10

INDICES OF ROBUSTNESS FOR SAMPLE REPRESENTATION

*Kenneth Frank**
Kyung-Seok Min†

Social scientists are rarely able to gather data from the full range of contexts to which they hope to generalize (Shadish, Cook, and Campbell 2002). Here we suggest that debates about the generality of causal inferences in the social sciences can be informed by quantifying the conditions necessary to invalidate an inference. We begin by differentiating the target population into two subpopulations: a potentially observed subpopulation from which all of a sample is drawn and a potentially unobserved subpopulation from which no members of the sample are drawn but which is part of the population to which policymakers seek to generalize. We then quantify the robustness of an inference in terms of the conditions necessary to invalidate an inference if cases from the potentially unobserved subpopulation were included in the sample. We apply the indices to inferences regarding the positive effect of small classes on achievement from the Tennessee class size study and then consider the breadth of external validity. We use the statistical test for whether there is a difference in effects between two subpopulations as a baseline to evaluate robustness, and we

The authors with to acknowledge the helpful comments of two anonymous reviewers, as well as Wei Pan. The opinions expressed in this paper are solely those of the authors. Direct correspondence to Kenneth Frank, Department of Sociology, Michigan State University, 462 Erickson Hall, East Lansing, MI 48824-1034; e-mail: kenfrank@msu.edu.

*Michigan State University

†Korea Institute of Curriculum and Evaluation

consider a Bayesian motivation for the indices and compare the use of the indices with other procedures. In the discussion we emphasize the value of quantifying robustness, consider the value of different quantitative thresholds, and conclude by extending a metaphor linking statistical and causal inferences.

1. INTRODUCTION

1.1. *"But Do Your Results Pertain to. . .?"*

Social scientists are faced with a dilemma because they are rarely able to gather data from the full range of contexts to which they hope to generalize (Shadish, Cook, and Campbell 2002). On the one hand, overly broad generalizations can be misleading when applied to populations that were not well represented by a sample. On the other hand, confining generalization to a target population from which a sample was randomly drawn can limit research results from informing the full range of policies for which they might be relevant. The challenge "But do your results pertain to . . .?" is essential, yet a quandary for social scientists.

Given this problem, the generality of any inference in social sciences is likely to be debated. But current debates are typically qualitative—either a sample represents a target population or it does not. And because generality is rarely certain, debates cast in qualitative terms will often be divisive. Proponents will claim that results generalize, and opponents will claim they do not. Furthermore, while there will rarely be consensus for any given policy, those in the middle must adjudicate in the qualitative terms in which the debate is cast.

Here we suggest that debates about the generality of causal inferences in the social sciences can be informed by quantifying the conditions necessary to invalidate an inference. In this sense we build on recent work in sensitivity analyses (Copas and Li 1997; Frank 2000; Gill and Robins 2001; Robins 1987; Rosenbaum 1987, 2001). But unlike other sensitivity analyses that focus on the robustness of inferences with respect to internal validity, we focus on the robustness of inferences with respect to external validity. Thus, after controlling for all relevant confounding variables (either through a randomized experiment or statistical control), we ask how heterogeneous parameters must be to invalidate inferences regarding effects.

We begin by differentiating the target population into two subpopulations: (1) a potentially observed subpopulation from which all of a sample is drawn, and (2) a potentially unobserved subpopulation from which no members of the sample are drawn (cf. Cronbach 1982) but which is part of the population to which policymakers seek to generalize. We then quantify the robustness of an inference from the observed data in terms of recomposition with the potentially unobserved subpopulation.

1.2. *From Causal Inference to Policy: The Effect of Small Classes on Academic Achievement*

The typical causal inference begins when an estimated effect exceeds some quantitative threshold (e.g., defined by statistical significance or an effect size). For the primary example of this article, consider the results from the Tennessee class size studies, which randomly assigned students to small and large classrooms to evaluate the effectiveness of small classes (Cook 2002; Finn and Achilles 1990; U.S. Department of Education 2002). As reported by Finn and Achilles (1990), the mean difference in achievement on the Stanford Achievement Test for reading for small classes (teacher pupil ratios of 1:13–17, $n = 122$) versus all other classes (teacher-pupil ratios of 1:22–25, some with an aid, $n = 209$) was 13.14 with a standard error of 2.34.[1] This difference is statistically significant. Finn and Achilles then drew on their analysis (including the statistical inference as well as estimates of effect sizes) to make a causal inference: "This research leaves no doubt that small classes have an advantage over larger classes in reading and mathematics in the early primary grades" (p. 573).

If Finn and Achilles' causal inference is correct, it might be reasonable to develop educational policy to reduce class size (e.g., the U.S. Elementary and Secondary Education Act of 2000, which allocated $1.3 billion for class size reduction). Attention then turns to the validity of

[1]These results were obtained from Finn and Achilles (1990, table 5), where the mean for other classes is based on the regular and aide classes combined 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 and corresponding standard error were based on the mean_difference/effect_size.

the causal inference. First, though implementation of the random assignment may not have been perfect (Hanushek 1999) as is often the case (Shadish et al. 2002, chaps. 9 and 10), random assignment of classes to conditions likely reduced most differences between classrooms assigned to be small or not (Cook 2002; Nye, Hedges, and Konstantopoulos 2000). Therefore any overestimate in the effect of small classes is unlikely to be attributed to preexisting differences between the small classrooms and other classrooms (in fact, Nye et al. suggest that deviations from intended treatment may have led to an *underestimate* of the effects of small classes). This is the power of randomization to enhance internal validity (Cook and Campbell 1979).

Attention then turns to the generality of the results beyond the particular sample. Critically, Finn and Achilles (1990) analyzed only a set of volunteer schools, all from Tennessee. Thus, in the most restricted sense, their findings generalize only to schools from Tennessee in the mid-1980s that were likely to volunteer. And yet restricted generalization places extreme limits on the knowledge gained from social science research, especially experiments on the scale of the Tennessee class size study (Shadish et al. 2002:18; Cronbach 1982). Do the results of the Tennessee study mean nothing regarding the likely effects of small classes in other contexts?

The challenge is how to establish external validity by bridging between the sample studied to any given target population. Anticipating challenges to external validity, Finn and Achilles (1990, pp. 559–60) noted that the schools studied were very similar to others in Tennessee in terms of teacher-pupil ratios and percentages of teachers with higher degrees. In the language of Shadish et al. (2002), social scientists can then use this surface similarity as one basis for generalizing from the volunteer sample to the population of schools in Tennessee. But those challenging the generality of the findings could note that the volunteer schools in the study were slightly advantaged in terms of per-pupil expenditures and teacher salaries (Finn and Achilles 1990:559) and Hanushek (1999) adds that the treatment groups were affected by nonrandom and differential attrition (although Nye et al. [2000] argue that this likely had little effect on the estimates). Thus, even for this well-designed study, there is a serious and important debate regarding the generality of the causal inference.

Critically, the debate regarding the generality of the findings beyond the interactions for which Finn and Achilles (1990) tested is either

disconnected from the statistical analyses used to establish the effect or essentially qualitative—the sample is characterized as representative or not. For example, the statistical comparison of schools in the Tennessee class size study with other schools in Tennessee may suggest surface similarity, but it does not quantify how results may be different if a sample more representative of all schools in Tennessee had been used. Similarly, critics suggesting that education in Tennessee is not like that in regions such as California (e.g., Hanushek 1999) use qualitative terms; they do not quantify the differences between their target population and the sample necessary to invalidate the inference that small classes generally improve achievement. Thus, in this article, we develop indices of how robust an inference is by quantifying the sample conditions necessary to make an inference invalid.

In Section 2 we present theoretical motivations for robustness indices; in Section 3 we define an ideal or perfectly representative sample that includes cases from a potentially unobserved population as well as the observed cases; in Section 4 we derive robustness indices for the representation of a sample in terms of the sample recomposition; in Section 5 we apply our indices to the Tennessee class size study; in Section 6 we relate our indices to discussions of the theoretical breadth of external validity; in Section 7 we consider a baseline for our indices in terms of whether there must be a statistical difference between estimates from the observed and unobserved populations to make the original inference invalid; in Section 8 we consider a Bayesian motivation for our indices; in Section 9 we compare with other procedures. In the discussion we emphasize the value of quantifying robustness, use of various quantitative thresholds for inference and consider possible extensions. The conclusion extends a metaphor of a bridge between statistical and causal inference (Cornfield and Tukey 1956).

2. THEORETICAL MOTIVATION FOR ROBUSTNESS INDICES

Our work builds on recent extensions of sensitivity analysis (e.g., Diprete and Gangl 2004; Frank 2000; Gill and Robins 2001; Pan and Frank 2004; Robins 1987; Robins, Rotnisky, and Scharfstein 2000; Rosenbaum 1986, 2002; Scharfstein 2002) to quantify the thresholds at which inferences are invalidated. For example, Rosenbaum (2002) shows that "to attribute the higher rate of death from lung cancer to an unobserved

covariate rather than to the effect of smoking, that unobserved covariate would need to produce a six-fold increase in the odds of smoking, and it would need to be a near perfect predictor of lung cancer" (p. 114).

Similar to Rosenbaum, Frank (2000) indexed the robustness of statistical inferences to the impact of potentially confounding variables that are unobserved. Cast in terms of the general linear model, Frank defined the impact of a confounding variable on an estimated regression coefficient and its standard error in terms of $r_{v \cdot y} \times r_{v \cdot x}$, where $r_{v \cdot y}$ is the correlation between an unmeasured confound, v, and the outcome y and $r_{v \cdot x}$ is the correlation between v and x, a predictor of interest. Maximizing under the constraint: impact $= r_{v \cdot y} \times r_{v \cdot x}$, Frank then developed a single index of how large the impact must be to invalidate a statistical inference.

In general, like the indices of Rosenbaum (2002), Robins (1989), and Frank (2000), the indices we will develop extend sensitivity analysis by quantifying the conditions necessary to invalidate an inference. Furthermore, like Rosenbaum's approach, we explore how extreme values would establish limits or bounds on significance levels, while like Frank's approach, we develop our indices in terms of the general linear model. But critically, we differentiate our approach from that of Rosenbaum, Robins, and Frank because we focus here on the representation of the sample, instead of on alternative explanations associated with selection bias as exemplied by control functions (Gill and Robins 2001; Robins 1987; Rosenbaum 1986, 2002) or confounding (Frank 2000). That is, our focus is more on external validity whereas most previous work has focused on internal validity.

In motivation and derivation, our indices also resemble those associated with assessment of publication bias in meta-analysis (e.g., Rosenthal 1979). We will attend to unobserved cases similar to those in the file drawer, distinct from the data used to obtain an estimate. But our indices will differ from the fail-safe n substantively and technically. Substantively, publication bias is induced because those studies with smaller effects are less likely to be published and therefore less likely to be observed by the meta-analyst (e.g., Hedges 1992). In contrast, our indices will quantify the concern of the skeptic regarding representation of the sample, without reference to a specific censoring mechanism.

Technically, because we develop our indices in terms of zero order and partial correlation coefficients, our approach is directly linked to the general linear model (our indices also have a direct extension to

the multivariate case; see Orwin [1983]), unlike the fail-safe n, which is specialized for meta-analysis. Furthermore, the file drawer problem, of course, refers to meta-analysis in which the individual cases are themselves studies, whereas our indices refer to single studies in which the individual cases are people. We comment more on this difference when comparing our approach with recent extensions of the fail-safe n (in Section 9.3).

3. AN IDEAL SAMPLE OF POTENTIALLY OBSERVED AND POTENTIALLY UNOBSERVED SUBPOPULATIONS

The prima facie challenge to generalizing to a target population in the social sciences is as follows: When subjects were not randomly sampled from some larger population, the results might not be generalized beyond the sample. To delve deeper, consider the structural model for the analysis by Finn and Achilles (1990):

$$Achievement = \beta_0 + \beta_1\, small\ class, \tag{1}$$

where *small class* takes a value of 1 if the classroom was small, 0 otherwise. Using the baseline model in (1), we introduce the concept of an ideal sample as one for which the estimate equals the population parameter. In this case, the ideal sample is one for which $\hat{\beta}_1^{\text{ideal}} = \beta_1$. Of course, if a sample is randomly drawn and a consistent estimator is used, $\mathrm{E}(\hat{\beta}_1) = \hat{\beta}_1^{\text{ideal}} = \beta_1$. In other words, $\hat{\beta}_1$ will not equal β_1 only because of sampling error. But here we will focus on the systematic difference between $\hat{\beta}_1$ and $\hat{\beta}_1^{\text{ideal}}$ that is due to differences in the composition of the samples.

