Regular Article

Sociological Methods & Research
2022, Vol. 51(4) 1947–1968
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0049124120914922
journals.sagepub.com/home/smr
SAGE

The probability of a robust inference for internal validity

Tenglong Li[1] and Ken Frank[1]

Abstract

The internal validity of observational study is often subject to debate. In this study, we define the counterfactuals as the unobserved sample and intend to quantify its relationship with the null hypothesis statistical testing (NHST). We propose the probability of a robust inference for internal validity, that is, the PIV, as a robustness index of causal inference. Formally, the PIV is the probability of rejecting the null hypothesis again based on both the observed sample and the counterfactuals, provided the same null hypothesis has already been rejected based on the observed sample. Under either frequentist or Bayesian framework, one can bound the PIV of an inference based on his bounded belief about the counterfactuals, which is often needed when the unconfoundedness assumption is dubious. The PIV is equivalent to statistical power when the NHST is thought to be based on both the observed sample and the counterfactuals. We summarize the process of evaluating internal validity with the PIV into a six-step procedure and illustrate it with an empirical example.

Keywords

observational study, causal inference, hypothesis testing, internal validity, Bayesian statistics, robustness indices, sensitivity analysis

[1] Michigan State University, East Lansing, MI, USA

Corresponding Author:
Tenglong Li, Department of Biostatistics, Boston University, 801 Massachusetts Avenue, Boston, MA 02118, USA.
Email: litenglong0311@gmail.com

Causal inferences are often made based on observational studies, which allow researchers to collect relatively large amounts of data with low cost per research question, compared to randomized experiments (Rosenbaum 2002; Schneider et al. 2007; Shadish, Cook, and Campbell 2002). Internal validity, which refers to whether one can make causal inference between two variables given they are correlated, is frequently challenged and difficult to assess for observational studies (Imai, King, and Stuart 2008; Imbens and Rubin 2015; Murnane and Willett 2011; Rosenbaum 2002, 2010; Rosenbaum and Rubin 1983a; Shadish et al. 2002). To characterize concerns about internal validity in observational studies, we adopt the concept of potential outcome, which refers to the outcome of every subject under every possible treatment (Holland 1986; Rubin 2007, 2008). A fundamental issue is that a subject can only choose one treatment at a time and thus only one potential outcome is observable. This renders all other potential outcomes missing (Imbens and Rubin 2015; Rubin 2005). Essentially, causal inference is treated as a missing data problem where the missing outcomes are assumed to be missing at random (MAR) conditional on a set of covariates, an assumption known as "unconfoundedness" (Imbens 2004; Rosenbaum and Rubin 1983b). Given the difficulty of justifying the unconfoundedness assumption, one may suspect the missing potential outcomes (i.e., counterfactual outcomes) are not MAR conditional on controlled covariates (Heckman 2005; Rosenbaum 1987; Rosenbaum and Rubin 1983a). This implies a missing confounder may exist and consequently the missing potential outcomes may not be comparable to the observed outcomes.

It is noteworthy that observational studies only approximate the missing outcomes based on this assumption; however, if important variables are omitted, such approximation would be misleading. The robustness of a causal inference is defined in this context as whether a causal relationship between two variables can still hold when the unconfoundedness assumption fails. The robustness of a causal inference is evaluated based on one's belief about counterfactual outcomes or missing confounders in order to make a decision about whether this inference is trustworthy (Frank 2000; Frank et al. 2013). We leverage this logic to quantify the robustness of a causal inference based on one's belief about the mean counterfactual outcomes for the treated subjects and the controlled subjects. To do this, we first define counterfactual outcomes as the unobserved sample and incorporate such unobserved sample into the observed sample to form the ideal sample, which, as indicated by its name, is ideal for making a causal inference (Frank et al. 2013; Rubin 2004, 2005; Sobel 1996). We focus the mean counterfactual outcomes (rather than individual values of counterfactual outcomes) because they are sufficient statistics for causal inference in a simple context. We further define the probability of a robust inference for internal validity (henceforth, we abbreviate it as the PIV) based on the ideal sample as the robustness index of internal validity. Our analytical procedure aims to bound the PIV of an inference based on one's belief and inform the robustness of a causal inference based on such bound(s). We apply our approach to Hong and Raudenbush (2005) which estimated a negative effect of kindergarten retention on reading achievement. Although Hong and Raudenbush analyzed a nationally representative sample mitigating concerns about external validity, the treatments (i.e., retained in kindergarten vs. promoted to the first grade) were not randomly assigned in this observational study, raising potential concerns about internal validity (Allen et al. 2009; Frank et al. 2013; Hong 2010; Schafer and Kang 2008).

