{y:t \cdot y:c} = 0$, which is reasonable when subjects are randomly assigned to treatment conditions and no matching is necessary to address the assumption of independence.

Cause, and estimation of the treatment effect

Define Δ_o as the estimate of the treatment effect from the observed data: $\bar{y}_o^t - \bar{y}_o^c$, and define Δ_u as the estimate of the treatment effect from the unobserved data: $\bar{y}_u^t - \bar{y}_u^c$. Without loss of generality assume throughout the remainder of this paper that $\Delta_o > 0$ and that the true treatment effect is positive, while complementary cases can be developed by symmetry.

To focus on the conditions indicative of cause, namely Δ_o , Δ_u , and $r_{y:t \cdot y:c}$, make the assumption that $(s_u^c)^2$, $(s_o^t)^2$, $(s_u^t)^2$ and $(s_o^c)^2$ estimate a common variance σ^2 . Thus a full expression for the estimated pooled variance is:

$$s_{pooled\ u+o}^2 = \frac{(n_o^t - 1)(s_o^t)^2 + (n_o^c - 1)(s_o^c)^2 + (n_u^t - 1)(s_u^t)^2 + (n_u^c - 1)(s_u^c)^2}{(n_o^t - 1) + (n_o^c - 1) + (n_u^t - 1) + (n_u^c - 1)} \quad (3)$$

$$= \frac{(n_o^t - 1)[(s_o^t)^2 + (s_u^c)^2] + (n_o^c - 1)[(s_o^c)^2 + (s_u^t)^2]}{2(n_o^t + n_o^c - 2)} .$$

Note if the treatment does not alter the variance then $(\sigma_u^c)^2 = (\sigma_o^t)^2$ and $(\sigma_u^t)^2 = (\sigma_o^c)^2$. That is, the variance for observed treatment units does not change when the same units received the control in the unobserved data, and the variance for observed control units does not change when the same units received the treatment in the unobserved data. This condition is violated if, for example, one treatment has a dilating effect, which could be addressed by using non-parametric techniques (Rosenbaum, 2002, chapter 5). If the condition does hold, (3) reduces to a standard pooled estimate of variance from the observed data.

Next, for initial development, assume that there were as many cases receiving the treatment as the control in the originally observed sample: $n_o^t = n_u^t = n_o^c = n_u^c = n/2$. This assumption will be relaxed in the next section, but is employed here to simplify expressions. Equation (2) can now be reduced to a function of n and s as well as Δ_o , Δ_u and $r_{y:t \cdot y:c}$:

$$t = \frac{\Delta_o + \Delta_u}{2s \sqrt{\frac{2(1 - r_{y:t \cdot y:c})}{n}}} \quad (4)$$

Now, the expression in (4) can be used to explore the conditions necessary to alter a statistical inference. In particular, setting $t = t^\#$, the critical value of the distribution for a given probability level (e.g., .05) and degrees of freedom $n-1$, and solving for Δ_u , yields

$$\Delta_u = 2t^\#s \sqrt{\frac{2(1 - r_{y:t \cdot y:c})}{n}} - \Delta_o \quad (5)$$

Though equation (5) can then be used to determine the values of Δ_u that would alter an inference, it does not directly generate a scale free metric for an index. That is, Δ_u is in the scale of the outcome variable, and thus different values of Δ_u cannot be compared across studies, nor we can we discuss the size of Δ_u in an abstract or theoretical sense.

To generate a scale free index compare Δ_u to Δ_o . In particular, define $M = (\Delta_o - \Delta_u) / 2\Delta_o$. Thus M is a measure of the difference between treatment effects in the observed and counterfactual cases. Assume Δ_o is > 0 and statistically significant, but that the inference is challenged with the claim that $\Delta_u < \Delta_o$. Given this claim, $M > 0$ because $(\Delta_o - \Delta_u)$ is positive as is $2\Delta_o$. Furthermore, the scaling factor $2\Delta_o$ represents a maximal difference between Δ_o and Δ_u necessary to alter an inference. If $\Delta_o - \Delta_u = 2\Delta_o$ then $\Delta_u = -\Delta_o$, the combined estimate is 0, and an inference rejecting the null hypothesis

must be altered. Thus we need not consider $\Delta_u < -\Delta_o$. As a result, $(\Delta_o - \Delta_u) < 2\Delta_o$ and $M < 1$. Thus, though the construction of M is similar to measures of relative bias (Krull and MacKinnon 2001), dividing by 2 creates a well defined range (i.e., $0 < M < 1$) for expressing robustness.

M and the Assumptions of Causal Inference

Drawing on the logic of the counterfactual, violation to independence can be addressed by estimating treatment effects while controlling for potentially confounding variables (Rubin 1978; Sobel 1996, 1998). Thus, estimates of the treatment effect are unbiased if the independence assumption is satisfied, *conditional* on covariates. The challenge that an inference is not robust because the independence assumption is violated then implies that the estimate of Δ_o is biased because there are *unmeasured* confounding variables.

To quantify the challenge based on violation of independence, consider the following model for y :

$$y = \mathbf{B}'\mathbf{z} + \mathbf{\Lambda}'\mathbf{w} + \tau d \quad , \quad (6)$$

where \mathbf{z}_i represents a vector of measured covariates and \mathbf{B} represents a vector of coefficients; \mathbf{w} represents a vector of *unmeasured* covariates, and $\mathbf{\Lambda}$ their corresponding coefficients; and d is an indicator taking a value of 1 if the unit received the treatment in the observed data, 0 if it received the control.

As in the example of estimates of the Catholic school effect (and as will be explored in the extensions in the next section), estimates of treatment effects can be directly obtained while controlling for the observed covariates \mathbf{z} . Challenges to inference based on unmeasured violation of

independence then can be attributed to $\Lambda'w$. Define $\mu = \Lambda'E(w_o^t - w_o^c)$. That is, μ is the predicted difference in y between the treatment and control groups in the observed data that can be attributed to unmeasured covariates w . Then the challenge to a causal inference based on unmeasured violation of the independence assumption implies that $\mu > 0$. That is, that those units that received the treatment in the observed sample had higher expected values on y than those that received the control because of differences in unmeasured characteristics. Note that because the observed treatment units are the same as the unobserved control units, $w_u^c = w_o^t$, and, by the same logic, $w_u^t = w_o^c$. Therefore $\Lambda'E(w_u^t - w_u^c) = -\mu$.

Focusing exclusively on the effect of violation of independence, and assuming the estimate of the treatment effect is obtained from a model already controlling for relevant measured covariates, $E(\Delta_o) = \tau + \mu$ and $E(\Delta_u) = \tau - \mu$ and

$$E(M) = \frac{E(\Delta_o) - E(\Delta_u)}{2E(\Delta_o)} = \frac{(\tau + \mu) - (\tau - \mu)}{2(\tau + \mu)} = \frac{\mu}{\tau + \mu} \quad (7)$$

Thus M is proportional to the extent to which the estimated treatment effect can be attributed to violations of independence, as represented by μ .

Charges that the assumption of additivity has been violated imply that the effect of the treatment varies over sub-populations. This can be conceptualized two ways using the counterfactual. First, adhering to the strict interpretation of the counterfactual in terms of paired cases on the same unit, additivity is violated if the effect of the treatment is different for those who received the treatment in the observed data than it would be for those who received the control. Second, following Rosenbaum and Rubin (1983), the counterfactual can also be approximated by

matching cases from different units. Using the matching conceptualization, additivity is violated if the effect of the treatment is different for the population from which the matches are drawn than for the population from which the original sample is drawn. Thus the matching conceptualization represents more directly the concept of external validity as it relates to how the treatment effect varies between two populations.