To quantify the systematic difference between an observed sample and an ideal sample, we define $b = \hat{\beta}_1 - \hat{\beta}_1^{\text{ideal}}$. We can then quantify robustness with a question: How great would b have to be to invalidate a causal inference? In the particular example presented here, how great would the difference have to be between the estimated effect of small classes in the Tennessee class size experiments and the estimated effect from a sample that is ideal for some target population to invalidate the inference that students learn more in small classes in that target population?

Defining b through the comparison of $\hat{\beta}_1$ and $\hat{\beta}_1^{\text{ideal}}$ quantitatively expresses the notion of constancy of effect that is essential to causal

inference. Gilbert and Mosteller (1972: p. 376) explain it this way: "When the same treatment, under controlled conditions, produces good results in many places and circumstances, then we can be confident we have found a general rule. When the payoff is finicky—gains in one place, losses in another—we are wary because we can't count on the effect." In a similar vein, Shadish et al. (2002, p. 87) list their threats to external validity in terms of variable effects as represented by interactions. In absolute terms there is constancy of effect only when $\hat{\beta}_1 = \hat{\beta}_1^{\text{ideal}}$. But in the pragmatic terms of robustness, we seek to quantify how large the difference between $\hat{\beta}_1$ and $\hat{\beta}_1^{\text{ideal}}$ must be such that the inference from $\hat{\beta}_1$ would not be made from $\hat{\beta}_1^{\text{ideal}}$.

Now, drawing on mixture models (McLachlan and Peel 2000), we assume that $\hat{\beta}_1^{\text{ideal}} = (1-\pi)\hat{\beta}_1^{\text{ob}} + \pi\hat{\beta}_1^{\text{un}}$, where $\hat{\beta}_1^{\text{ob}}$ is the estimate of β_1 from the observed sample (e.g., the Tennessee schools from the 1980s that volunteered for the study); $\hat{\beta}_1^{\text{un}}$ is the estimate for cases that should be included in an ideal sample but which were unobserved (e.g., non-volunteer schools in Tennessee); and π represents the proportion of the ideal sample that is constituted by the unobserved cases.[2] (The distinction between the observed and unobserved populations concerns the mechanisms of selection into the sample, which we discuss in more detail in Section 6.)

To focus on the systematic difference between $\hat{\beta}_1^{\text{ideal}}$ and $\hat{\beta}_1$ that is generated by sample composition, note that the sampling error in $\hat{\beta}_1^{\text{ob}}$ recurs in $\hat{\beta}_1^{\text{ideal}}$ and assume $\hat{\beta}_1^{\text{un}}$ with $\text{E}(\hat{\beta}_1^{\text{un}}) = \beta_1^{\text{un}}$. Now the focal research question of this article can be phrased in terms of the unobserved quantities: What combination of $\hat{\beta}_1^{\text{un}}$ (the relationship between class size and achievement in the unobserved population of schools) and π (the proportion of unobserved schools occurring in an ideal sample) is necessary to invalidate the original inference?[3] Critically, to focus on the effect of sample composition on $\hat{\beta}_1^{\text{ideal}}$, we assume that there is no bias in $\hat{\beta}_1^{\text{un}}$ that can be attributed to omitted confounding variables. In our example, this could be accomplished if $\hat{\beta}_1^{\text{un}}$ were estimated from a randomized experiment like that used to estimate $\hat{\beta}_1^{\text{ob}}$.

[2]Note that π need not correspond to a population parameter; it is simply the proportion of the unobserved sample that occurs in an ideal sample.

[3]Alternatively, we could define the bias of $\hat{\beta}_1^{\text{ob}}$ as $\hat{\beta}_1^{\text{ob}} - \beta_1 = \pi(\hat{\beta}_1^{\text{ob}} - \beta_1^{\text{un}})$, and our question could be rephrased as "How much bias in $\hat{\beta}_1^{\text{ob}}$ must there be to invalidate the original inference?"

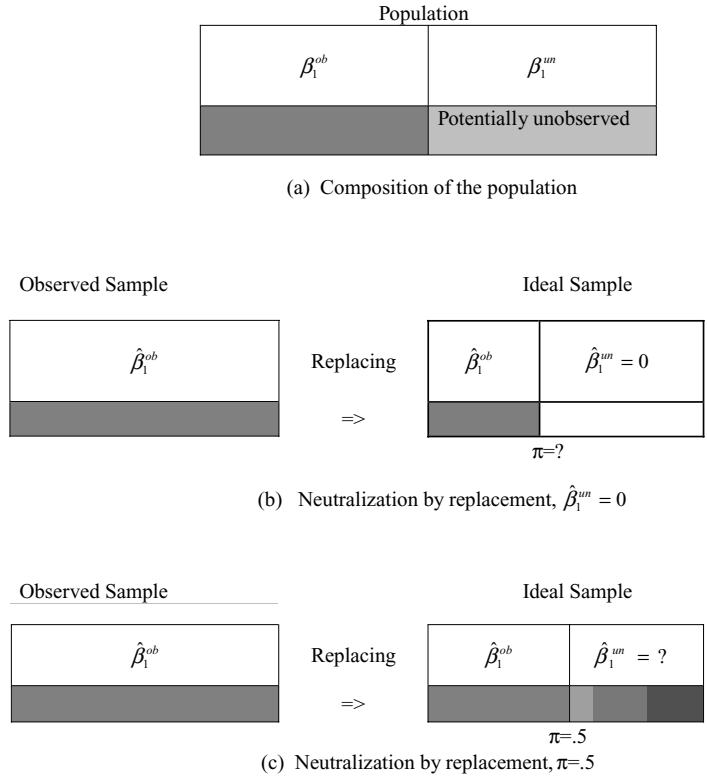

FIGURE 1. Population and sample coefficients and proportions for replacement.

As shown in Figure 1, our conceptualization does more than make the typical distinction between units that were sampled and those that were not. In our framework, we consider the population to consist of a potentially observed and unobserved subpopulation as in Figure 1(a). On the left of Figures 1(b) and 1(c), any observed sample consists of units drawn only from the potentially observed subpopulation. The ideal samples are shown on the right. As examples, an ideal sample might be achieved by replacing an unknown proportion ($\pi = ?$) with cases for which $\hat{\beta}_1^{\text{un}} = 0$ (as shown via the clear box below $\hat{\beta}_1^{\text{un}}$ in part 1(b), where shading indicates the magnitude of the coefficient) or by replacing half the sample ($\pi = .5$) with cases for which $\hat{\beta}_1^{\text{un}}$ is unknown (as shown by the multiple possible shades below $\hat{\beta}_1^{\text{un}}$ in 1(c)). Recomposition through the proportion replaced and the value of $\hat{\beta}_1^{\text{un}}$

will be explored in the development of our indices.[4] Critically, the movement from the left to the right side of parts 1(b) and 1(c), from observed to ideal sample, is hypothetical in conceptualization—the sample on the right, by definition, will never be observed.

4. INDICES OF ROBUSTNESS FOR SAMPLE REPRESENTION

Given that a regression coefficient is positive and statistically significant, in this section we use the distinction between $\hat{\beta}_1^{\text{ob}}$ and $\hat{\beta}_1^{\text{ideal}}$ to derive indices of the robustness of a causal inference to concerns regarding the representation of the sample.[5] We motivate our indices by asking "How robust is an inference to hypothetical modifications of the sample to make it more representative of a particular target population?"

We derive our indices in terms of sample correlations because they are a commonly understood metric of effect size. But we note that the statistical test for a correlation or partial correlation is equivalent to that for a regression coefficient (Cohen and Cohen 1983; Fisher 1924). As basic notation, define r_{xy}^{ob} as the statistically significant sample correlation coefficient for the observed cases, r_{xy}^{un} as the sample correlation coefficient for the unobserved cases, and r_{xy}^{ideal} as the correlation coefficient for the ideal sample based on a combination of observed and unobserved cases. Similarly, ρ_{xy}^{ob}, ρ_{xy}^{un}, and ρ_{xy} are the correlations in the potentially observed, potentially unobserved, and combined populations, respectively.

4.1. *Neutralization by Replacement*

Inferences for correlation coefficients are based on sample sizes, means, and variances of the predictor of interest (X) and the outcome (Y). To quantify the robustness of an inference with respect to the representation of a sample, we begin by assuming that means and variances of X and Y are the same for potentially observed and unobserved samples. If

[4]If one has more substantial information regarding the observed sample or population, values other than $\hat{\beta}_1^{\text{un}} = 0$ and $\pi = .5$ in parts (b) and (c), respectively, could be used.

[5]See technical Appendix A for a quick reference to all of the indices developed in this article.

the means were not identical, differences in means could be accounted for by adding hypothetical indicators of whether the data were potentially observed (e.g., volunteered) or not to model (1), which adjust for different central tendencies. The assumption of homogeneous variances is consistent with standard assumptions made for inferences from general linear models such as regression or analysis of variance. Moreover, the key point here is that the framework for generating the indices is purely hypothetical, and in this hypothetical context we focus on the indicator of a relationship, the covariance—not on the characteristics of context, such as means and variances. Nonetheless, in Section 4.3, we relax the assumptions of equal means and variances.

Given that the means and variances of X and Y are the same for potentially observed and potentially unobserved samples, r_{xy}^{ideal} is a simple weighted average of r_{xy}^{un} and r_{xy}^{ob}. Thus if n^{ob} is the number of originally observed cases, and n^{un} is the number of cases to be replaced by unobserved cases with correlation r_{xy}^{un}, then

$$r_{xy}^{ideal} = \left[\left(n^{ob} - n^{un}\right)r_{xy}^{ob} + n^{un}r_{xy}^{un}\right]/n^{ob}. \tag{2}$$

Defining π as the proportion of the sample that is replaced, $\pi = n^{un}/n^{ob}$, then

$$r_{xy}^{ideal} = (1-\pi)r_{xy}^{ob} + \pi r_{xy}^{un}. \tag{3}$$

To establish the conditions under which the original inference would be invalid, we compare the value of r_{xy}^{ideal} to a quantitative threshold of the same sign, $r^{\#}$. Thus the inference based on $r_{xy}^{ob} \geq r^{\#}$ is invalid if $r_{xy}^{ideal} < r^{\#}$, which implies

$$(1-\pi)r_{xy}^{ob} + \pi r_{xy}^{un} < r^{\#}. \tag{4}$$

The quantity $r^{\#}$ could be defined by an effect size, such as .2, considered large enough to be the basis of causal inference. In fact, we will consider thresholds based on specific effect sizes throughout this article.

To define $r^{\#}$ by statistical significance, begin by noting (Cohen and Cohen 1983:52)

$$t = \frac{r\sqrt{n-q}}{\sqrt{1-r^2}}, \tag{5}$$

where q is the number of parameters estimated (including the intercept, and the parameters for X and any other covariates) and t is the ratio for assessing the statistical significance of r. We then obtain the value of r that is just statistically significant, the threshold $r^{\#}$, by setting $r = r^{\#}$ $t = t_{\text{critical}}$ and solving for $r^{\#}$:

$$r^{\#} = \frac{t_{\text{critical}}}{\sqrt{(n^{ob} - q) + t^2_{\text{critical}}}}. \tag{6}$$

The threshold in (6) also could be interpreted as an effect size that has a 50 percent probability of being statistically significant in an entirely new sample.

Regardless of the threshold used to define $r^{\#}$, the relative trade-off between replacement proportion and the missing correlation can be represented by solving (4) for π:

$$\pi > (r^{\#} - r^{ob}_{xy})/(r^{un}_{xy} - r^{ob}_{xy}). \tag{7}$$

For example, for values of $r^{ob}_{xy} = .1$, .3 and .5 ($n^{ob} = 500$) the relationship between π and r^{un}_{xy}, using statistical significance with $p \leq .05$ as a threshold (i.e., $r^{\#} = .09$),[6] is shown in Figure 2. The area above a given curve indicates the region in which the initial inference is invalid, and the area below a given curve indicates the region in which the inference is valid. Thus, we refer to curves such as those shown as in Figure 2 as robustness curves. We see that π increases exponentially with $r^{un}_{xy,}$ with π approaching 1 when $r^{un}_{xy} = .09$, the value of $r^{\#}$. Note also that the curvature is more pronounced for lower values of r^{ob}_{xy}, indicating that for smaller r^{ob}_{xy}, robustness increases rapidly only when r^{un}_{xy} approaches $r^{\#}$.

Critically, Figure 2 suggests two key points as a basis for quantifying robustness. First, the y-intercept, occurring at the dotted line defined by $r^{un}_{xy} = 0$, corresponds to the challenge that the sample did not represent a critical subpopulation for which $\rho^{un}_{xy} = 0$. Assuming $r^{un}_{xy} = \rho^{un}_{xy} = 0$,[7] (7) becomes

[6]We will consider thresholds defined by effect size as well as statistical significance in the empirical examples.