A Survey of Similar Approaches

Sensitivity Analysis

Sensitivity analysis (Rosenbaum 1986, 1987, 1991, 2002, 2010; Rosenbaum and Rubin 1983a) addresses the influence of a missing confounder on the estimates and inference for regression and nonparametric tests, and more importantly, it connects the violation of unconfoundedness assumption to the violation of random assignment in matched pairs. Therefore, it informs the internal validity of a matching design. Other literature on sensitivity analysis has similar orientation toward missing confounders (Copas and Li 1997; Hosman, Hansen, and Holland 2010; Lin, Psaty, and Kronmal 1998; Masten and Poirier 2018; Robins, Rotnitzky, and Scharfstein 2000; VanderWeele 2008). The PIV shares the objective of checking the sensitivity of results to potential violation of the unconfoundedness assumption with the sensitivity analysis, but the PIV is not limited to a single type of design (like matching) or estimation (like regression). In fact, the PIV can be employed in any design that deemed appropriate for observational studies.

Bayesian Sensitivity Analysis

Bayesian sensitivity analysis (BSA; McCandless and Gustafson 2017; McCandless, Gustafson, and Levy 2007; McCandless et al. 2012) parameterizes the models for explaining the outcome and the unmeasured confounder carefully, so that it can identify the key parameters of confounding effect and examine their impacts on the estimate of treatment effect under a Bayesian framework. BSA has two main advantages: First, the data augmentation in Bayesian modeling allows one to build a model for the unobserved confounder and repeatedly draw random samples of it. As a result, one would get expected distributions of the confounding and treatment effect parameters. Additionally, BSA offers modeling flexibility through prior specification. Comparing to BSA, the implementation and interpretation of the analysis for the PIV would be much easier as BSA is built on complicated Markov chain Monte Carlo algorithms.

The Robustness Indices of Causal Inferences

The robustness indices of causal inferences (Frank 2000; Frank et al. 2013) quantify the strength of internal validity in terms of the impact of an unmeasured confounding variable or the proportion of observed cases can be replaced by the null cases that an inference can afford. The PIV is inherently connected to both papers as it starts with the decision rules and the missing data perspective shared by Frank et al. (2013) and relies on the relationship between the estimate of average treatment effect and null hypothesis statistical testing (NHST), which has been studied by Frank (2000). The PIV is different from the robustness indices because it requires a bounded belief about counterfactual outcomes (or a missing confounder), and it is a probabilistic index which is shown to be equivalent to the statistical power.

Manski's Bounds of Treatment Effect

Bounding treatment effect is proposed by acknowledging the issue of nonidentification of the estimate of average treatment due to counterfactual outcomes (Manski 1990, 1995; Manski and Nagin 1998). Different bounds of treatment effect can be obtained by imposing different assumptions on the counterfactuals, and the bounds of treatment effect would be tightened by making stronger assumption(s). Both the PIV and the bounds of treatment effect proposed by Manski consider the situations when the unconfoundedness assumption is implausible so that one has to form a belief about counterfactual outcomes. Different from the PIV, Manski's bounds are not built on NHST and the parametric (normality) assumption. Rather, the bounds offer insights about the worth of a causal inference through exploring loss-based alternatives rooted in the context of program evaluation. Furthermore, Manski's bounds leverage nonlinear relationships to determine constraints on parameter values, whereas the PIV is built on comparison of means and quantifies the likelihood an inference would hold, assuming normality.