Of course, charges that additivity has been violated can be addressed by estimating interactions or separate effects for various sub-populations. Even after doing so, additivity may still be violated in unknown ways within sub-samples. To quantify the violation of additivity, define τ as the treatment effect in the overall population and assume that $E(\Delta_o)=\tau+\alpha$ and $E(\Delta_u)=\tau-\alpha$. If an inference is invalid because additivity is violated, $\alpha > 0$ indicating that the effect is larger (for some unknown reason) for the population from which the observed cases is drawn than for the overall population. Then focusing on violation of additivity, assuming independence is satisfied (i.e., $\mu=0$):

$$E(M) = \frac{E(\Delta_o) - E(\Delta_u)}{2E(\Delta_o)} = \frac{(\tau+\alpha) - (\tau-\alpha)}{2(\tau+\alpha)} = \frac{\alpha}{\tau+\alpha} . \quad (8)$$

Thus M is proportional to the extent to which the estimated treatment effect can be attributed to violations of additivity as expressed through α .

Finally, combining the results from (6) and (8) yields:

$$E(M) = \frac{E(\Delta_o) - E(\Delta_u)}{2E(\Delta_o)} = \frac{(\tau+\alpha+\mu) - (\tau-\alpha-\mu)}{2(\tau+\alpha+\mu)} = \frac{\alpha+\mu}{\tau+\alpha+\mu} . \quad (9)$$

In equation (9), if violations of the assumptions can only *reduce* the treatment effect (as is charged by critics of causal inferences), α and μ are both positive. Given the assumption that the treatment effect, τ , is also positive, equation (9) implies $0 \leq M \leq 1$. Thus M can be intuitively interpreted as the proportion of the estimated effect that can be attributed to violations of independence *or* additivity.

M also can be interpreted as a violation of the stable unit treatment value added (SUTVA) assumption. SUTVA implies that the observed outcome is *stable* in the sense that the unit would have the same outcome for all other allocations such that unit i receives treatment t (Rubin 1990 page 282). This is a critical assumption in moving from causal inference for an individual to causal inference for a group (Rubin 1986, 1990). SUTVA can be violated “if there exists interference between units” (Rubin 1990, page 282), such that one unit’s value on a possible outcome depends on the assignment of another unit. In our examples, SUTVA is violated if the effect of one student’s assignment to a Catholic school or small class depends another student’s assignment. This could occur if there are certain students who benefit from studying together because each student’s performance would be better if the other were assigned to the same school or classroom.

Define Δ'_o as the difference in means for the originally observed cases *after* the unobserved cases have been added to the sample. In originally defining M , it was assumed that $E(\Delta'_o) = E(\Delta_o)$ but that the overall statistical inference changed because $E(\Delta_u) \neq E(\Delta_o)$. Instead, assume $E(\Delta_u) = E(\Delta_o)$, that is that the expected treatment effect for the unobserved data equals the expected treatment effect for the originally observed data, but that $E(\Delta'_o) \neq E(\Delta_o)$, that is that the expected difference in means in the observed cases changes as a result of assigning the unobserved counterfactuals to treatments. Then define $M' = (\Delta_o - \Delta'_o) / (2 \Delta_o)$ and note that the calculations for M' are identical to those for M , when substituting Δ'_o for Δ_u . Thus M also is sensitive to violations of SUTVA.

Indices Based on M

Substituting M in (5) and solving for M yields

$$M = 1 - \frac{t^{\#} s \sqrt{\frac{2}{n}(1 - r_{y:t \cdot y:c})}}{\Delta_o} . \quad (10)$$

Thus if M is larger than the right side of (10) then Δ_{u+o} , the estimated treatment effect in the combined sample including counterfactual cases, is not statistically significant. The right hand side of (10) then defines the threshold for violation of the assumptions of inference or $\text{TVAI}(\Delta_o)$; the larger the $\text{TVAI}(\Delta_o)$ the more robust the statistical to violations of the assumptions of causal inference.

Because $r_{y:t \cdot y:c}$ appears on the right hand side of (10), the expression does not produce a single valued index. This is consistent with Dawid's (2000) critique of the counterfactual model as not identifiable; causal inference from the counterfactual depends on Δ_u and $r_{y:t \cdot y:c}$ both of which are unobserved. Dawid (2000) describes the assumption $r_{y:t \cdot y:c} = 0$ as a convention that yields interpretable causal inferences of the counterfactual from a frequentist perspective. But this assumption reduces the paired test to an independent samples test, and thus loses the power of the counterfactual in pairing multiple possible outcomes for a given unit. Furthermore, Dawid argues that specifying $r_{y:t \cdot y:c} = 0$ renders the treatment effect so variable as to violate the assumption of additivity. On the other hand, as $r_{y:t \cdot y:c} \rightarrow 1$ the treatment has near uniform effects and the model is, in Dawid's terms, deterministic. Thus it appears that there is no value of $r_{y:t \cdot y:c}$ that is completely defensible.

Dawid's critique of the counterfactual model as unidentified is in fact a limitation of causal inference that is merely manifest in Δ_u and $r_{y:t \cdot y:c}$. The limitation could as easily be applied to

concerns raised regarding internal as well as external validity (Shadish et al. 2002), concerns regarding confounding variables as well as interaction effects (Frank 2000; Frank and Min, under review), the independence assumption as well as the additivity assumption (Holland 1986), etc.. Ultimately there are multiple assumptions necessary for any causal inference, and the critique of the counterfactual as unidentifiable is not unique.

Recognizing that many assumptions are needed for causal inference, one could represent sensitivity in terms of combinations of parameters. Indeed I will do this for the examples of Catholic schools and class sizes. But this falls short of the goal of generating single valued indices that can be easily interpreted for a given application and can be compared across applications.

Current use of the counterfactual provides guides my choice of focus on M or $r_{y:t \cdot y:c}$. In particular, Rosenbaum and Rubin (1983) exploit the counterfactual to establish a paired data set, with those receiving the treatment matched to those receiving the control, where matches are typically based on the propensity for receiving the treatment. Thus the pairing of the counterfactual is incorporated into the design. Analysis can then focus on the components of M , which are indicative of treatment effects. Following Rosenbaum and Rubin (1983) I define indices for specific theoretically motivated values of $r_{y:t \cdot y:c}$ or for distributions of $r_{y:t \cdot y:c}$, allowing me to focus on M as indicative of a treatment effect.

Begin by drawing on our understanding of $r_{y:t \cdot y:c}$ as a correlation between possible outcomes to select theoretically motivated values of $r_{y:t \cdot y:c}$. In particular, setting $r_{y:t \cdot y:c} = .5$ characterizes moderately well matched pairs, pragmatically splitting the difference between Dawid's extremes of nonadditivity and determinism. Setting $r_{y:t \cdot y:c} = .5$ defines the TVAI* (Δ_o)⁵:

⁵Alternatively, the TVAI* (Δ_o) can be obtained as an average of TVAI (Δ_o) for $r_{y:t \cdot y:c} = -1$ and $r_{y:t \cdot y:c} = 1$.

$$TVAI^*(\Delta_o) = 1 - \frac{t^{\#}s}{\Delta_o\sqrt{n}} . \quad (11)$$

Interestingly, note that the $TVAI^*(\Delta_o)$ is a function of the quantities defining a confidence interval for Δ_o . I will comment more on this in the next section.

For those uncomfortable with assigning $r_{y:t \cdot y:c}$ a single theoretical value, consider $r_{y:t \cdot y:c}$ a nuisance parameter and integrate the right hand side of (10) over $r_{y:t \cdot y:c}$ from -1 to 1 (assuming $r_{y:t \cdot y:c}$ is uniformly distributed across the interval):

$$\int_{-1}^1 \left(1 - \frac{t^{\#}s}{(\Delta_o)\sqrt{n}} \sqrt{\frac{2}{n}(1-r_{y:t \cdot y:c})} \right) dr_{y:t \cdot y:c} = 2 \left(1 - \frac{4}{3} \frac{t^{\#}s}{(\Delta_o)\sqrt{n}} \right) . \quad (12)$$

Because the width of the interval is two, the mean value, $TVAI^{**}(\Delta_o)$, is then

$$TVAI^{**}(\Delta_o) = 1 - \frac{4}{3} \frac{t^{\#}s}{\Delta_o\sqrt{n}} . \quad (13)$$

By the mean value theorem $TVAI^{**}(\Delta_o)$ occurs when $r_{y:t \cdot y:c} = 1/9$.