[7]By definition, we cannot observe sample statistics for the unobserved population. Therefore we assume that the unobserved sample statistics equal the

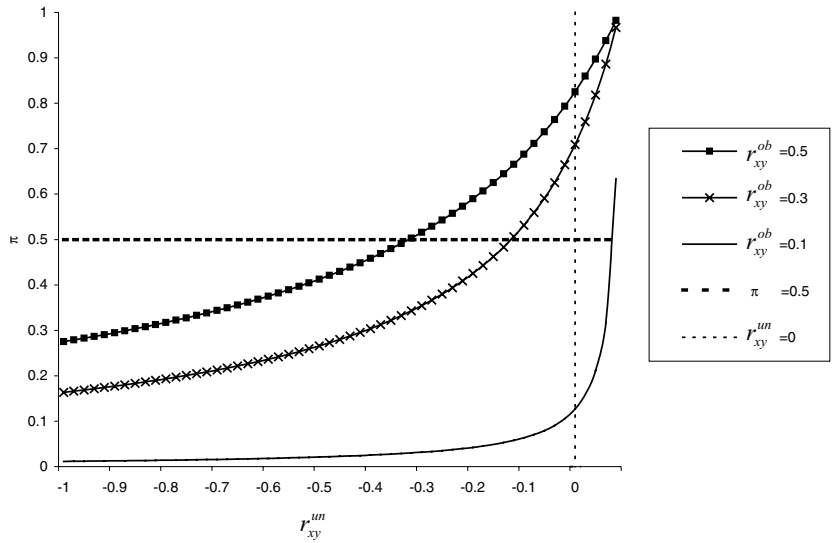

FIGURE 2. Robustness curves defined by the sample proportion (π) and correlation coefficient (r_{xy}^{un}) of unobserved cases ($n^{ob} = 500$ and $\alpha = 0.05$) necessary to make the overall inference invalid.

$$\pi > 1 - \mathrm{r}^{\#}/r_{xy}^{ob}. \tag{8}$$

Thus, assuming $r_{xy}^{un} = \rho_{\mathrm{xy}}^{\mathrm{un}} = 0$, if π is greater than $1 - r^{\#}/r_{xy}^{ob}$, the inference would be invalid if π of the sample were replaced. The right-hand side of (8), $1 - \mathrm{r}^{\#}/r_{xy}^{ob}$, defines the *i*ndex of *e*xternal *v*alidity for $r_{xy}^{un} = 0$ and *r*eplacement, or $IEVR(\pi, r_{xy}^{un} = 0)$.[8] Thus $IEVR(\pi, r_{xy}^{un} = 0)$ defines the

unobserved population parameters. Alternatively, we could write $\mathrm{E}[r_{xy}^{un}] = \rho_{\mathrm{xy}}^{\mathrm{un}} = 0$ and correspondingly develop all further expressions based on expected values. But, for the ease of notation we assume there is no sampling error and develop our indices using symbols for the unobserved sample correlation.

[8]It can also be that the observed sample consists of two subsamples: one for which $r_{xy} > 0$ and the other for which $r_{xy}^{un} = 0$, mixed in proportion to $\pi^{+} = n^{+}/n$, the proportion in the observed sample with $r_{xy} > 0$. Clearly the overall inference would not change to the extent that cases for which $r_{xy} = 0$ were replaced by unobserved cases for which $r_{xy} = 0$. Thus we could calculate the proportion of the observed sample for which $r_{xy} > 0$ that would need to be replaced to make the

proportion of the observed sample that must be replaced with cases for which the nil hypothesis is true to make the inference invalid.

Generally, $IEVR(\pi, r^{un}_{xy} = 0)$ is large when $r^{\#}/r^{ob}_{xy}$ is small, and therefore when r^{ob}_{xy} is much larger than $r^{\#}$; the index reflects the extent to which r_{xy} exceeds $r^{\#}$. The index also is well bounded: $0 \leq IEVR(\pi, r^{un}_{xy} = 0) < 1$. The left inequality holds because $\mathrm{r}^{\#}/r^{ob}_{xy}$ cannot be greater than 1 because the starting condition for our derivation is that r^{ob}_{xy} is greater than the threshold defined by $r^{\#}$. The right inequality holds because $\mathrm{r}^{\#}/r^{ob}_{xy}$ cannot be less than zero because both quantities take the same sign by definition of $r^{\#}$. As a particular example, for $r^{ob}_{xy} = .50$ and $n^{ob} = 500$ as in Figure 2, the $IEVR(\pi, r^{un}_{xy} = 0) = .82$, indicating that 82 percent of the original sample would have to be replaced with cases for which $r^{un}_{xy} = 0$ to invalidate the original inference. In contrast, only 71 percent of the cases would have to be replaced if $r^{ob}_{xy} = .3$ and 12 percent if $r^{ob}_{xy} = .1$.

To develop a second index, focus on the midpoint at $\pi = .5$ defined by the dashed lines in Figure 2. Computationally, begin with (4), substitute $\pi = .5$ and solve for r^{un}_{xy}:

$$r^{un}_{xy} < 2\mathrm{r}^{\#} - r^{ob}_{xy}. \tag{9}$$

Thus if $r^{un}_{xy} < 2r^{\#} - r^{ob}_{xy}$, the original inference would be invalid if half the sample were replaced with cases from the potentially unobserved subpopulation. We refer to $2r^{\#} - r^{ob}_{xy}$ as the *i*ndex of *e*xternal *v*alidity for $\pi = .5$ and replacement, or $IEVR(\pi = .5, r^{un}_{xy})$. Note that the $IEVR(\pi = .5, r^{un}_{xy})$ is a linear function of r^{ob}_{xy} as can be observed from (9). Examples in Figure 2 indicate that for $r^{ob}_{xy} = .5$ the $IEVR(\pi = .5, r^{un}_{xy}) = -.32$; for $r^{ob}_{xy} = .3$ the $IEVR(\pi = .5, r^{un}_{xy}) = -.12$; and for $r^{ob}_{xy} = .1$ the $IEVR(\pi = .5, r^{un}_{xy}) = .08$. In each case, if r^{un}_{xy} is less than the $IEVR(\pi = .5, r^{un}_{xy})$, the inference would be invalidated if half the sample were replaced with cases from the unobserved subpopulation.[9]

inference invalid, merely by defining $\pi = n^{un}/n^{+}$ instead of $\pi = n^{un}/n^{ob}$ as above. But this would imply that there are existing, nonignorable, interactions in the observed data, and so we do not present this as our main index.

[9]Interestingly, for $r^{ob}_{xy} = .1$, $IEVR(\pi = .5, r^{un}_{xy}) = .08$, indicating that there must be little difference between r^{ob}_{xy} and r^{un}_{xy} for the inference to be valid.

Both the $IEVR(\pi, r_{xy}^{un} = 0)$ and the $IEVR(\pi = .5, r_{xy}^{un})$ are indices of the robustness of an inference to concerns regarding the representation of a population. The critical difference is that $IEVR(\pi, r_{xy}^{un} = 0)$ accepts the nil hypothesis, that $r_{xy}^{un} = \rho_{xy}^{un} = 0$, and then considers sample recomposition through π. On the other hand, $IEVR(\pi = .5, r_{xy}^{un})$ does not accept the nil, but begins with the hypothesis that the potentially unobserved sample is as large as the potentially observed sample, and then determines r_{xy}^{un} needed to generate a different inference from the ideal sample than from the observed sample. Thus the two indices quantify different aspects of sample recomposition.

Critically, because the sampling distribution for a partial correlation is equivalent to that for a zero-order correlation (Cohen and Cohen 1983; Fisher 1924), either $IEVR(\pi, r_{xy}^{un} = 0)$ or $IEVR(\pi = .5, r_{xy}^{un})$ can be extended readily to models containing covariates, $\boldsymbol{u}$. In particular, replace r_{xy}^{ob} with $r^{ob}{}_{xy|\boldsymbol{u}}$ and note that the degrees of freedom are adjusted by q as in (5) and (6) (see Cohen and Cohen 1983:103–7; Fisher 1924). Thus the indices can be used in conjunction with the general linear model, especially models that use covariates to make differences between treatment and control ignorable (Winship and Sobel 2004).

We note three assumptions we made as we developed our indices. First, our scenario began with an estimate of a regression coefficient from a general linear model. As such, we assumed that Y is a continuous variable and the relationship between X and Y is linear. Furthermore, we assumed that the model is correctly specified in that all relevant confounding variables have been controlled for and thus there is no remaining spurious component in the estimated relationship between X and Y (cf. Frank's index [2000] for robustness due to omitted confounding variables or Rosenbaum's index [1986, 2002] for selection bias). Correspondingly, the extension to the partial correlation in the preceding paragraph allows us to accommodate models that employ statistical control.

Second, the mixture model implies that Y may not be normally distributed (e.g., Y may be bimodal because there are two distinct subpopulations). Thus if the unobserved data were in fact observed, we could estimate β_1^{ob}, β_1^{un}, and π via a finite mixture model (see McLachlan and Peel 2000). But our scenario is purely hypothetical; by definition the observed sample does not include units from the potentially unobserved subpopulation. Therefore we cannot estimate the mixture model. Moreover, because model fit only improves with the

number of components estimated in a mixture model, standard significance tests assuming only a single component are conservative, and therefore the indices we developed based on significance tests will be conservative. Thus, we use the mixture model primarily as a rhetorical device to illuminate the assumptions of, and challenges to, inference (and, in Section 8, we consider a Bayesian motivation for the indices). The derivation of our indices is, however, based on the maximum likelihood estimates that would be obtained if the unobserved sample were available for estimation via a mixture model (Day 1969:464; Lindsay and Basak 1993; McLachlan and Peel 2000).

Third, though we accept the initial inference regarding β_1, we recognize that there is always the possibility of a Type I or Type II error due to sampling variation. (Type I is a rejection of the null hypothesis when in fact it is true; Type II is a failure to reject the null hypothesis when in fact it is false.) Essentially, the concern regarding Type I and Type II errors relative to our indices is neither greater nor less than it would be for any statistical inference, as the initial inference is the baseline for our indices.

4.2. *Neutralization by Addition (Contamination)*

Instead of replacing potentially observed cases with potentially unobserved cases as in the previous subsection, consider instead augmenting a data set with further observations. The effect on the sample is described as contamination by Donoho and Huber (1983).[10] The overall sample size increases, with the ideal sample achieved by adding an unknown number of cases with $\hat{\beta}_1^{\text{un}}$ (or r_{xy}^{un}) $= 0$ and π unknown or with $\pi = .5$ and $\hat{\beta}_1^{\text{un}}$ (or r_{xy}^{un}) unknown. The fundamental relationship between π and r_{xy}^{un} remains as defined in (7).

To develop expressions for indices for added cases we reexpress (2) in terms of the sizes of the potentially observed and potentially unobserved samples without the constraint of preserving the overall

[10]In our framework the new cases contaminate if they make the inference invalid, just as for Donoho and Huber (1983) the cases contaminate if they "break" the estimator, although both types of contamination are merely objective phenomena from a statistical standpoint.

sample size:

$$r_{xy}^{ideal} = \left(n^{ob} r_{xy}^{ob} + n^{un} r_{xy}^{un}\right) / \left(n^{ob} + n^{un}\right). \tag{10}$$

As in (7), to obtain an index for $r_{xy}^{un} = \rho_{xy}^{un} = 0$ based on contamination, begin by setting $r_{xy}^{ideal} < \mathrm{r}^{\#}$. If effect size is used to define the threshold of inference, then we merely specify the value of $r^{\#}$ in terms of a specific effect size, such as .2. On the other hand, calculations are more complex if $r^{\#}$ is specified in terms of statistical significance because the threshold changes with the sample size. In particular, noting that $r^{\#}$ is now a function of n^{ob} and n^{un}, substitute $\mathrm{r}^{\#}$ for r_{xy}^{ideal} in (10), then using (5) reexpress $r^{\#}$ in terms of n and t_{critical}, solve for n^{un} and define it as n^{un*}:

$$n^{un*} = \frac{n^{ob}\left[n^{ob}\left(r_{xy}^{ob}\right)^2 - 2t_{\text{critical}}^2 + r_{xy}^{ob}\sqrt{(n^{ob})^2\left(r_{xy}^{ob}\right)^2 + 4t_{\text{critical}}^2(t_{\text{critical}}^2 - q)}\right]}{2t_{\text{critical}}^2}. \tag{11}$$

Then if $n^{un} > n^{un*}$, the original inference is invalid.[11] Calculating π as the proportion $n^{un*}/(n^{ob} + n^{un*})$ then defines an *i*ndex of *e*xternal *v*alidity for $r_{xy}^{un} = 0$ and *c*ontamination, or $IEVC(\pi, r_{xy}^{un} = 0)$. That is, $IEVC(\pi, r_{xy}^{un} = 0)$ indicates what proportion of cases in the ideal sample must come from the unobserved population to invalidate the inference from the observed data. Thus if $\pi \geq n^{un*}/(n^{ob} + n^{un*})$, then the inference from the observed data is invalid.