Replication Probability

Various replication probabilities have been proposed for two main reasons: First, they purpose safeguarding readers from the misguidance and misinterpretation of p values. Second, they are used to accentuate that the true scientific significance is about replicability rather than statistical significance (Boos and Stefanski 2011; Greenwald et al. 1996; Killeen 2005; Posavac 2002; Shao and Chow 2002). The PIV is in fact the probability of replicating a significant result in observational study, and it is more akin to p_{rep} (Killeen 2005; Iverson, Wagenmakers, and Lee 2010) which is the probability of obtaining an effect with the same sign as the observed one. Different from p_{rep} and all other replication probabilities, the PIV takes counterfactual outcomes into consideration and therefore it is not a function of p value. Therefore, it does not inherit any weakness from p value like most proposed replication probabilities do (Doros and Geier 2005).

Counterfactual Outcomes as the Unobserved Sample

Research Setting

This article targets observation studies with two groups, that is, the treatment group and the control group. Furthermore, we only consider observational studies with representative samples so that we can focus on internal validity. This article focuses on the simple group-mean-difference estimator (referred to as the simple estimator henceforth) of an average treatment effect, which computes the difference between the adjusted mean treated outcome and the adjusted mean control outcome. The adjusted means can be calculated based on propensity score matching or stratification and perceived as valid estimators of true means of treated outcome and control outcome when the unconfoundedness assumption holds.

Definitions

> **Definition 1:** *The unobserved sample* refers to the collection of the counterfactual outcomes of all sampled subjects. *The unobserved treated sample* refers to the collection of the counterfactual outcomes of the sampled subjects who actually received the control. *The unobserved control sample* refers to the collection of the counterfactual outcomes of the sampled subjects who actually received the treatment.

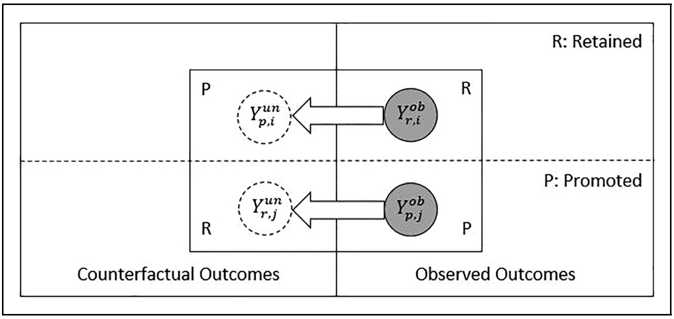

Figure 1. The unobserved sample in Hong and Raudenbush (2005) for the simple estimator.

Example: The unobserved sample of Hong and Raudenbush (2005) is the collection of counterfactual reading scores of sampled students in their study. Specifically, this unobserved sample can be decomposed into the unobserved control sample which is the collection of reading scores of retained students had they all been promoted to first grade and the unobserved treated sample which is the collection of reading scores of promoted students had they all been retained in kindergarten.

Figure 1 illustrates the conceptualization of the unobserved sample in Hong and Raudenbush (2005) for the simple estimator. The observed outcome $Y^{ob}_{r,i}$ symbolizes the reading score of any retained student whose counterfactual outcome is $Y^{un}_{p,i}$. Likewise, the observed outcome $Y^{ob}_{p,j}$ represents the reading score of any promoted student whose counterfactual outcome is $Y^{un}_{r,j}$. The unobserved sample consists of counterfactual outcomes $Y^{un}_{p,i}$ and $Y^{un}_{r,j}$.

Finally, we define the ideal sample as follows:

> **Definition 2**: *The ideal sample* refers to the combination of the observed sample and the unobserved sample. *The ideal treated sample* refers to the combination of the observed treated sample and the unobserved treated sample. *The ideal control sample* refers to the combination of the observed control sample and the unobserved control sample.