More generally, consider restricting $0 < r_{y:t \cdot y:c} < 1$, the region in which the benefits of the counterfactual are realized in the successful matching of possible outcomes. We can then use a Beta distribution to express prior beliefs about $r_{y:t \cdot y:c}$. Define the standard Beta distribution (over the interval 0,1) as

$$f(r_{y:t \cdot y:c}) = \frac{(r_{y:t \cdot y:c})^{a-1}(1-r_{y:t \cdot y:c})^{b-1}}{B(a,b)} \quad (14)$$

where

$$B(a,b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt \quad (15)$$

Then different thresholds can be obtained by integrating

$$\int_0^1 \left(1 - \frac{t^{\#s}}{(\Delta_o)} \sqrt{\frac{2}{n} (1 - r_{y:t \cdot y:c})} \right) f(r_{y:t \cdot y:c}) dr_{y:t \cdot y:c} \quad (16)$$

As $a=b$ and both approach ∞ , the Beta distribution converges on a the single point $r_{y:t \cdot y:c} = .5$, and (10) reduces to (11) yielding the TVAI* (Δ_o). On the other hand, when $a=b=0$ the Beta distribution is uniform yielding

$$TVAI^{***}(\Delta_o) \approx 1 - \frac{.94t^{\#s}}{\Delta_o \sqrt{n}} \quad (17)$$

occurring when $r_{y:t \cdot y:c} = 5/9$ (which is very close to the theoretically motivated $r_{y:t \cdot y:c} = .5$ for the TVAI* [Δ_o]). Thus when $a=b$ and when $a > 0$ and $b > 0$, $1/2 \leq r_{y:t \cdot y:c} \leq 5/9$. Thus the mean value of $r_{y:t \cdot y:c}$ has a very restricted range when prior beliefs regarding $r_{y:t \cdot y:c}$ are symmetric.

Of course, prior beliefs may be that $r_{y:t \cdot y:c}$ is not symmetric. That is, researchers may believe that the mass of $r_{y:t \cdot y:c}$ is concentrated near 0 (with $b > a$) or near 1 (with $a > b$)⁶. Because the definite integrals vary only in the coefficient, κ , by which $t^{\#s}/[\Delta_o(n)^{1/2}]$ is multiplied, the mean value of $r_{y:t \cdot y:c}$ can be calculated as $r_{y:t \cdot y:c} = 1 - \kappa^2$.

⁶Researchers can generate the shape of the pdf for a Beta distribution for any a or b at <http://www.mechmat.univ.kiev.ua/probability/Projects/StatGraph/BetaGraph.html>

Using the approach as described in the preceding paragraph and as in (14) through (17), I calculated the mean value of $r_{y:t \cdot y:c}$ for all integer combinations of a and b for $1 \leq a \leq 10$ and $1 \leq b \leq 10$. The result is that the mean value of $r_{y:t \cdot y:c}$ is almost completely determined ($R^2 = .98$) by the following formula:

$$\hat{r}_{y:t \cdot y:c} = .52687 + .04806a - .05107b - .00518(a-5.5)^2 + .00553(b-5.5)^2 \quad . \quad (18)$$

Note that $r_{y:t \cdot y:c}$ increases with a , but with negative curvature, and $r_{y:t \cdot y:c}$ decreases with b , with positive curvature. Note also that the linear and quadratic effects of a are similar in magnitude to those of b , though not identical.

Using (18) researchers need not evaluate integrals as in (16). Instead researchers can specify the parameters of a Beta distribution that represent their beliefs about $r_{y:t \cdot y:c}$ and then calculate the mean value of $r_{y:t \cdot y:c}$ integrating over the whole distribution using (18). From this the TVAI can be calculated from (10). As examples, when $a=3$ and $b=10$ (with the mode close to zero and positive skew), $\hat{r}_{y:t \cdot y:c} = .233$ and $\hat{\kappa} = .876$, while when $a=10$ and $b=3$ (with the mode close to one and negative skew), $\hat{r}_{y:t \cdot y:c} = .792$ and $\hat{\kappa} = .456$.

Extensions

Including Covariates

Any of the above indices can easily be generalized to analyses that control for covariates, \mathbf{z} , that may be used to establish conditional independence (Rubin 1974; Sobel 1996, 1998). For example, when using multiple regression to control for covariates, the regression coefficient for an indicator of the treatment (i.e., 1 for the treatment, 0 for the control), represents the difference in

means of the treatments, conditional on the covariates. The residual variance of y also is conditioned on the covariates and the sampling distribution for a partial correlation is the same as that for a correlation, save the change in degrees of freedom which can be reflected by using $n-q-1$ instead of n , where q is the number of parameters estimated in the model (see Cohen and Cohen, 1983, pages 103-107). Thus these conditional values can be used instead of the unconditional values beginning with equation (2).

Unequal Sample Sizes

The assumption $n_o^t = n_o^c = n_u^t = n_u^c = n/2$ was made so that statistical inferences would not be influenced by unequal numbers of observed treatment and control cases. But this may not apply in many sampling schemes. If the assumption did not hold, differences in sample means, beginning with equation (4), can be replaced by weighted means. For example

$$\frac{(n_o^t \bar{y}_c^t - n_o^c \bar{y}_o^c)}{n} \text{ replaces } \Delta_o \quad . \quad (19)$$

Correspondingly, Δ_{u+o} is a weighted average of the two components of the counterfactual (Winship and Morgan 1999), the estimated treatment effect for those who received the treatment and the estimated treatment effect for those who received the control:

$$\Delta_{u+o} = \frac{n_o^t (\bar{y}_o^t - \bar{y}_u^c) + n_o^c (\bar{y}_u^t - \bar{y}_o^c)}{n_o^t + n_o^c} \quad . \quad (20)$$

Using (20) one may generate two indices, one for $\bar{y}_o^t - \bar{y}_u^c$ using $n=n_o^t$ and one for $\bar{y}_u^t - \bar{y}_o^c$ using $n=n_o^c$. Conservatively, an overall index can be obtained using the minimum of n_o^t and n_o^c , while the indices developed in the previous section are based on the mean of n_o^t and n_o^c .

Relationship to a Confidence Interval and Alternate Thresholds

As previously noted, below equation (11), there is a functional relationship between TVAI* (Δ_o) and the lower bound of a confidence interval. Specifically, assuming covariates in the model,

$$TVAI^*(\Delta_o) = \left(s_x \sqrt{1 - r_{xz}^2} \left(\frac{\text{lower bound of the confidence interval for } \Delta_o}{\Delta_o} - 1 \right) + 1 \right) \quad (21)$$

If $s_x = 1$ and there are no covariates in the model, then the right side of (21) reduces to:

$$\frac{\text{lower bound of the confidence interval for } \Delta_o}{\Delta_o} \quad (22)$$

Like the TVAI, the ratio on the right of (22), referred to as $LB(\Delta_o)/(\Delta_o)$, is bounded by 0 and 1. First, the lower bound of a confidence interval is always less than (Δ_o) and thus the ratio on the right of (22) is less than 1. Furthermore, when the null hypothesis is rejected, the lower bound of the confidence interval is greater than zero (assuming $\Delta_o > 0$) and thus the ratio on the right side of (22) is > 0 .

Beyond having a fixed range and being a special case of the TVAI*, the $LB(\Delta_o)/(\Delta_o)$ offers an intuitive index of robustness in its own right. The $LB(\Delta_o)/(\Delta_o)$ favors the larger effect for two intervals of the same width and also favors the narrower width for two effects of the same size. Thus

the ratio intuitively combines the key quantities of the point estimate and the width of a confidence interval into a single index of robustness.