Next, the fundamental relationship defining the replacement index for $\pi = .5$ is not a function of the sample size. Furthermore, if $\pi = .5$, then $n^{un} = n^{ob}$. The combined sample size is thus known, as is the new significance level:

$$r^{\#\#} = \frac{t_{\text{critical}}}{\sqrt{(2n^{ob} - q) + t_{\text{critical}}^2}}. \tag{12}$$

Thus replacing $r^{\#}$ in (9) with $r^{\#\#}$ in (12), if $r_{xy}^{un} < 2r^{\#\#} - r_{xy}^{ob}$ then the inference would be altered if the sample were doubled by adding

[11] We could also correct the critical value of t for changes in degrees of freedom (see Min and Frank 2002), although this correction is likely to be small; for $n > 60$; $t(\mathrm{df} = 60) = 2.000$ is only .04 greater than $t(df = \infty) = 1.96$.

cases with r_{xy}^{un}. The quantity $2r^{\#\#} - r_{xy}^{ob}$ then defines the *i*ndex of *e*xternal *v*alidity for $\pi = .5$ and *c*ontamination, or $IEVC(\pi = .5, r_{xy}^{un})$. That is, if $r_{xy}^{un} \leq 2r^{\#\#} - r_{xy}^{ob}$, then the original inference based on the observed data is invalid.

4.3. *A General Formula for r_{xy}: Relaxing Assumptions of Equal Means and Variances*

In the preceding subsections we assumed that the means and variances of X and Y were the same for potentially observed and unobserved samples, arguing that these were easily adjusted for or that they were analogous to population assumptions for the general linear model. Furthermore, we cannot use a complex mixture model that relaxes these assumptions to derive closed form expressions for robustness indices because "explicit formulas for parameter estimates [for mixture finite models] are typically not available" (McLachlan and Peel 2000:25). We can, however, draw on mixture models to present a general expression for r_{xy}, the correlation in a combined sample, that does not assume equal means and variances between subsamples.

Drawing on Day (1969: 466), and defining s_{xy} as a covariance, an expression for a sample covariance for the mixture of an observed and unobserved component is

$$s_{xy} = (1-\pi)s_{xy}^{ob} + \pi s_{xy}^{un} + (1-\pi)\pi(\bar{x}^{ob} - \bar{x}^{un})(\bar{y}^{ob} - \bar{y}^{un}).$$

Using this general result, as derived in Appendix B, an expression for r_{xy} that does not assume equal means and variances is:[12]

$$r_{xy} = \frac{(1-\pi)r_{xy}^{ob}s_x^{ob}s_y^{ob} + \pi r_{xy}^{un}s_x^{un}s_y^{un} + (1-\pi)\pi\left(\bar{x}^{ob} - \bar{x}^{un}\right)\left(\bar{y}^{ob} - \bar{y}^{un}\right)}{\sqrt{\begin{array}{l}\left[(1-\pi)\left(s_x^{ob}\right)^2 + \pi\left(s_x^{un}\right)^2 + (1-\pi)\pi\left(\bar{x}^{ob} - \bar{x}^{un}\right)^2\right] \\ \times\left[(1-\pi)\left(s_y^{ob}\right)^2 + \pi(s_y^{un})^2 + (1-\pi)\pi\left(\bar{y}^{ob} - \bar{y}^{un}\right)^2\right]\end{array}}}. \quad (13)$$

Note that the numerator is a function of weighted covariances plus a correction based on $(1-\pi)\pi(\bar{x}^{ob} - \bar{x}^{un})(\bar{y}^{ob} - \bar{y}^{un})$. The correction

[12]Expressions for the first four moments for the distribution of r_{xy}, using Fisher z to transform r_{xy} to be approximately normally distributed, can be found in Day (1969) and Lindsay and Basak (1993).

contributes positively to r_{xy} if $(\bar{x}^{ob} - \bar{x}^{un})$ and $(\bar{y}^{ob} - \bar{y}^{un})$ take the same sign. This holds because the correction represents a between sample (observed versus unobserved) contribution to the correlation. In the example of the class size study, the correction would decrease the relationship between small classes and achievement if the classes in the unobserved sample (e.g., nonvolunteer classes) had lower achievement, making $(\bar{y}^{ob} - \bar{y}^{un})$ positive, and there were more small classes in the unobserved sample than in the observed sample, making $(\bar{x}^{ob} - \bar{x}^{un})$ negative (the resulting product of $[\bar{x}^{ob} - \bar{x}^{un}]$ and $[\bar{y}^{ob} - \bar{y}^{un}]$ would then be negative).

To focus on the assumption of equal variances, consider $\bar{x}^{ob} = \bar{x}^{un}$ and $\bar{y}^{ob} = \bar{y}^{un}$. Then

$$r_{xy} = \frac{(1-\pi) r_{xy}^{ob} s_x^{ob} s_y^{ob} + \pi r_{xy}^{un} s_x^{un} s_y^{un}}{\sqrt{\begin{array}{l}\left\{\left(s_x^{un}\right)^2 + (1-\pi)\left[\left(s_x^{ob}\right)^2 - \left(s_x^{un}\right)^2\right]\right\} \\ \times \left\{(s_y^{un})^2 + (1-\pi)\left[(s_y^{ob})^2 - (s_y^{un})^2\right]\right\}\end{array}}}, \tag{14}$$

which represents a simple weighted correlation coefficient, with weights proportional to π and $(1-\pi)$. Thus *a priori* beliefs about the values of the unobserved variances can be inserted into (14). That is, this equation can be used to broaden the scope of external validity to account for different scales used to measure treatments or outcomes between the observed and unobserved samples. Finally, note that if the variances in the unobserved sample equal those of the observed sample, then (14) reduces to the special case we focus on in (3).

5. EXAMPLE: THE INFERENCE THAT SMALL CLASSES IMPROVE ACHIEVEMENT

Recall that in our example the inference made by Finn and Achilles (1990) that smaller classes improve achievement was based on a nonrandom sample of volunteer elementary schools in Tennessee in the mid-1980s. We also note that Finn and Achilles and others explored various interaction effects. In particular, Finn and Achilles reported no significant differences of the class size effect by location or grade, and Nye et al. (2002) reported that differences by prior level of achievement were not statistically significant (there were some small differences by ethnicity, on which we will comment later in the discussion). Thus, given

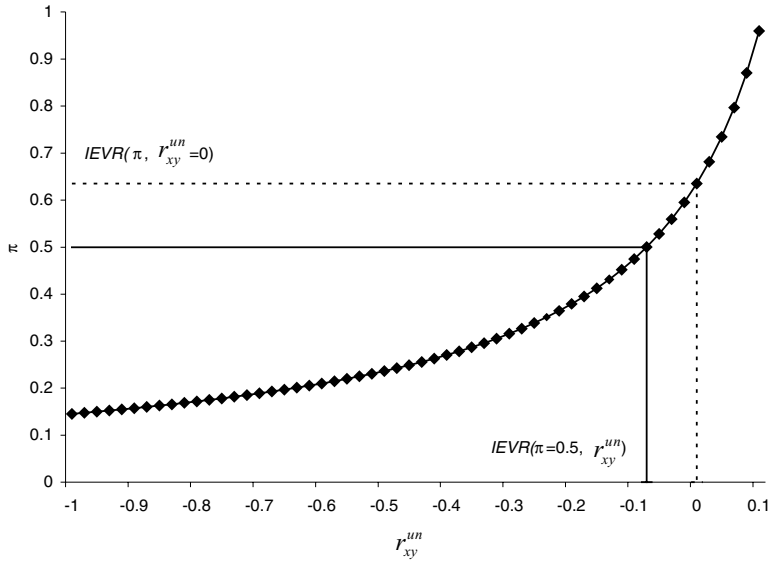

FIGURE 3. Robustness curve and $IEVR(\pi, r_{xy}^{un} = 0)$ and $IEVR(\pi = .5, r_{xy}^{un})$ for the Tennessee class size study by Finn and Achilles (1990).

fairly stable effects across subsamples, we now ask: What must be π and r_{xy}^{un} to invalidate the overall inference that would be made from an ideal sample representing some population of schools other than those that volunteered in Tennessee in the mid 1980s?

A robustness curve for Finn and Achilles' (1990) results is shown in Figure 3 ($r_{xy}^{ob} = .296$, $r^{\#} = .107$). The dashed lines indicate that the intersection at $r_{xy}^{un} = 0$ occurs for $\pi = .64$, which defines the $IEVR$ $(\pi, r_{xy}^{un} = 0)$ as presented in (8). Thus 64 percent or more of the volunteer schools would have to be replaced with a sample for which $r_{xy}^{un} = 0$ to invalidate the original inference from the observed data. As a complement, the solid lines indicate that the intersection at $\pi = .5$ occurs at $-.08$, which is the $IEVR(\pi = .5, r_{xy}^{un})$ as defined in (9). Thus, assuming half the sample were replaced, r_{xy}^{un} would have to be less than $-.08$ to invalidate the original inference. If $r^{\#} = .2$ were used to establish the threshold, the $IEVR(\pi = .5, r_{xy}^{un})$ would equal .10 and $IEVR(\pi, r_{xy}^{un} = 0)$ would be 32 percent. Thus, as might be expected for a relatively large sample, robustness is not as great if the threshold is defined by a moderate effect size instead of by statistical significance.

Following Rosenbaum's (2002) approach of placing a bound on the overall estimate, if $r_{xy}^{un} = -1$, then approximately 15 percent of the sample would have to be replaced to invalidate the inference. Refer to this as $\pi\ |(r_{xy}^{un} = -1)$, the lower bound on π. Clearly there is no specific upper bound for π as r_{xy}^{un} approaches 1, because if $r_{xy}^{un} > r_{xy}^{ob}$ the inference would be valid regardless of the value of π. Thus $.15 \leq \pi \leq 1$.

Regarding bounds for r_{xy}^{un}, clearly the inference changes only if $r_{xy}^{un} < \mathrm{r}^{\#}$. Thus there is no need to consider $r_{xy}^{un} \geq \mathrm{r}^{\#}$ in terms of the robustness of the inference. Finally, as implied by the original bound on π, if $\pi < .15$ then there is no value of r_{xy}^{un} that can make the original inference invalid. Thus $-1 \leq r_{xy}^{un} \leq \mathrm{r}^{\#}$; in this case $-1 \leq r_{xy}^{un} \leq .107$.

Now consider adding cases instead of replacing them. Then 2153 cases for which $r_{xy}^{un} = 0$ would have to be added to alter the inference in an ideal sample. The new cases would comprise 87 percent of an ideal sample, as defined by the $IEVC(\pi, r_{xy}^{un} = 0)$ in (11). Alternatively, if the sample size were doubled by adding cases from the potentially unobserved population, the inference would be invalid only if r_{xy}^{un} were less than $-.14$, as defined by the $IEVC(\pi = .5, r_{xy}^{un})$ in Section 4.2.

Calculation of the robustness indices by no means resolves the debate regarding the validity of the inference that small classes have a positive effect on achievement. But the debate has now been quantified in terms of the relationship between class size and achievement in cases from the unobserved population and proportional representation in an ideal sample. Now, those making the causal inference cannot merely claim that attempts were made to recruit all Tennessee schools and that the volunteer schools were similar to others—they must claim that those schools that volunteered were at least representative of 36 percent of the population or that $r_{xy}^{un} > -.08$. Similarly, critics of the causal inference cannot merely suggest that there were potential threats to external validity (such as nonrandom sampling and differential attrition). They must argue that such threats would have rendered 64 percent of the sample nonrepresentative or that if half the sample were replaced to construct an ideal sample, $r_{xy}^{un} < -.08$.

6. ROBUSTNESS INDICES AND THE BREADTH OF EXTERNAL VALIDITY

Our framework and resultant indices can be interpreted in two ways in terms of the breadth of external validity (cf. compare Cronbach [1982]

with Shadish et al. [1991]). In the narrowest sense, our indices quantify how robust inferences are to generalizations to the population from which the cases were directly sampled. In our example, the narrow domain refers to schools in Tennessee in the mid-1980s, including those who did not volunteer as well as those who did volunteer. Thus the key distinction between the two subpopulations derives from the mechanics of sampling that caused attrition or nonresponse. In the broadest sense, the indices can be interpreted in terms of external validity beyond the immediate sampling frame. In the example of the Tennessee class size study, Hanushek (1999) refers to aggressive efforts to reduce class size in California in the late 1990s that were motivated in part by the Tennessee class size studies (CSR 1998). Here the distinction between the subpopulations is based on the intent of the researchers and policymakers.

Clearly, classrooms in California in the late 1990s were not part of the original sample frame for a study conducted in Tennessee in the mid-1980s. Moreover, classrooms in California in the 1990s may have been less advantaged than those that volunteered for the Tennessee experiments as California tended to have low per pupil expenditures relative to the rest of the nation. (Ed Source [2003], recalling also that classrooms in the Tennessee study were more advantaged than the state as a whole in terms of per pupil expenditures and teacher salaries.) If small classes were especially helpful for advantaged classrooms, then the effect of small classes in California could be less than that found by Finn and Achilles (1990) in Tennessee, and the inference may not generalize. But our index quantifies *how much* lower the effect from a replacement sample in California would have to be to invalidate the overall inference that small classes improve achievement in some combination of schools representing Tennessee in the mid-1980s and California in the mid-1990s.[13]

Regardless of whether they define the target population in the restrictive or broad sense, social scientists must debate external validity

[13]In this particular example, any overestimate of the class size effect for California may be compensated by the higher percentages of minorities, for whom class sizes were more effective, in California than in Tennessee (25 percent of the classrooms reported on by Finn and Achilles [1990] are minority, whereas Hispanics alone account for 25 percent of the students in California [NCES 2003]). Thus in considering the extension of Finn and Achilles' results to California, policymakers and social scientists must balance the possibility of weaker effects of small classes for less advantaged schools against the stronger effects of small classes for minorities.

through scientific reasoning. For example, Shadish et al. (2002:353–54) list five principles of generalized causal inferences including surface similarity; (1) judging the apparent similarities between the things scientists study and the targets of generalization; (2) making discriminations that might limit generalization; (3) ruling out irrelevancies that do not change the generalization; (4) making interpolations and extrapolations; and (5) providing causal explanations. These are all assertions that depend on scientific reasoning.

Our indices then work in conjunction with the principles for generalization. For the more restricted interpretation of external validity for the example, Finn and Achilles (1990) established that surface similarity between the schools in the study and all schools in Tennessee is moderate to high and that most differences between the sample and the target populations were minimal. Thus relatively small values of the indices may be taken as indicators of robustness. In contrast, when social scientists consider broader generalizations, such as to California in the mid-1990s, the sample and target population may differ substantially. And it would be more difficult to rule out factors that may reduce r_{xy}^{ideal} below a threshold. Correspondingly, higher values of the indices are required to argue that an inference is robust when we seek to make broader generalizations.

7. WHETHER THE DIFFERENCE BETWEEN r_{xy}^{ob} AND r_{xy}^{un} WOULD BE STATISTICALLY SIGNIFICANT

Thinking of external validity in terms of nonadditive effects provides a framework for comparing r_{xy}^{ob} against r_{xy}^{un}. In particular, $|r_{xy}^{ob} - r_{xy}^{un}|$ could be compared against the criterion necessary to reject the hypothesis of no interaction between source of data (observed versus unobserved population) and size of effect. Formally, define the Fisher z of r: $z(r) = .5[\ln(1 + r) - \ln(1 - r)]$, and define $w = |z(r_{xy}^{ob}) - z(r_{xy}^{un})|$. Then, following Cohen and Cohen (1983:53–55), the interaction between source of data and size of correlation is statistically significant if $w > 1.96\sqrt{\frac{1}{(1-\pi)(n^{ob}-q)} + \frac{1}{\pi(n^{ob}-q)}}$, where q equals three plus the number of parameters estimated in the model. Defining $m = 1.96\sqrt{\frac{1}{(1-\pi)(n^{ob}-q)} + \frac{1}{\pi(n^{ob}-q)}}$, r_{xy}^{ob} would be statistically different from

r_{xy}^{un} if

$$r_{xy}^{ob} > \frac{r^{\#} - e^{2m}(1 + r^{\#}) - 1 - \sqrt{\begin{array}{l} 4\left(e^{2m} - 1\right)(\pi - 1)\left(e^{2m}\pi + r^{\#} + e^{2m}r^{\#} - \pi\right) \\ \quad + (1 - r^{\#} + e^{2m}(1 + r^{\#}))^2 \end{array}}}{2\left(e^{2m} - 1\right)(\pi - 1)}. \quad (15)$$

Thus the right-hand side of (15) defines the index of *e*xternal *v*alidity beyond interaction (*IEVBI*). For the Tennessee class size example, setting $\pi = .5$, the *IEVBI* is .215, which is exceeded by $r_{xy}^{ob} = .296$. Therefore, if half the sample were replaced, r_{xy}^{ob} would have to be significantly different from r_{xy}^{un} to alter the overall inference regarding the effect of small schools. Thus, either the overall inference would not change if half the cases were replaced, or the inference would change only if r_{xy}^{ob} were significantly different from r_{xy}^{un}. But, if r_{xy}^{ob} were significantly different from r_{xy}^{un}, then inferences should be made about the two populations separately, and thus the original *i*nference based on r_{xy}^{ob} applies at least to the population from which the observed cases were drawn.

Generally, the *IEVBI* offers an important clarification to debates regarding the presence of interactions on making causal inferences, with some arguing for the ability to make a causal inference even if the effect varies across subsamples (e.g., Cook and Campbell 1979) and others arguing that effects must always be reported by subsample (e.g., Cronbach 1982). If r_{xy}^{ob} is greater than the *IEVBI*, then perhaps the Campbell and Cronbach camps could agree that one would either report an overall effect (if r_{xy}^{un} were greater than $IEVC[\pi = .5, r_{xy}^{un}]$) or one would report separate effects (if r_{xy}^{un} were less than $IEVC[\pi = .5, r_{xy}^{un}]$). When r_{xy}^{ob} is less than the *IEVBI*, the inference is murkier; when small interactions in the data could alter the overall inference they must decide whether to report an overall effect (Campbell) or effects by subgroups (Cronbach). Perhaps this murkier situation accurately reflects the small value of r_{xy}^{ob} relative to the threshold for inference.

Of course, the above calculations assume that we use statistical significance to determine whether there is a discernable difference in the effect between the observed and unobserved samples. Alternatively, w could be compared against the criterion necessary to distinguish one component from another in terms of bimodality or significance tests

from finite mixture models that do not assume an indicator of the source of the data has been measured (see McLachlan and Peel 2000:p 11). Generally, we could describe an inference as robust in an absolute sense if it could be invalidated only if the unobserved sample must be discernibly different from the observed sample, using statistical significance, effect size, or other criteria to operationalize discernable.

8. A BAYESIAN MOTIVATION FOR THE INDICES

While we motivated our indices by considering how observed and unobserved samples combined to form an ideal estimate from a mixture model, we could also have motivated our indices from a Bayesian perspective. In particular, we define the likelihood in terms of the observed sample and the prior in terms of the sample from the potentially unobserved population. Following Lee (1989:169), the Fisher z transformation of each sample correlation is normally distributed with variance $1/n$ and is an unbiased estimate of the Fisher z of the population correlation. Therefore, the estimated posterior mean for the ideal sample is (Lee 1989:p 175):

$$\begin{aligned} z(r_{xy}^{ideal}) &= \mathrm{Var}(r_{xy}^{ideal})\left[\mathrm{Var}^{-1}\left(r_{xy}^{ob}\right)z\left(r_{xy}^{ob}\right) + \mathrm{Var}^{-1}\left(r_{xy}^{un}\right)z\left(r_{xy}^{un}\right)\right] \\ &= 1/(n^{ob} + n^{un})\left[n^{ob}z\left(r_{xy}^{ob}\right) + n^{un}z\left(r_{xy}^{un}\right)\right] \\ &= (1-\pi)z\left(r_{xy}^{ob}\right) + \pi\, z\left(r_{xy}^{un}\right). \end{aligned} \tag{16}$$

This latter expression is the Bayesian analog to our original expression for r_{xy}^{ideal} from the mixture model in (3).

Using the Bayesian approach, the posterior distribution for ρ_{xy} for the whole population is $\sim N(z[r_{xy}^{ideal}\,], [n^{ob} + n^{un}\,]^{-1})$. This posterior can then be used to quantify robustness by considering the value of r^{un} necessary to make r_{xy}^{ideal} fall within a 95 percent highest posterior density (HPD) interval (which is similar to a frequentist 95 percent confidence interval). Applying the Bayesian approach to the example of the Tennessee class size study, $IEVC^{Bayesian}\,(\pi, r_{xy}^{un} = 0) = 2129$ and $IEVC^{Bayesian}\,(\pi = .5, r_{xy}^{un}) = -0.15$, which differ slightly from $IEVC(\pi, r_{xy}^{un} = 0)$ of 2153 and the $IEVC(\pi = .5, r_{xy}^{un})$ of -0.14 (as calculated in Section 5).

Though the Bayesian formulation may be intuitive for some, we favor the frequentist approach for three reasons. First, while the

Bayesian formulation applies to the *IEVC* wherein a new estimate is obtained based on combining information from the posterior and the prior, it does not apply as well to the *IEVR* in which some of the observed information is *replaced* with that from the prior. Second, our calculations based on the mixture model are exact for r_{xy}^{ideal}, whereas the Bayesian approach is based on an approximation. Third, though the Bayesian framework is quite popular in statistics, social scientists are still inclined to apply frequentist interpretations to their analyses. For example, in reviewing Volume 70 (2005) of the *American Sociological Review,* only 3 out of 29 articles using inferential statistics made explicit use of a Bayesian approach (Cole 2005; Karinek et al. 2005; Mallard 2005).[14] Thus we appeal to the more common frequentist framework in considering the conditions necessary to alter an inference. On the other hand, few social scientists adhere to a strict frequentist interpretation as consideration of an array of possible effects is often implicit in sensitivity analyses, choices of covariates, and analyses by subsample. In this sense, our indices bridge between the frequentist and Bayesian interpretations because our indices ask the frequentist to consider alternative inferences for different samples. Ultimately, the comparability of the expressions and values in the empirical example suggests that the purely Bayesian and frequentist approaches will generate similar impressions of the robustness of inferences.

9. COMPARISON WITH OTHER PROCEDURES

9.1. *Cross-Validation Studies*

Our approach is like that of cross-validation indices (e.g., Cudeck and Browne 1983) in that we consider two separate samples. But both samples are observed in the construction of cross-validation indices (in fact the cases are often randomly separated), with the cross-validation index calculated by assessing the fit in one sample based on the parameter estimates from another. In contrast, the supplemental sample

[14]This ignores applications of multilevel models that can be interpreted as "empirical Bayes" but also can be interpreted from weighted least squares or generalized least squares perspectives (Raudenbush and Bryk 2002) and in which the authors interpreted *p*-values from a frequentist perspective.

necessary to construct the ideal estimate is unobserved. Thus a cross-validation index indicates a particular model is good relative to others if the model fits well in the observed cross-validation sample, whereas for us an inference is robust if the hypothetical unobserved sample would have to be considerably different from the observed sample to alter a statistical inference.

9.2. *Breakdown Points*

Our indices are similar to those for defining the breakdown point of an estimator in that they consider the statistical effects of altered samples (indeed our language of contamination and replacement is that of Donoho and Huber [1983]). Breakdown points refer to the properties of estimators (e.g., the least squares or maximum likelihood estimates of β_1) and are defined by the smallest amount of contamination that may cause an estimator to take on arbitrarily large aberrant values (Donoho and Huber 1983:157). Thus, for example, the breakdown point for the least squares estimate of β_1 is one, because one extreme observation can infinitely alter an estimate. But our indices differ from breakdown points because they apply to an inference for a specific sample, instead of to the estimator, independent of a sample. For example, we report the $IEVR(\pi, r_{xy}^{un} = 0)$ for the Finn and Achilles (1990) class size effect as 64 percent, while $IEVR(\pi, r_{xy}^{un} = 0)$ could be smaller or larger for another study, but the breakdown point for the least squares estimate, like all other least squares estimates, is one.