Expressing a robustness index in terms of the ratio in (22) also facilitates consideration of criterion for decision-making other than statistical significance from a classical frequentist interpretation. First, one could define the ratio by a Bayesian confidence region instead of the frequentist interval. Second, one could move beyond statistical significance by replacing the lower bound of the interval with an alternative criterion. For example, if one preferred to make a causal inference based on effect size, then one could replace the quantities used for statistical inference with those used to define an effect size. For example, define the effect size robustness (ESR) for an effect size of .2 as:

$$ESR(\Delta_o,.2)=1-\frac{.2s}{\Delta_o} . \quad (23)$$

Like the other indices, the larger the value of $ESR(\Delta_o,.2)$, the more Δ_o exceeds the threshold, the more robust is the inference. Also like the other indices, $ESR(\Delta_o,.2)$ will be bounded by 0 and 1 as long as the effect (e.g., Δ_o) surpasses the baseline positive criterion (e.g., .2s).

Expressing robustness generally in terms of a ratio of a cut-off criterion to Δ_o expands the framework for interpreting the indices. Nonetheless, each index is defined relative to a dichotomous decision, while the indices reflect the uncertainty of decision-making. That is, representing the robustness of an inference recognizes that although a threshold was exceeded, the decision could be altered by unknown quantities. In the next section I calculate the indices for inferences regarding the effects of Catholic schools and class size on academic achievement. These indices contextualize each causal inference and provide a basis for comparing the robustness of the causal inferences.

Application of Indices to Inferences for Catholic Schools and Small Classes

In this section I will apply the indices of robustness to causal inferences regarding the effects of Catholic schools and class size. As noted in the introduction, critics of the statistical and causal inferences regarding Catholic schools often suggest that the observed relationship is spurious. In other words, the assumption of independence has not been satisfied, or that the results are biased due to selective assignment to school sector⁷. This is a typical concern raised for an observational study.

The concern regarding possible violation of independence can be most directly addressed by drawing on the longitudinal design of High School and Beyond (HS&B) and the subsequent National Educational Longitudinal Study beginning in 1988 (NELS:88). For example, of the analyses of the Catholic school effect using the more recent NELS:88 data (Altonji, Elder and and Taber 2000; Figlio and Stone 1999; Gamoran 1996; Goldhaber 1996; Grogger and Neal 2001), arguably the most current and methodologically comprehensive can be found in Morgan (2001). Morgan begins by reproducing Coleman's analyses, defining the outcome as a difference score between twelfth and tenth grades and then controlling for the key covariate, tenth grade achievement (Morgan also explores models that include other covariates, propensity scores and possible mediating constructs such as educational expectations and parental involvement, course-taking, school climate and track).

Essentially confirming Coleman's results, Morgan found that the overall effect of Catholic schools on math achievement was statistically significant, controlling for tenth grade math achievement and family background and demographics (a set of fourteen variables). As reported in Table 1, Morgan's estimated difference between Catholic and public schools was .99 on a

⁷Coleman et al. were also critiqued for failing to account for the clustering of students within schools which now is typically accounted for with multilevel or random effects models (e.g., Lee and Bryk 1989) and which were employed in the analyses presented below.

standardized math test, with standard error of .33⁸. The corresponding effect size (based on the difference in means relative to the residual standard deviation in Achievement) is .23, and the 95% confidence interval is (.34, 1.64).

Insert Table 1 about here

Though Coleman et al. and Morgan controlled for pre-tests and analyzed nationally representative data, there may still be concerns regarding the assumptions of causal inference. For example, the covariates did not include parental educational expectations and intended engagement of parents, both of which might be higher for those students who attended Catholic schools, and whose effects on achievement emerge after tenth grade. Thus the independence assumption could be violated. Furthermore, the effects of Catholic schools may be stronger for those of lower socioeconomic status or lower propensity to attend Catholic schools, violating the additivity assumption.

How much of the observed effect of Catholic schools would have to be attributed to violations of assumptions such that the inference would be altered in an analysis including the unobserved, counterfactual data? Assuming moderate matching of the counterfactual and observed data (i.e., $r_{y:Catholic \cdot y:public} = .5$), more than 73% of the observed effect would have to be attributed to violations of assumptions such that the inference for the combined data would be different from that for the observed data. Correspondingly, $\Delta_u^{Catholic}$ would have to be less than -.45 to alter the inference, as

⁸Morgan's standard errors reflect dependencies associated with the nesting of students within schools – see note to Morgan's Table 1, page 345 – and thus the sample size is based on the number of schools, approximately 973. Using the multilevel modeling software of Bryk, Raudenbush and Congdon (2002), the regression coefficient was 1.0 with standard error .34 and t-ratio of 2.96, confirming Morgan's results.

shown in (24):

$$\begin{aligned}\Delta_o &= \hat{\tau} + (\hat{\alpha} + \hat{\rho}) = (1 - TVAI^*)\Delta_o + (TVAI^*)\Delta_o = .27 + .72 = .99 \text{ and} \\ \Delta_u &= \hat{\tau} - (\hat{\alpha} + \hat{\rho}) = (1 - TVAI^*)\Delta_o - (TVAI^*)\Delta_o = .27 - .72 = -.45 \quad .\end{aligned}\tag{24}$$

Assuming only weak matching (i.e., $r_{y:Catholic \cdot y:public} = 1/9$), more than 64% of the estimated effect, with $\Delta_u^{Catholic} < -.27$, would have to be attributable to violations of assumptions to alter the inference.

The threshold for the TVAI indices is defined by the conditions necessary to alter a statistical inference. But in the extensions I generated indices based on alternative thresholds. For the Catholic schools example, if one were willing to make an inference only if the estimated effect were less than an effect size of .2, then $\Delta_u^{Catholic}$ would have to be less than -.72, with ESR $(\Delta_o, .2s) = .14$, to alter the inference in the combined data. Note that because the effect size of .2 would still be statistically significant, the ESR $(\Delta_o, .2s)$ is a higher threshold than the TVAI, and thus less robust with respect to violations of assumptions. Finally the $LB(\Delta_o)/\Delta_o$, based on the intuitive comparison of the lower bound of the confidence interval to the estimated effect, is .34.

Though the robustness indices can be used to inform scientific and policy debates regarding the effects of Catholic schools that were inferred from observational studies, recently the federal government (U.S. Department of Education 2002), some state governments (e.g., Tennessee) and private foundations (e.g., Mathematica) have emphasized the value of randomized experiments to reduce concerns regarding violations of the independence assumption. The Tennessee class size

study affords an important example of the use of randomized experiments to determine causal inferences for educational policy (Finn and Achilles 1990)⁹.

In the Tennessee class size study, classrooms were randomly assigned to have small teacher-pupil ratios (1:13-17, 122 classes), regular teacher-pupil ratios (1: 22-25, 111 classes) or regular teacher-pupil ratios with an aide (98 classes). As reported in the bottom row of Table 1, the mean difference in achievement on the Stanford Achievement Test for reading for small classes versus all others was 13.14 with a standard error of 2.34¹⁰. The corresponding effect size is .68 and the 95% confidence interval is (8.55, 17.73).

As noted in the introduction, some have challenged the inferences regarding class size, arguing that because the effects vary by characteristics of the students and classrooms, the study lacks external validity (e.g., Hanushek 1999). But how much of the estimated effect of small classes would have to be attributed to violations of assumptions such that the inference would be altered in an analysis including the unobserved, counterfactual data? Assuming moderate matching of the counterfactual and observed data (i.e., $r_{y:small\ class \cdot y:other\ class} = .5$), more than 84% of the observed effect would have to be attributed to violations of assumptions to alter the inference. Correspondingly, $\Delta_u^{Small\ class}$ would have to be less than -8.97 to alter the inference. Assuming only weak matching (i.e., $r_{y:small$

⁹ In research related to policies similar to those implied by Coleman et al., Paul Peterson and colleagues have recently assessed the effects of vouchers using randomized experiments (e.g., Greene 2000; Howell and Peterson 2000; Mayer, et al. 2002; Wolf 2000; see McEwan 2000, for a review). But the average results are often not statistically significant across the whole of the sample (e.g., Mayer et al.), making the robustness indices undefined.