9.3. *Extension of Fail-Safe* n

We note that calculations of the fail-safe *n* in meta-analysis have been extended to characterizations of the likely underlying sampling distribution of effect sizes. For example, trim and fill procedures (e.g., Duval and Tweedie 2000) use funnel plots to examine evidence of publication bias under the assumption that the distribution of effect sizes is symmetric. Thus if one tail appears censored, the procedure trims from the other tail and fills in the censored tail until the distribution appears symmetric. Following this approach, we could consider indices based on replacing those observations that have the largest residuals from the overall trend. This is a refinement of our $IEVR(\pi, r_{xy}^{un} = 0)$ in which

we focus on replacing cases with correlation with exactly r_{xy}^{ob}. Importantly, such focus on individual data points may be more defensible in the meta-analytic context where each point represents many cases and is thus measured with higher precision than in the typical regression analysis in which each point represents only a single observation. Thus we leave consideration of replacement of specific data points to further research on the relationship between statistical influence and statistical inference.

9.4. *Approaches for Missing Data*

Our approach may be compared with others that have been applied to missing data, such as maximum likelihood estimation or multiple imputation (e.g., Allison 2000; D'Agostino and Rubin 2000; Daniels and Hogan 2000; Dehejia and Wahba 1999; Little 1992; see the review by Collins, Shafer, and Kam 2001). These approaches seek to improve point estimates and confidence intervals by modeling the pattern of missing data. But our focus is on cases for which all variables are missing—the cases are purely hypothetical. And clearly such hypothetical data cannot alter estimates and inferences since they contain no information. Ultimately, our indices complement the use of other missing data procedures, as we can apply our indices to quantify robustness after using other missing data procedures to improve estimation.

9.5. *Comparison with Econometric Forecasting*

Our characterization of broad external validity corresponds with Heckman (2005) and the emphasis by Manski (1995, chap. 1) on the importance of forecasting effects of new treatments or in new populations. Drawing on Marschak (1953), Heckman emphasizes that econometric analyses allow forecasting of results better than the Rubin (1974)/Holland (1986) causal model (which is based on matching or randomized experiments but applies to the general linear model) because the econometric approach takes into account how and why members of different populations might choose different treatments. Thus effects in a new population are generalized from evidence in the sample most representative of that population, thereby better accounting for the likely choices made by members of the population as well as the resulting treatment

effects. In this light, our indices, quantified in terms of the general linear model, extend the conceptualization of the Rubin/Holland causal model toward a forecasting function because they allow policymakers and researchers to consider how different a population must be from the population studied such that the inferences from the observed data are invalid for forecasting to that population.

10. DISCUSSION

In the behavioral and social sciences, we can be certain of external validity only when observations are randomly sampled or when data are missing completely at random (Little and Rubin 1987). But social scientists rarely analyze perfectly random samples (e.g., without attrition). Correspondingly, critics in the social sciences may challenge the generality of inferences whenever there is uncertainty as to how representative a sample is of some target population.

Of course, the first response to such concerns should be to include all relevant subpopulations in a sample and compare observed relationships, testing for interaction effects (Cronbach and Snow 1977). Where interaction effects are detected, different estimates of the effects in each subpopulation would be reported. In our example, Finn and Achilles (1990) followed this procedure, testing for interactions of class size with location of the school, grade, and predominant race in the school (and Nye et al. [2002] tested for interactions by prior achievement). These basic methods may be extended by more elaborate techniques such as the exploration of treatment effects by strata of propensity scores (e.g., Rosenbaum and Rubin 1983; Morgan 2001).

But sometimes it is not possible to obtain data for all relevant subpopulations. For example, the Tennessee class size study did not report results based on per pupil expenditures and teacher salaries. Nor were there funds or political motivation to include schools from other states, nor did the research span over decades to the current date.

Furthermore, even inferences made from some subsamples could be challenged. For example, even inferences made from the Tennessee study for minorities may not apply to minorities across Tennessee in the 1980s or to minorities in other states or other time periods. Thus, even when analyses are broken down by subsample there may still be the concern that an inference from an observed subsample would differ from

that of a perfectly representative subsample (Birnbaum and Mellers 1989; Cronbach and Snow 1977; Greenland 2000). In the extreme, accepting inference only if there are no interactions can lead to an infinite reduction to effects for single units, which requires the impossible counterfactual data (Holland 1986).

To inform inevitable debates regarding the external validity of an inference, we have developed our indices to quantify how much of a departure from a perfectly representative sample is required to invalidate an inference. Regarding the inference from the Tennessee class size study that small classes improve achievement, the index of external validity for $r_{xy}^{un} = 0$ [$IEVR(\pi, r_{xy}^{un} = 0)$] indicated that 64 percent or more of the volunteer schools would have to have been replaced with a sample for which $r_{xy}^{un} = \rho_{xy}^{un} = 0$ to invalidate the inference. Note that the $IEVR(\pi, r_{xy}^{un} = 0)$ of 64 percent can be compared with Finn and Achilles' (1990:559) sampling percentage of about 33 percent, indicating that if the nil hypothesis holds for the unobserved schools, the inference from the observed data would be invalid.

As a complement to the $IEVR(\pi, r_{xy}^{un} = 0)$, the $IEVR(\pi = .5, r_{xy}^{un})$ indicated that if half the sample were replaced, r_{xy}^{un} would have to have been less than $-.08$ to invalidate the inference from the observed data. As a basis of comparison, the estimated effects of small classes were uniformly positive across categories of urbanacity (Finn and Achilles 1991, table 3), levels of prior achievement (Nye et al. 2002, table 1), and samples with different attrition patterns (Hanushek 1999, table 5). Thus the requirement that small classes would have to have a negative effect in the unobserved population to invalidate the inference is extreme when compared to the range of estimated effects. Furthermore, the unobserved estimate of $r_{xy}^{un} = -.08$ would have to be significantly different statistically from the observed estimated of $r_{xy}^{ob} = .296$ to invalidate the inference, suggesting that the inference is robust in an absolute sense defined by statistical inference; either the overall inference would not change if half the cases were replaced, or the inference would change only if r_{xy}^{un} were significantly different from r_{xy}^{ob}, implying that the original inference applies at least to a discernible subpopulation.[15]

[15]Solving (4) for $r_{xy}^{un} : r_{xy}^{un} < \frac{r^{\#} - (1-\pi) r_{xy}^{ob}}{\pi}$ can also be used to assess the robustness of an inference with respect to bias induced by attrition. For example, if there were 20 percent attrition in the Tennessee class size study, r_{xy}^{un} would have to be less than $-.70$ for the inference based on the observed data to be invalid.

Our indices can be interpreted either in a narrow or broad sense of external validity. In the narrowest sense, we may consider generalizing Finn and Achilles' (1990) results to elementary schools in Tennessee in the mid-1980s, drawing on the high surface similarity in terms of time and place and some background characteristics between those schools that did and did not volunteer for the Tennessee class size study. Similarly, it may be relatively straightforward to rule out many likely factors differentiating volunteer from nonvolunteer populations that would render 64 percent of the observed data as nonrepresentative. On the other hand, it may well be that less than 64 percent of the sample would contribute to a sample that is representative of schools in different places and at different times (e.g., California in the mid-1990s).

Because the indices are developed with respect to the general linear model, they can be applied to other analyses that ultimately employ the general linear model or variations of it. For example, we could apply the indices to analyses based on propensity scores used to focus on specific treatment effects but where there still may be concerns regarding external validity (e.g., Morgan 2001). Similarly we can quantify the robustness of inferences from meta-analyses with respect to generalizing to other populations or into the future (Worm et al. 2006).

In contrast to the classic approach to experimental design (e.g., Fisher 1924), our logic is decidedly *post hoc*, imploring researchers to consider how results might have been affected by an alternative composition of the sample. Furthermore, our analysis is in terms of common procedures associated with the general linear model. This distinguishes our approach from approaches based on nonparametric statistics (e.g., Rosenbaum 2002), nonlinear relationships (e.g., Manski 1990), or the fail-safe *n* problem in meta-analysis (Rosenthal 1979), while the hypothetical nature of our framework distinguishes our approach from procedures that draw on observed characteristics of the missing cases (e.g., cross-validation studies and multiple imputation).

10.1. *Quantitative Thresholds and Decision-Making*

Recognizing that statistical significance is not the only criterion for making a causal inference, we developed our indices for any quantitative threshold. Most generally, 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 for a sample of different composition.

Although we have recognized alternative thresholds, some may still be uneasy with using statistical inference as one basis for causal inference (e.g., Hunter 1997). In making a causal inference, we should rely on effect sizes, confidence intervals, causal mechanisms, and the nuances of implementation (Wilkinson et al. 1999). 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. Consider 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" (p. 25). Therefore statistical significance is treated as an essential condition for causal inference, and thus it is reasonable to define thresholds for robustness in terms of statistical inference.

Ultimately, what are the key objections to moving from statistical analysis to causal inference? First, the observed relationship may be spurious because there may be an omitted confounding variable (correlation does not equal causation [Holland 1986]). This concern is quantified via Rosenbaum's (2002) index of selection bias and Frank's (2000) index of robustness to the impact of a confounding variable. Or the effect may vary across contexts (Gilbert and Mosteller 1972; Winship and Sobel 2004). This is the focus of the indices developed here. Drawing on Holland (1986), we see that the combination of existing indices for spurious relationships (Rosenbaum 2002; Frank 2000) and the indices presented here for representation of a sample quantify the primary concerns in moving from statistical analysis to causal inference. If the key statistical thresholds are unlikely to be exceeded when confounding variables are included or alternative samples are used, then the statistical, and thus the causal, inferences are robust.

10.2. *Limitations and Extensions*

We wish to emphasize the *post hoc* nature of our interpretation of the indices. The indices quantify what *would have happened if* the sample had been more representative of a target population. This informs the question of construct validity—of evidence of an underlying mechanism that operates across contexts (Cook and Campbell 1979). We recognize that the indices are less informative for the *adaptation* of treatments to

alternative contexts. For example, though small classes appeared to have some small effects in California, *implementation* of small classes resulted in the hiring of higher percentages of unqualified teachers, especially for students who were most disadvantaged (Bohrnstedt and Stecher 2002; Hanushek 1999). Thus issues of implementation must be considered even if causal inference is robust.

We note that our indices are limited to applications of the general linear model. Using this model has the advantage that we can calculate how unobserved quantities affect parameter estimates and statistical inference in closed form. But we anticipate great value in extending the indices to a broader set of models. For example, Harding (2003) extended Frank's index for confounding variables to logistic regression. Furthermore, because we developed our indices by drawing on the functional relationship between a t-ratio and a correlation coefficient (as in equation [4]), we could theoretically extend the indices to any statistical procedure that reports t-ratios—for example, to multilevel models that correct standard errors and estimates for the nesting of observations (Raudenbush and Bryk 2002; Seltzer, Kim, and Frank 2006).

11. CONCLUSION

In spite of the value of robustness indices, it is worth emphasizing the robustness indices do *not* sustain new causal inferences. In the example, the original inference that small classes improve achievement was not modified. Nor do the robustness indices replace the need for improved research designs or better theories. If we accept that causal inferences are to be debated (Abbott 1998), what robustness indices do is quantify the terms of the debate. Therefore instead of "abandoning the use of causal language" (Sobel 1998: 345, see also Sobel 1996:p 355) we quantify the robustness of an inference and interpret it relative to the design of a study.

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

APPENDIX A: A QUICK GUIDE FOR CALCULATING ROBUSTNESS INDICES FOR SAMPLE REPRESENTATION

Known quantities:

r_{xy}^{ob}: the correlation between the treatment (x) and the outcome (y) in the observed sample (.296 in the example);
n^{ob}: the observed sample size (331 in the example);
$r^{\#}$: the threshold for a sample correlation for making an inference ($r^{\#}$ for statistical significance in the example = .107, as obtained from equation [6] in the main text);
t_{critical}: the critical value of a t-distribution used for inference (1.96 in the example), and q is the number of parameters estimated (including the intercept, the parameter for x and parameters for any other covariates, = 2 in the example).

Unknown quantities necessary to construct the ideal sample:

r_{xy}^{un}: the correlation between the treatment (x) and the outcome (y) in the unobserved sample;
π: the proportion of the ideal sample that is constituted by the unobserved cases.

The general expression for the relationship of interest in an ideal sample is (equation [3] in the main text):

$$r_{xy}^{ideal} = (1 - \pi) r_{xy}^{ob} + \pi r_{xy}^{un}.$$

To determine a robustness index, set $r_{xy}^{ideal} \leq \mathrm{r}^{\#}$, set one of the unknown quantities to a desired value, and solve for the other unknown quantity.

For Neutralization by Replacement (Replacing a Portion of a Sample)

Q. Assuming there is no effect in the unobserved sample ($r_{xy}^{un} = 0$), what proportion of the original sample (π) must be replaced to invalidate the inference that small classes have an effect on achievement?

The *i*ndex of external validity for $r_{xy}^{un} = 0$ and *r*eplacement, or $IEVR(\pi, r_{xy}^{un} = 0) = \mathrm{r}^{\#}/r_{xy}^{ob}$ (= .64 in the example).