¹⁰ These results were obtained from Finn and Achilles, table 5, where the mean for other classes is based on the regular and aide classes combine proportional to their sample sizes. Effect size was taken at the classroom level to address concerns regarding the nesting of students within schools. The pooled standard deviation was calculated as $mean_difference/effect_size$.

class ·y:other class = 1/9), still more than 79% of the estimated effect, with $\Delta_u^{Small\ class} < -7.57$, would have to be attributable to violations of assumptions to alter the inference.

Figure 1 provides a graphical representation of the paired observations as in the counterfactual (n=50 simulated cases). The observed treatment effect was constructed by creating treatment and paired controls such that the controls were selected from a normal deviate with standard deviation=19.36 and the treatment was constructed using a second random normal deviate such that $\Delta_o^{Small\ class} = 13.14$ and $s = 19.36$ and $r_{y:t \cdot y:c} = .5$ (the values of 13.14 and 19.36 were as estimated by Finn and Achilles). Similarly, the unobserved cases were constructed such that $\Delta_u^{Small\ class} = -8.97$, as calculated from the TVAI, with $s = 19.36$ and $r_{y:t \cdot y:c} = .5$.

The vertical dimension of figure 1 corresponds to values of achievement for small classes and the horizontal dimension corresponds to the paired values for regular classes. The squares represent the observed treatment effects and the triangles represent the unobserved treatment effects. The forty-five degree line occurs at $y_{small\ classes\ i} = y_{other\ classes\ i}$. Points falling below the line indicate negative treatment effects and points falling above the line indicate positive treatment effects.

The data in figure 1 provide a graphical interpretation to the TVAI. Note that all but four of the triangles appear above the forty-five degree line, indicating the general positive observed effect of the treatment. If violations to assumptions altered the inference regarding the effect of small classes, then the unobserved treatment effect would have to be as represented by the squares, all but seven of which are below the forty-five degree line. Thus for the combined data, class size would not be statistically significant. Most importantly, figure 1 shows the strong separation between observed treatment effects and unobserved treatment effects that would have to occur to alter the inference. Recall that little of this separation can be attributed to differences between subjects receiving treatment or control in the original data, because Finn and Achilles implemented a randomized

experiment. Thus the separation must primarily be attributed to violation of the additivity assumption – small classes would have to have a different and strikingly negative effect in the unobserved counterfactuals to alter the overall inference.

Insert Figure 1 about here

Using the alternative criteria, if one were willing to make an inference only if the estimated effect were less than an effect size of .2, then $\Delta_u^{Small\ class}$ would have to be less than -11.73, with ESR $(\Delta_o, .2s) = .71$, to alter the inference in the combined data. Finally the $LB(\Delta_o)/\Delta_o$ is .65. Thus, regardless of the index, the inference for small classes is more robust than that for Catholic schools.

Relationship between TVAI and $r_{y:t \cdot y:c}$ for the effects of Catholic schools and small classes

The functional relationship between the TVAI and $r_{y:t \cdot y:c}$ for the causal inferences for Catholic schools and small classes is shown in figure 2. For the specific findings of each study, the lines indicate how large the TVAI must be to alter the inference given a value of $r_{y:t \cdot y:c}$. Generally, the TVAI increases monotonically with $r_{y:t \cdot y:c}$; the greater the correlation between pairs of observations, the greater must be proportion of the estimated treatment effect that can be attributed to violations of assumptions to alter the statistical inference. Furthermore, the TVAI $\rightarrow 1$ as $r_{y:t \cdot y:c} \rightarrow 1$ because the proportion of the estimated treatment effect attributable violations of assumptions must approach 100% to alter the inferences in perfectly matched data (i.e., $r_{y:t \cdot y:c} = 1$). Finally, note that because integration over each distribution of $r_{y:t \cdot y:c}$ as in (16) maps to a single value of $r_{y:t \cdot y:c}$ by the mean value theorem, all possible beta distributions of $r_{y:t \cdot y:c}$ are represented by the curve in figure 2.

Insert Figure 2 about here

The region above each line is where the TVAI is great enough to alter inferences, given $r_{y:t \cdot y:c}$. Thus the fact that the dashed line for small classes is above the dotted line for Catholic schools indicates that, generally, the causal inference for small classes is more robust with respect to violations of assumptions than the causal inference for Catholic schools (this also holds true for the indices based on confidence intervals and effect sizes). This is in spite of the fact that the sample size for Catholic schools is larger indicating that robustness is not strictly a function of sample size. The vertical lines at $r_{y:t \cdot y:c} = .11$ (1/9) and $r_{y:t \cdot y:c} = .5$ indicate the TVAI** and TVAI* respectively. Note that the TVAI** for Catholic schools is at .64 and for small classes is .79 while the TVAI* for Catholic schools is .73 and for small classes is .84. Thus the difference in robustness between the two causal inferences is greater for smaller values of $r_{y:t \cdot y:c}$, making TVAI** the more discriminating index.

Discussion

Strictly speaking, we can only be certain of a cause if we were to observe the counterfactual cases for individuals. But if units have something in common, if they are exchangeable, then social scientists may look for underlying indications of cause in a statistical analysis of the aggregate. Yet, it is precisely when analyzing the aggregate that we expose ourselves to the possibility that differences among individuals could spuriously produce treatment effects (violation of independence), or that the treatment effect varies in unknown ways across a population (violation of additivity).

Because social scientists are rarely certain that the assumptions of inference have been satisfied, I have quantified how large the violations to assumptions would have to be to alter an

inference. Is the observed relationship between treatment and outcome strong enough to be indicative of an underlying principle even though the relationship varies across units, could partially be attributed to other factors, or could be altered if others were assigned to the treatment? The answers are intuitively phrased in terms of the proportion of an estimated treatment effect that must be attributable to violations of assumptions to alter the inference.

In this paper I have exploited the matching feature of the counterfactual to consider theoretical values or distributions of $r_{y:t \cdot y:c}$. I could then develop the indices in terms of the comparison of Δ_u and Δ_o . Drawing on the counterfactual, I showed how difference between Δ_u and Δ_o could be induced by differences in characteristics of those who received the treatment and control, thus violating the independence assumption and compromising internal validity. Or the difference between Δ_u and Δ_o could be induced by variable treatment effects, violating the additivity assumption and thus compromising external validity. Thus the counterfactual not only provides a theoretical framework for considering $r_{y:t \cdot y:c}$, but also captures both key assumptions of inference in the comparison of Δ_u to Δ_o .

Ultimately, robustness indices have great potential for application compared with other procedures. Robustness indices can be calculated whenever a statistical inference is made from a general linear model, without need for an instrumental variable (as in the case of selection models), large pools of subjects who received treatment or control (as are needed for matching or to estimate interactions), or extensive covariates (as are used to generate propensity scores or as applied in multiple imputation). Furthermore, while the indices are similar to bounds placed on coefficients (e.g., Manski and Nagin 1998), the indices can be defined directly in terms of the conditions needed to alter a statistical and thus causal inference. Finally, the particular robustness indices developed here directly represent the uncertainty about how violation of assumptions would affect an inference,

in contrast to general sensitivity analyses that consider a range of possible estimates (and resulting inferences) or other recently developed robustness indices that focus on the independence assumption.

Beyond the strict numerical comparison, the robustness of the inferences for Catholic schools and small classes should be interpreted relative to the strengths of the study designs. In particular, the assumption of independence is more likely to be violated than the assumption of additivity when making inferences regarding Catholic schools from nationally representative observational studies. As a complement, the assumption of additivity is more likely to be violated than the assumption of independence when making inferences for the effects of small classes from a randomized experiment on a volunteer sample.

Because each design dramatically reduces the extent to which one assumption is violated, inferences can be altered primarily through violations of the other assumption. That is, of the more than 73% of the estimated effect that must be attributed to violations of assumptions to alter the inference regarding Catholic schools, most must be due to violation of the independence assumption. Furthermore, the violation to independence is *conditional* on covariates, including a prior measure of mathematics achievement, which were controlled in the model used to estimate the Catholic schools effect. Similarly, of the 84% of the estimated effect that must be attributed to violations of assumptions to alter the inference regarding small classes, most must be attributed to violation of additivity because independence is satisfied to the extent that the randomized experiment was implemented. Thus because the design of each study reduces some possible violations of assumptions, the robustness indices establish a high threshold for violations to the remaining assumptions to alter the inference.

Though the metric of the TVAI is intuitive, it is further informative to compare it to reductions in estimated treatment effects due to existing controls for violations of assumptions. For example, $\Delta_o^{Catholic}$ dropped from 1.31 to .99 once Morgan controlled for family background and demographic characteristics, a drop of .32 or 25%. Using the TVAI*, the proportion of the estimated treatment effect that could be attributed to uncontrolled violations of assumptions (.72) would have to be 2.25 greater than the proportion that was attributed to measured demographics, including socioeconomic status. Given that demographic and background controls are among the most important in Morgan's analysis (Morgan's estimate of the treatment effect is fairly stable until controlling for climate, track, and course-taking which are theoretically the *mediating* mechanisms for the Catholic school effect), then the requirement that any conditional violations of assumptions reduce the treatment effect more than twice as much as do demographic and background controls variables supports the contention that the inference regarding Catholic is moderately robust.

Ultimately, the TVAI pertains to construct validity. True, estimated effects may be partially attributed to alternative explanations, and true, effects may vary across sub-populations. But the question here is, are the biases great enough to alter statistical inference, and, with that, are they great enough to deny that there is some underlying mechanism at work that is strong enough to infer causality in spite of possible bias in estimation? In the first example of this manuscript, how much of the observed relationship between Catholic schools and achievement would have to be attributed to violations of assumptions to void the inference that Catholic schools help students learn more, perhaps because of school climate, assignment of students to the academic track and intensive course taking? In the second example, how much of the relationship between small classes and achievement would have to be attributed to violations of assumptions to void the inference that small classes help students learn more, perhaps because students in small classes complete basic instruction more

quickly, have more time for covering additional basic material, have more in-depth instruction, and generally more access to teachers and resources (Word et al 1990)? The TVAI quantifies the answers.

Conclusion

In spite of the value of the robustness indices it is worth emphasizing the robustness indices do *not* definitively sustain new causal inferences. In fact, the original inferences regarding the effects of Catholic schools and class size on academic achievement were not modified. Nor do the robustness indices replace the need for improved research designs or better theories. If one accepts that causal inferences are to be debated (Abbott 1998), what the robustness indices do is quantify the terms of the debate in terms of the proportion of a treatment effect that could be attributed to violations of assumptions. Therefore instead of “abandoning the use of causal language” (Sobel 1998, page 345, see also Sobel 1996, page 355) we can quantify the robustness of a causal inference and interpret relative to the design of a study.

Metaphorically, assumptions support the bridge between statistical and causal inference (Cornfield and Tukey 1956). And the robustness indices characterize the strength of that bridge according to a counterfactual blueprint. Large values, defined relative to the study design and theoretical understandings of the phenomenon, support a causal inference, and then action. Small values suggest trepidation for even the smallest of decisions. Ultimately, no causal inference is certain, but robustness indices help us choose which bridges to cross.

Table 1
Indices of Robustness for Causal Inferences for
Catholic Schools and Class Size on Academic Achievement

	Standard Inference					Robustness			
	Δ_o (se)	t	n	Confidence Interval	Effect Size	counterfactual TVAI*	confidence interval TVAI**	LB(Δ_o)/ Δ_o	Effect Size ESR(Δ_o ,.2s)
Morgan (replication of Coleman et al.) <i>Catholic versus public schools</i>	.99 ^a (.33)	3.00	973	[.34, 1.64]	.23	.73	.64	.34	.14
Finn and Achilles <i>Small classes versus others</i>	13.14 (2.34)	5.62	331	[8.55,17.73]	.68	.84	.79	.65	.71

^a conditional on covariates.

Δ_o is the observed treatment effect.

Δ_u is the unobserved treatment effect for the counterfactual cases.

LB is the lower bound of a 95% confidence interval.

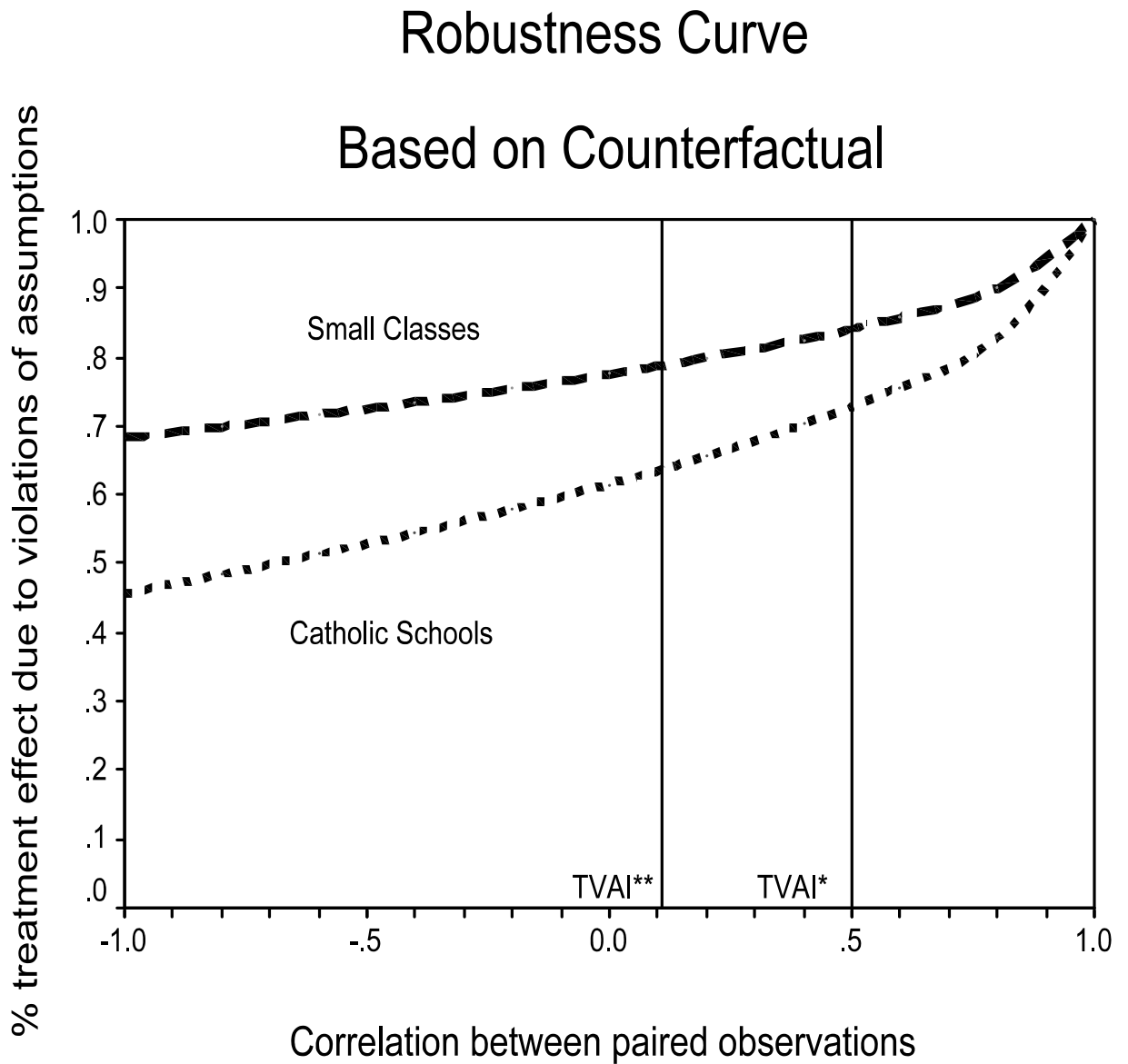
TVAI is the threshold for violation of the assumptions of inference.