A. Assuming $r_{xy}^{un} = 0$, replace at least 64 percent of the sample ($\pi \geq .64$) to invalidate the inference that small classes have an effect on achievement.

Q. Assuming half the sample were replaced ($\pi = .5$), what must be the effect in the unobserved sample (r_{xy}^{un}) to invalidate the inference that small classes have an effect on achievement?

The index of external validity for $\pi = .5$ and *r*eplacement, or *IEVR* ($\pi = .5, r_{xy}^{un}$) = $2r^{\#} - r_{xy}^{ob}$ (= −.08 in the example).

A. If half the sample were replaced, r_{xy}^{un} must be less than or equal to −.08 to invalidate the inference that small classes have an effect on achievement.

For Neutralization by Contamination (Adding to a Sample)

The index of external *v*alidity for $r_{xy}^{un} = 0$ and *c*ontamination, or $IEVC(\pi, r_{xy}^{un} = 0)$, is the same as $IEVR(\pi, r_{xy}^{un} = 0)$, but it is based on $\pi = n^{un*}/(n^{ob} + n^{un*})$, where

$$n^{un*} = \frac{n^{ob}\left[n^{ob}\left(r_{xy}^{ob}\right)^2 - 2t_{\text{critical}}^2 + r_{xy}^{ob}\sqrt{\left(n^{ob}\right)^2\left(r_{xy}^{ob}\right)^2 + 4t_{\text{critical}}^2\left(t_{\text{critical}}^2 - q\right)}\right]}{2t_{\text{critical}}^2}.$$

In the example, $n^{un*} = 2153$, which would account for 87 percent of the cases in the ideal sample.

The index of external validity for $\pi = .5$ and contamination, or $IEVC(\pi = .5, r_{xy}^{un}) = 2r^{\#\#} - r_{xy}^{ob}$, where

$$r^{\#\#} = \frac{t_{\text{critical}}}{\sqrt{(2n^{ob} - q) + t_{\text{critical}}^2}}.$$

In the example, $IEVC(\pi = .5, r_{xy}^{un}) = -.14$, indicating r_{xy}^{un} would have to be less than or equal to −.14 to invalidate the inference if the sample size were doubled by adding cases from the potentially unobserved population.

Whether the Difference Between r_{xy}^{ob} and r_{xy}^{un} Would Be Statistically Significant

The index of external validity beyond interaction (*IEVBI*) is equal to

$$\frac{\mathrm{r}^{\#} - e^{2m}(1+\mathrm{r}^{\#}) - 1 - \sqrt{\begin{array}{l}4(e^{2m}-1)(\pi-1)(e^{2m}\pi+\mathrm{r}^{\#}+e^{2m}\mathrm{r}^{\#}-\pi) \\ +(1-\mathrm{r}^{\#}+e^{2m}(1+\mathrm{r}^{\#}))^2\end{array}}}{2(e^{2m}-1)(\pi-1)},$$

where $m = 1.96\sqrt{\dfrac{1}{(1-\pi)(\mathrm{n^{ob}}-\mathrm{q})} + \dfrac{1}{\pi(\mathrm{n^{ob}}-\mathrm{q})}}$.

In the example, assuming $\pi = .50$, the *IEVBI* = .215 which is less than r_{xy}^{ob} of .296. This indicates that r_{xy}^{un} would have to be significantly different statistically from r_{xy}^{ob} to invalidate the inference.

See http://www.msu.edu/~kenfrank/research.htm#causal for a spreadsheet that calculates some of these indices.

APPENDIX B: FULL EXPRESSION FOR r_{xy}

We develop here an expression for r_{xy} that does not assume that the means and variances of the potentially unobserved population equal those of the potentially observed population. First, we calculate a covariance constructed from two samples with different means and variances. Symbols are defined as follows.

1. n = sample size
r_{xy} = correlation
x_i = X component
y_i = Y component
$\bar{x}$ = mean of X
$\bar{y}$ = mean of Y
s_{xy} = covariance of X and Y
s_x = standard deviation of X
s_y = standard deviation of Y for the representative sample that is a combination of potentially observed and potentially unobserved samples
2. Corresponding statistics for the potentially observed sample are n^{ob}, r_{xy}^{ob}, x_i^{ob}, y_i^{ob}, $\bar{x}^{ob}$, $\bar{y}^{ob}$, s_{xy}^{ob}, s_x^{ob}, and s_y^{ob}.

3. Corresponding statistics for the potentially unobserved sample are n^{un}, r_{xy}^{un}, x_i^{un}, y_i^{un}, $\bar{x}^{un}$, $\bar{y}^{un}$, s_{xy}^{un}, s_x^{un}, and s_y^{un}.

To begin,

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \frac{1}{n}\left[\sum_{i=1}^{n^{ob}}\left(x_i^{ob} - \bar{x}\right)\left(y_i^{ob} - \bar{y}\right) + \sum_{i=1}^{n^{un}}\left(x_i^{un} - \bar{x}\right)\left(y_i^{un} - \bar{y}\right)\right]. \tag{B-1}$$

Define $\pi = \frac{n^{un}}{n}$ and $(1-\pi) = \frac{n^{ob}}{n}$. Allowing for different means between the observed and unobserved samples implies

$$\sum_{i=1}^{n^{ob}}\left(x_i^{ob} - \bar{x}\right)\left(y_i^{ob} - \bar{y}\right) = \sum_{i=1}^{n^{ob}}\left[\left(x_i^{ob} - (1-\pi)\bar{x}^{ob} - \pi\,\bar{x}^{un}\right)\left(y_i^{ob} - (1-\pi)\bar{y}^{ob} - \pi\,\bar{y}^{un}\right)\right]. \tag{B-2}$$

Using the identity (A-B-C)(D-E-F) = (A-B)(D-E)-F(A-B)-C(D-E-F), the right side of the above equation equals:

$$\sum_{i=1}^{n^{ob}}\left[\left(x_i^{ob} - (1-\pi)\bar{x}^{ob}\right)\left(y_i^{ob} - (1-\pi)\bar{y}^{ob}\right)\right] + \sum_{i=1}^{n^{ob}}\left\{\left(-\pi\,\bar{y}^{un}\left[x_i^{ob} - (1-\pi)\bar{x}^{ob}\right] - \pi\,\bar{x}^{un}\left[y_i^{ob} - (1-\pi)\bar{y}^{ob} - \pi\,\bar{y}^{un}\right]\right\}. \tag{B-3}$$

The above equation can then be decomposed into $W^{ob} + Q^{ob}$, where

$$W^{ob} = \sum_{i=1}^{n^{ob}}\left[\left(x_i^{ob} - (1-\pi)\bar{x}^{ob}\right)\left(y_i^{ob} - (1-\pi)\bar{y}^{ob}\right)\right] \quad \text{and}$$
$$Q^{ob} = \sum_{i=1}^{n^{ob}}\left\{-\pi\,\bar{y}^{un}\left[x_i^{ob} - (1-\pi)\bar{x}^{ob}\right] - \pi\,\bar{x}^{un}\left[y_i^{ob} - (1-\pi)\bar{y}^{ob} - \pi\,\bar{y}^{un}\right]\right\}. \tag{B-4}$$

Now, the *observed* sample covariance can itself be decomposed:

$$\begin{aligned} n^{ob}s_{xy}^{ob} &= \textstyle\sum_{i=1}^{n^{ob}} \left(x_i^{ob} - \bar{x}^{ob}\right)\left(y_i^{ob} - \bar{y}^{ob}\right) \\ &= \textstyle\sum_{i=1}^{n^{ob}} \left[x_i^{ob} - (1-\pi)\bar{x}^{ob} - \pi\bar{x}^{ob}\right] \\ &\qquad \times \left[y_i^{ob} - (1-\pi)y_i^{ob} - \pi y_i^{ob}\right] \quad . \end{aligned} \tag{B-5}$$

Again using the identity (A-B-C)(D-E-F) = (A-B)(D-E)-F(A-B)-C(D-E-F), the right side of the above equation equals

$$\begin{aligned} &\textstyle\sum_{i=1}^{n^{ob}} \left\{\left[x_i^{ob} - (1-\pi)\bar{x}^{ob}\right]\left[y_i^{ob} - (1-\pi)\bar{y}^{ob}\right]\right\} \\ &+ \textstyle\sum_{i=1}^{n^{ob}} \left\{-\pi\bar{y}^{ob}\left[x_i^{ob} - (1-\pi)\bar{x}^{ob}\right] - \pi\bar{x}^{ob}\left[y_i^{ob} - (1-\pi)\bar{y}^{ob} - \pi\bar{y}^{ob}\right]\right\}. \end{aligned} \tag{B-6}$$

The above equation equals $W^{ob} + Z^{ob}$, where

$$\begin{aligned} &Z^{ob} = \\ &\textstyle\sum_{i=1}^{n^{ob}} \left\{-\pi\bar{y}^{ob}\left[x_i^{ob} - (1-\pi)\bar{x}^{ob}\right] - \pi\bar{x}^{ob}\left[y_i^{ob} - (1-\pi)\bar{y}^{ob} - \pi\bar{y}^{ob}\right]\right\}. \end{aligned} \tag{B-7}$$

Substituting $n^{ob}s_{xy}^{ob} - Z^{ob}$ for W^{ob} above,

$$\sum_{i=1}^{n^{ob}} \left(x_i^{ob} - \bar{x}\right)\left(y_i^{ob} - \bar{y}\right) = n^{ob}s_{xy}^{ob} - Z^{ob} + Q^{ob}. \tag{B-8}$$

Similarly,

$$\sum_{i=1}^{n^{mi}} \left(x_i^{un} - \bar{x}\right)\left(y_i^{un} - \bar{y}\right) = n^{un}s_{xy}^{un} - Z^{un} + Q^{un}, \tag{B-9}$$

where

$$
\begin{aligned}
Q^{un} &= \sum_{i=1}^{n^{un}} \left\{ -(1-\pi)\bar{y}^{ob}\left(x_i^{un} - \pi\bar{x}^{un}\right) \right. \\
&\quad \left. -(1-\pi)\bar{x}^{ob}\left[y_i^{un} - (1-\pi)\bar{y}^{ob} - \pi\bar{y}^{un}\right]\right\} \quad \text{and} \\
Z^{un} &= \textstyle\sum_{i=1}^{n^{un}} \left\{ -(1-\pi)\bar{y}^{un}\left(x_i^{ob} - \pi\right)\bar{x}^{ob}\right) \\
&\quad -(1-\pi)\bar{x}^{un}\left[y_i^{un} - (1-\pi)\bar{y}^{un} - \pi\bar{y}^{un}\right]\right\}.
\end{aligned}
\tag{B-10}
$$

Thus

$$
\begin{aligned}
s_{xy} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{1}{n}\left(n^{ob}s_{xy}^{ob} - Z^{ob} + Q^{ob} + n^{un}s_{xy}^{un} - Z^{un} + Q^{un}\right).
\end{aligned}
\tag{B-11}
$$

Now, $-(Z^{ob} + Z^{un}) + (Q^{ob} + Q^{un})$
$= n(1-\pi)\pi(\bar{x}^{ob}\bar{y}^{ob} + \bar{x}^{un}\bar{y}^{un} - \bar{x}^{ob}\bar{y}^{un} - \bar{x}^{un}\bar{y}^{ob})$.

Therefore, $s_{xy} = (1-\pi)s_{xy}^{ob} + \pi s_{xy}^{un} + (1-\pi)\pi(\bar{x}^{ob} - \bar{x}^{un})(\bar{y}^{ob} - \bar{y}^{un})$.