TVAI* assumes the correlation between observed cases and unobserved cases paired on the same unit is .5.

TVAI** assumes the correlation between observed cases and unobserved cases paired on the same unit can take any value between -1 and 1.

ESR is effect size robustness.

Figure 1

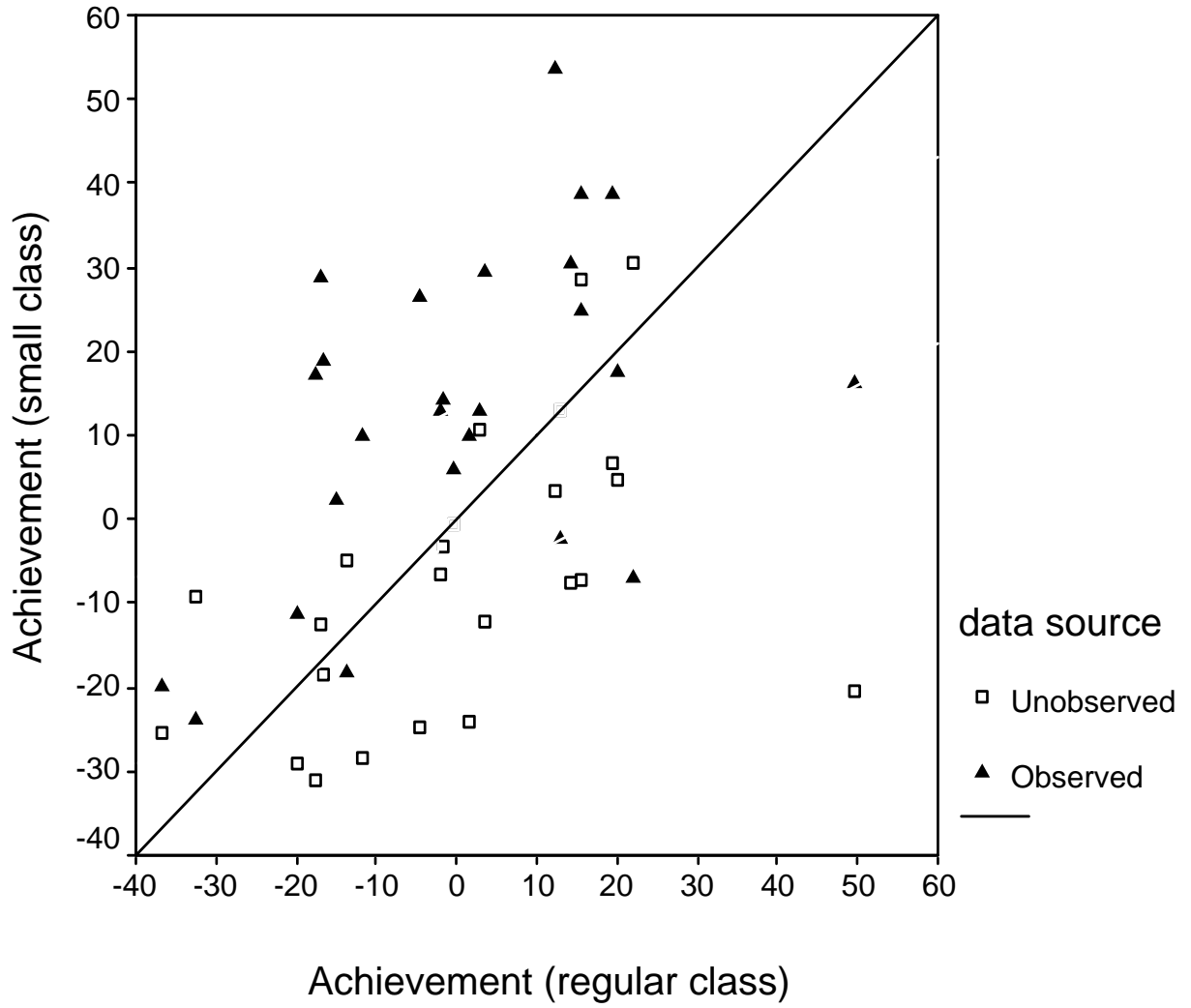


TVAI is the threshold for violation of the assumptions of inference .

TVAI* assumes the correlation between observed cases and unobserved cases paired on the same unit is .5.

TVAI** assumes the correlation between observed cases and unobserved cases paired on the same unit can take any value between -1 and 1.

Figure 2
Paired Observations for Counterfactual: Class Size and Achievement



REFERENCES

- Abbott, A. (1998, November), The Causal Devolution. *Sociological Methods & Research*, 27(2), 148-181.
- Alexander, K. L., & Pallas, A. M. (1985), School sector and cognitive performance: When is a little a little? *Sociology of Education*, 58(2), 115-128.
- Alexander, K. L., & Pallas, A. M. (1983, October), Private Schools and Public Policy: New evidence on cognitive achievement in public and private schools. *Sociology of Education*, 56, 170-182.
- Altonji, J., Elder, T., & Taber, C. (2000), *Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools*, unpublished manuscript., Northwestern University.
- 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, 444-472.
- Angrist, J., & Lavy, V. (1997). *Using Maimonides rule to estimate the effect of class size on scholastic achievement.*, NBER, working paper 5888.
- Bakan, D. (1966), The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.
- Boozer, M. R., Cecilia. (1995). *Intraschool variation in class size: Patterns and Implications.*, NBER, working paper 5144.
- Bryk, Anthony S., Valerie E. Lee, and Peter B. Holland. 1993. *Catholic Schools and the Common Good*. Cambridge, MA: Harvard University Press.
- Bryk, A., Raudenbush, S., and Congdon, R. 2002. Hierarchical Linear Models. Scientific Software Inc. Chicago, IL.
- California Board of Education. 2001. Class Size Reduction. <http://www.cde.ca.gov/csr/>
- Campbell, Donald T. and Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In: J. Hellmuth, Ed. The Disadvantaged Child. Vol 3. Compensatory Education: a national debate. New York: Brunner/Mazel; 1970.
- Carver, R. (1978, August), The case against statistical significance testing. *Harvard Educational Review*, 48(3), 378-399.
- Chubb, J., & Moe, T. (1990), *Politics, markets, and America's schools*. Washington, D.C.: Brookings Institution.
- Cochran, W.G. 1965. "The Planning of Observational Studies of Human Populations." *Journal of the Royal Statistical Society Ser. A*,(Part 2):234-66.
- Cohen, J., & Cohen, P. (1983), *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, N.J.: Erlbaum.
- Cohen, J. (1990, December), Things I have learned (So far), *American Psychologist*, 45(12), 1304-1312.
- Cohen, J. (1994, December), The earth is round (p <.05), *American Psychologist*, 49(12), 997-1003.
- Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, J., et al. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