By similar calculations, $s_x^2 = (1-\pi)\left(s_x^{ob}\right)^2 + \pi\left(s_x^{un}\right)^2$
$+(1-\pi)\pi(\bar{x}^{ob} - \bar{x}^{un})^2$
and $s_y^2 = (1-\pi)\left(s_y^{ob}\right)^2 + \pi\left(s_y^{un}\right)^2 + (1-\pi)\pi\left(\bar{y}^{ob} - \bar{y}^{un}\right)^2$. (B-12)

The overall expression for the combined correlation (expressed in terms of observed and unobserved correlations) is then

$$
r_{xy} = \frac{(1-\pi)r_{xy}^{ob}s_x^{ob}s_y^{ob} + \pi r_{xy}^{un}s_x^{un}s_y^{un} + (1-\pi)\pi(\bar{x}^{ob} - \bar{x}^{un})(\bar{y}^{ob} - \bar{y}^{un})}{\sqrt{\begin{array}{l}[(1-\pi)\left(s_x^{ob}\right)^2 + \pi\left(s_x^{un}\right)^2 + (1-\pi)\pi(\bar{x}^{ob} - \bar{x}^{un})^2] \\ \times[(1-\pi)(s_y^{ob})^2 + \pi(s_y^{un})^2 + (1-\pi)\pi(\bar{y}^{ob} - \bar{y}^{un})^2]\end{array}}}.
\tag{B-13}
$$

If $s_x = s_x^{ob} = s_x^{un}$ and $s_y = s_y^{ob} = s_y^{un}$ then

$$r_{xy} = \frac{s_x s_y\left[(1-\pi)r_{xy}^{ob} + \pi r_{xy}^{un}\right] + (1-\pi)\pi(\bar{x}^{ob} - \bar{x}^{un})(\bar{y}^{ob} - \bar{y}^{un})}{\sqrt{\left[(1-\pi)\pi(\bar{x}^{ob} - \bar{x}^{un})^2 + s_x^2\right]\left[(1-\pi)\pi(\bar{y}^{ob} - \bar{y}^{un})^2 + s_y^2\right]}}. \tag{B-14}$$

If $\bar{x}^{ob} = \bar{x}^{un}$ and $\bar{y}^{ob} = \bar{y}^{un}$ then

$$r_{xy} = \frac{(1-\pi)r_{xy}^{ob}s_x^{ob}s_y^{ob} + \pi r_{xy}^{un}s_x^{un}s_y^{un}}{\sqrt{\begin{array}{l}\left\{\left(s_x^{un}\right)^2 + (1-\pi)\left[\left(s_x^{ob}\right)^2 - \left(s_x^{un}\right)^2\right]\right\} \\ \times\left\{\left(s_y^{un}\right)^2 + (1-\pi)\left[\left(s_y^{ob}\right)^2 - \left(s_y^{un}\right)^2\right]\right\}\end{array}}}. \tag{B-15}$$

If both means and variances are equal, we get $r_{xy} = (1-\pi)r_{xy}^{ob} + \pi r_{xy}^{un}$ as in equation (3) in the main text of this paper.

REFERENCES

Abbott, A. 1998. "The Causal Devolution." *Sociological Methods and Research* 27:148–81.

Allison, P. D. 2000. "Multiple Imputation for Missing Data." *Sociological Methods and Research* 28:301–9.

Birnbaum, M. H., and B. A. Mellers. 1989. "Mediated Models for the Analysis of Confounded Variables and Self-Selected Samples." *Journal of Educational Statistics* 14:121–40.

Bohrnstedt, G. W., and B. M. Stecher. 2002. *What We Have Learned About Class Size Reduction in California*. Sacramento, CA: California Department of Education.

Cohen, J., and P. Cohen. 1983. *Applied Multiple Regression/ Correlation Analysis for Behavioral Science*. Hillsdale, NJ: Lawrence Erlbaum.

Cole, W. M. 2005. "International Human Rights Covenants, 1966–1999." *American Sociological Review* 70(3):472–95.

Collins, L. M., J. Shafer, and C. Kam. 2001. "A Comparison of Restrictive Strategies in Modern Missing Data Procedures." (special issue) *Psychological Methods* 6:330–51.

Cook, T. D. 2002. "Randomized Experiments in Education: Why Are They So Rare?" Working Paper No. 02-19, Institute for Policy Research, Northwestern University.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago, IL: Rand McNally.

Copas, J., and H.G. Li. 1997. "Inference for Non-random Samples." *Journal of the Royal Statistical Society B* 59:55–96.

Cornfield, J., and J. W. Tukey. 1956. "Average Values of Mean Squares in Factorials." *Annals of Mathematical Statistics* 27:907–49.

Cronbach, L. J. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey Bass.

Cronbach, L. J., and R. E. Snow. 1977. *Aptitudes and Instructional Methods: A Handbook of Research on Interactions*. New York: Irvington.

CSR Research Consortium. 2000. "Class Size Reduction in California: The 1998–99 Evaluation Findings." (http://www.classize.org/summary/98-99/).

Cudeck, R., and M. Browne. 1983. "Cross-Validation of Covariance Structures." *Multivariate Behavioral Research* 18:147–67.

D'Agostino, R. B., Jr., and D. B. Rubin. 2000. "Estimating and Using Propensity Scores with Partially Missing Data." *Journal of the American Statistical Association* 95:749–59.

Daniels, M. J., and J. W. Hogan. 2000. "Reparameterizing the Pattern Mixture Model for Sensitivity Analyses Under Informative Dropout." *Biometrics* 56:1241–48.

Day, N. E. 1969. "Estimating the Components of a Mixture of Normal Distributions." *Biometrika* 56(3):463–74.

Dehejia, R. H., and S. Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94:1053–62.

Diprete, T., and M. Gangl. 2004. "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments." Pp. 271–310 in *Sociological Methodology*, vol. 34, edited by Ross M. Stolzenberg. Boston, MA: Blackwell Publishing.

Donoho, D. L., and P. J. Huber. 1983. "The Notion of Breakdown Point." Pp. 157–84 in *A Festschrift for Erich L. Lehman*, edited by P. J. Bickel, K. Doksun, and J. L. Hodge, Jr. Belmont, CA: Wadsworth International Group.

Duval, S., and R. Tweedie. 2000. "Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis." *Biometrics* 56:455–63.

Ed Source on Line. 2003. *Rankings and Estimates*. National Education Association.

Finn, J. D., and C. M. Achilles. 1990. "Answers and Questions About Class Size: A Statewide Experiment." *American Educational Research Journal* 27:557–77.

Fisher, R. A. 1924. "The Distribution of the Partial Correlation Coefficient." *Metron* 3:329–32.

———. 1973. *Statistical Methods for Research Workers*. New York: Hafner.

Frank, K. A. 2000. "Impact of a Confounding Variable on a Regression Coefficient." *Sociological Methods and Research* 29:147–94.

Gilbert, J. P., and F. Mosteller. 1972. "The Urgent Need for Experimentation." In *On Equality of Educational Opportunity*, edited by F. Mosteller and D. P. Moynihan. New York: Random House.

Gill, R. D., and J. M. Robins. 2001. "Causal Inference for Complex Longitudinal Data: The Continuous Case." *Annals of Statistics* 29(6):1785–811.
Greenland, S. 2000. "An Introduction to Instrumental Variables for Epidemiologists." *International Journal of Epidemiology* 29:722–29.
Hanushek, E. 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* 21:143–63.
Harding, D. 2003. "Counterfactual Models of Neighborhood Effect: The Effect of Neighborhood Poverty and High School Dropout on Teenage Pregnancy." *American Journal of Sociology* 109:676–719.
Heckman, J. J. 2005. "The Scientific Model of Causality." Pp. 1–97 in *Sociological Methodology*, vol. 35, edited by Ross M. Stolzenberg. Boston, MA: Blackwell Publishing.
Hedges, L. 1992. "Modeling Selection Effects in Meta-Analysis." *Statistical Science* 7:246–55.
Holland, P. W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–70.
Hunter, J. E. 1997. "Needed: A Ban on the Significance Test." *American Psychological Society* 8(1):3–7.
Karinek, K., B. Entwisle, and A. Jampaklay. 2005. "Social Ties and Urban Settlement among Tahi Immigrants." *American Sociological Review* 70(5):779–800.
Korinek, K., B. Entwisle, and A. Jampaklay. 2005. "Through Thick and Thin Layers of Social Ties and Urban Settlement Among Thai Migrants." *American Sociological Review* 70(5):779–800.
Lee, P. M. 1989. *Bayesian Statistics: An Introduction*. New York: Oxford University Press.
Lindsay, B. G., and P. Basak. 1993. "Multivariate Normal Mixtures: A Fast Consistent Method of Moments." *Journal of the American Statistical Association* (88):468–76.
Little, R. J. A. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87:1227–37.
Little, R. J. A., and D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: J. Wiley.
Mallard, G. 2005. "Interpreters of the Literary Canon and Their Technical Instruments: The Case of Balzac Criticism." *American Sociological Review* 70(6):992–1010.
Manski, C. F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review Papers and Proceedings* 80:319–23.
Manski, C. F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.
Marschak, J. 1953. "Econometric Measurements for Policy and Prediction." Pp. 1–26 in *Studies in Econometric Method*, edited by W. Hood and T. Koopmans. New York: Wiley.
McLachlan, G., and D. Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Min, K.-S., and K. A. Frank. 2002. "The Impact of Nonignorable Missing Data on the Inference of Regression Coefficients." Presented at the annual meeting of Mid-Western Educational Research Association, October 16, Columbus, OH.

Morgan, S. L. 2001. "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning," *Sociology of Education* 74:341–74.

National Center for Educational Statistics (NCES). 2003. http://nces.ed.gov/pubs2003/hispanics/Section1.asp.

Nye, B., L.V. Hedges, and S. Konstantopoulos. 2000. "Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment." *American Educational Research Journal* 37:123–51.

———. 2002. "Do Low-Achieving Students Benefit More from Small Classes? Evidence from the Tennessee Class Size Experiment." *Education Evaluation and Policy Analysis* 24(3):201–17.

Orwin, R. G. 1983. "A Fail-Safe N for Effect Size in Meta-Analysis." *Journal of Educational and Behavioral Statistics* 8:157–59.

Pan, W., and K. A. Frank. 2004. "A Probability Index of the Robustness of a Causal Inference." *Journal of Educational and Behavioral Statistics* 28:315–37.

Raudenbush, S. W., and A. S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Newbury Park, CA: Sage.

Robins, J. 1987. "A Graphical Approach to the Identification and Estimation of Causal Parameters in Mortality Studies with Sustained Exposure Periods." *Journal of Chronic Diseases* 40(2):1395–1615.

Robins J. M. 1989. "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies." Pp. 113–159 in *Health Services Research Methodology: A Focus on AIDS*, edited by L. Sechrest, H. Freeman, and A. Mulley. Rockville, MD: U.S. Department of Health and Human Services, National Center for Health Services Research and Health Care Technology Assessment.

Robins, J., A. Rotnisky, and D. Scharfstein. 2000. "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models." Pp. 1–95 in *Statistical Models in Epidemiology, the Environment and Clinical Trials (The IMA Volumes in Mathematics and Its Applications)*, edited by E. Halloran and D. Berry. New York: Springer-Verlag.

Rosenbaum, P. R. 1986. "Dropping Out of High School in the United States: An Observational Study." *Journal of Educational Statistics* 11(3):207–24.

———. 2002. "Attribution Effects to Treatment in Matched Observational Studies." *Journal of the American Statistical Association* 97:183–92.

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.

Rosenthal, R. 1979. "The "File-Drawer Problem" a Tolerance for Null Results." *Psychological Bulletin* 86:638–41.

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

Scharfstein, D. A. I. 2002. "Generalized Additive Selection Models for the Analysis of Studies with Potentially Nonignorable Missing Data." *Biometrics* 59:601–13.

Seltzer, M. H., J. Kim, and K. A. Frank. 2006. "Studying the Sensitivity of Inferences to Possible Unmeasured Confounding Variables in Multisite Evaluations." Paper presented at the American Educational Researcher Association, San Francisco, CA.

Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Shadish, W. R., T. D. Cook, and L. C. Leviton. 1991. *Foundations of Programme Evaluation: Theories of Practice*. London: Sage.

Sobel, M. E. 1995. "Causal Inference in the Social and Behavioral Sciences." Pp. 1–38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C. C. Clogg, and M. E. Sobel. New York: Plenum Press.

———. 1996. "An Introduction to Causal Interface." *Sociological Methods and Research* 24(3):353–79.

———. 1998. "Causal Inference in Statistical Models of the Process of Socioeconomic Achievement: A Case Study." *Sociological Methods and Research* 27:318–48.

Suppes, P. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Tennessee State Department of Education. 1990. "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project." Retrieved April 17, 2007. (http://www.heros-inc.org/summary.pdf).

U.S. Department of Education. 2002. *Report on Scientifically Based Research Supported by U.S. Department of Education*. Retrieved April 17, 2007 (http://www.excelgov.org/displayContent.asp?Keyword=prppcEvidence).

Wainer, H., and D. H. Robinson. 2003. "Shaping Up the Practice of Null Hypothesis Significance Testing." *Educational Researcher* 32:22–30.

Wilkinson, L., and Task Force on Statistical Inference. 1999. "Statistical Methods in Psychology Journals: Guidelines and Explanations." *American Psychologist* 54:594–604.

Winship, C., and M. Sobel. 2004. "Causal Inference in Sociological Studies." Pp. 481–503 in *Handbook on Data Analysis*, edited by M. Hardy and A. Bryman. Thousand Oaks, CA: Sage.

Worm, B., et al. 2006. "Impacts of Biodiversity Loss on Ocean Ecosystem Services." *Science* (314):787–90.