- Coleman, J. S., Hoffer, T., & Kilgore, S. (1982), *High School Achievement: Public, Catholic, and Private Schools Compared*. New York: Basic Books.
- Coleman, J. S., & Thomas Hoffer. (1987), *The Impact of Communities*. New York: Basic Books.
- Cook, T., & Campbell, D. T. (1979), *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin Company.
- Cornfield, J., & Tukey, J. W. (1956, Dec.), Average Values of Mean Squares in Factorials. *Annals of Mathematical Statistics*, 27(No. 4), 907-949.
- Cronbach, Lee J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.
- Cronbach, L.J. and Snow, R.E. (1977). *Aptitudes and Instructional Methods: A Handbook of Research on Interactions*. New York: Irvington
- Davies, O. L. *Design and Analysis of Industrial Experiments*. London: Oliver and Boyd; 1954.
- Dawid, A. P. (2000, June), Causal inference without counterfactuals. *American Statistical Association*, pp. 407-962.
- Figlio, D., & Stone, J. (1999), School choice and student performance: Are private schools really better? *Research in Labor Economics*.
- Finn, J. D., Charles M. Achilles. (1990, Fall), Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, No. 3, Vol. 27, pp. 557-577.
- Fisher, R. (1973), *Statistical Methods for Research Workers*. New York: Hafner.
- Frank, K. A. (2000), Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods & Research*, 29(2), 147-194.
- Frank, K. A. And Min, K.S. (Under review), "Indices for external validity: Neutralization of statistical significance by unobserved cases" Based on a paper presented at the annual meeting of the American Educational research Association, Frank, K.A. and Duke, N.K. "The Value of Large Scale Data Bases Versus Randomized Experiments for Educational Research."
- Gamoran, A. (1996), Curriculum standardization and equality of opportunity in Scottish secondary education: 1984-90. *Sociology of Education*, 69, 1-21.
- Gigerenzer, G. (1993), The superego, the ego, and the id in statistical reasoning. In G. Keren and C. Lewis (Ed.), *A Handbook for the Data Analysis in the Behavioral Sciences, Methodological Issues* (pp. 311-339), Hillsdale, NJ: Erlbaum.
- Glass, G., Cahen, L., Smith, M., & Filby, N. (1982). *School Class Size: Research and Policy*. Beverly Hills: Sage.
- Goldberger, A. S., & Glen G. Cain. (1982), The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report. *Sociology of Education*, 55, 103-22.
- Goldhaber, D. D. (1996), Public and private high schools: Is school choice an answer to the productivity problem? *Economics of Education Review*, 15(2), 93-109.
- Greene, J. P., Peterson, P. E., & Du, J. (1998), School choice in Milwaukee: A randomized experiment. *Learning from School Choice*, pp. 335-356.

- Greene, J. P. (2000), *A Survey of Results from Voucher Experiments: Where we are and what we know*. The Manhattan Institute.
- Grogger, J., & Neal. (2001), Further evidence on the benefits of Catholic secondary schooling. *Brookings-Wharton Papers on Urban Affairs, 1*.
- Hanushek, Eric A. (1999) "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects", *Educational Evaluation and Policy Analysis, 21*(2), pp. 143-163.
- Heckman, James J. (1997). "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making program evaluations." *Journal of Human Resources, 32*: 441-462.
- Heckman, J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions. Heckman and Singer New York: Cambridge University Press; pp. 156-245.
- Hoffer, T., Andrew M. Greeley. (1985), Achievement Growth in Public and Catholic Schools. *Sociology of Education, 58*, 74-97.
- Holland, P. W. (1986), Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-970.
- Howell, W. G., & Peterson, P. E. (2000), School choice in Dayton, Ohio: An evaluation after one year, Program on Education Policy and Governance. Harvard University.
- Hunter, J. E. (1997, January), Needed: A ban on the significance test. *American Psychological Society, 8*(1), 3-7.
- Krull, J., & MacKinnon, D.P. (2001). "Multivariate modeling of individual and group level mediated effects." *Multivariate Behavioral Research, 36*(2): 249-277.
- Lee, V., & Bryk, A. (1989), A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62*, 172-192.
- Lieberman, A. (1995). *The work of restructuring schools*. New York: Teachers College Press.
- Little, R., & Rubin, D. (2000), Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches. *Annual Review Public Health, 21*, 121-45.
- Manski, C., & Nagin, D. (1998), Bounding Disagreements about Treatment Effects. *Sociological Methodology, 28*, 99-137.
- Mayer, D., Peterson, P., Myers, D., Tuttle, C., & Howell, W. (2002), *School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarship Program*. Mathematic Policy Research Institute.
- McEwan, P. J. (2000), "The potential impact of large-scale voucher programs." *Review of Educational Research, Vol. 70, No. 2*, pp. 103-149.
- Mclaughlin, M., & Talbert, J. (1991). *The contexts of teaching in secondary schools: Teachers' realities*. New York: Oxford University Press.
- Meehl, P. E. (1978), Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834.

- Morgan, S. L. (2001), Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. *Sociology of Education*, 74, 341-374.
- Morgan, W. R. (1983), Learning and Student Life Quality of Public and Private School Youth. *Sociology of Education*, 56, 187-202.
- Morrison, D., & Henkel, R. (1970), *The Significance Test Controversy*. Chicago: Aldine.
- Neyman, J. (1990), On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statistical Science*, 5, 465-72.
- Nickerson, RS (2000). "Null hypothesis significance testing: A review of an old and continuing controversy" *Psychological Methods* (5), 241-301
- Noell, J. (1982), Public and Catholic Schools: A Reanalysis of "Public and Private Schools. *Sociology of Education*, Vol. 55, No. 2/3. (Apr. - Jul., 1982), pp. 123-132.
- Oakes, M. (1986), *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.
- Robins, J., Rotnisky, A., & Scharfstein, D. (2000), Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. E. Halloran and D. Berry (Ed.), (pp. 1-95).
- Rosenbaum, P. R. (1986), Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, 11(3), 207-224.
- Rosenbaum, P.R. and D. B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70(1): 41-55.
- Rozenboom, W. W. (1960), The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57(5), 416-428.
- Rubin, D. B. (1974), Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66, 688-701.
- Rubin, D. B. (1986), Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference." *Journal of American Statistical Association*, 83, 396.
- Rubin, D. B. (1990), Formal Modes of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference*, 25, 279-292.
- Scharfstein, D. a. I., RA. (2002), Generalized Additive Selection Models for the Analysis of Studies with Potentially Non-ignorable Missing Data. *Biometrics*.
- Schmidt, F. L. (1996), Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1(2), 115-129.
- Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, NY: Houghton Mifflin.
- Sobel, M. E. (1995), Causal Inference in the Social and Behavioral Sciences. In Gerhard Arminger, Clifford C. Clogg & Michael E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 1-38), New York: Plenum Press.
- Sobel, M. E. (1996, February), An introduction to causal inference. *Sociological Methods & Research*, 24(3), 353-379.
- Sobel, M. E. (1998, November), Causal inference in statistical models of the process of socioeconomic achievement: A case study. *Sociological Methods & Research*, 27(2), 318-348.

- Sohn, D. (1998), Statistical Significance and Replicability. *Theory & Psychology*, 8(3), 291-311.
- Stolzenberg, R., & Relles, D. (1997). Tools for Intuition about Sample Selection Bias and its Correction. *American Sociological Review*, 62(3), 494-507.
- U.S. Department of Education. (2001). Class Size Reduction Program.
<http://www.ed.gov/offices/OESE/ClassSize/index.html>
- U.S. Department of Education. (2002, November). *Report on scientifically based research supported by u.s. department of education*. Retrieved from
<http://www.excelgov.org/displayContent.asp?Keyword=prppcEvidence>.
- Wainer, H. and Robinson, D.H. (2003). "Shaping up the practice of null hypothesis significance testing." *Educational Researcher*, 32(7), 22-30.
- Wilkinson, L. & T. F. o. S. I. (1999), Statistical Methods in Psychology Journals. *American Psychologist*, 54(8), 594-604.
- Winship, C., & Morgan, S. (1999). The Estimation of Causal Effects from Observational Data. *Annual Review of Sociology*, 25, 659-707.
- Wolf, P. J., Howell. (2000), School choice in Washington, DC: An evaluation after one year, Program on Education Policy and Governance. Harvard University.
- Word, Elizabeth, Achilles, Charles A., Bain, Helen, Folger, John, Johnston, John, Lintz, Nan (1990). Project Star: Final Executive summary report" Tennessee State Department of Education