Drawing on the definitions above, we argue that it is the unobserved sample that induces the bias which undermines internal validity. The unobserved sample can be perceived as the gap between the observed sample and the ideal sample needed for insuring internal validity. The unconfoundedness assumption implies the unobserved sample is ignorable based on a set of covariates, that is, the unobserved sample will essentially be the same as the observed sample conditional on the set of covariates. Given this assumption is frequently and constantly challenged, our goal is to quantify the robustness of the inference by discovering how the unobserved sample affects the NHST.

Sample Statistics and Notation

This section introduces the notations of the sample statistics defined based on the observed, unobserved, and ideal samples. In general, the observed sample statistics are all fixed and known quantities since the observed sample is held fixed when we consider using the PIV. The unobserved and ideal sample statistics, on the other hand, are unknown quantities of main interest. In this context, δ is a random variable representing the population average treatment effect.

The observed sample statistics (known and fixed): $\hat{\delta}$ is the simple estimator of the average treatment effect based on the observed sample, with $\hat{\delta} = \bar{Y}_t^{ob} - \bar{Y}_c^{ob}$ where $\bar{Y}_t^{ob}$ denotes the adjusted mean outcome of the observed treated subjects and $\bar{Y}_c^{ob}$ denotes the adjusted mean outcome of the observed control subjects, such as adjusted based on a propensity score design. The terms $\hat{\sigma}_t^2$ and $\hat{\sigma}_c^2$ denote the variances of the treated and the control outcomes in the observed sample, respectively. The observed sample size is n^{ob} and the proportion of treated subjects in the observed sample is π.

The unobserved sample statistics (focused unknown): $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$ denote the mean outcomes in the unobserved treated sample and the unobserved control sample, respectively.

The ideal sample statistics (unknown due to the unobserved sample): $\hat{\delta}^{id}$ is the simple estimator of average treatment effect based on the ideal sample and $SE_{\hat{\delta}^{id}}$ is its standard error. $\bar{Y}_t^{id}$ and $\bar{Y}_c^{id}$ denote the mean outcomes in the ideal treated sample and the ideal control sample, respectively, and their difference is $\hat{\delta}^{id}$. Lastly, the variance of $\bar{Y}_t^{id}$ and $\bar{Y}_c^{id}$ are denoted by ϕ_t and ϕ_c, respectively.

We are interested in the distribution of δ, and the randomness of δ is mainly due to the counterfactual outcomes which are potentially different from the observed outcomes. Therefore, the distribution of δ needs to be defined based on the ideal sample, such that both counterfactual and observed outcomes are included. As a result, the distribution of δ is defined by the unobserved and observed sample statistics (or, equivalently, just the ideal sample statistics).

The PIV

The PIV is rooted in NHST context. To conduct a causal inference, the null hypothesis H_0: $\delta = \delta_0$ is assumed to be tested against the alternative hypothesis H_a: $\delta \neq \delta_0$ (δ_0 is usually 0).[1] Here we define $\delta^{\#}$ to be the threshold of rejecting the null hypothesis (and thus finding a significant effect), and for NHST $\delta^{\#}$ is just the product of a critical value C and the standard error of $\hat{\delta}$. Furthermore, the PIV is meaningful when the null hypothesis has been rejected based on the observed sample, and we are interested in whether the null hypothesis would be rejected if counterfactual outcomes were known.

Frank et al. (2013) provided the following decision rules on whether a causal inference will be invalidated due to limited internal validity: Given $\hat{\delta}$ is significantly positive, an inference will be invalidated if $\hat{\delta} > \delta^{\#} > \delta$. Given $\hat{\delta}$ is significantly negative, an inference will be invalidated if $\hat{\delta} < \delta^{\#} < \delta$. Since $\hat{\delta}$ is fixed and exceeds the threshold, the aforementioned decision rules can be simplified as $\delta < \delta^{\#}$ for a significantly positive $\hat{\delta}$ or $\delta > \delta^{\#}$ for a significantly negative $\hat{\delta}$. The decision rules can be also interpreted in the opposite way: An inference cannot be invalidated if $\delta > \delta^{\#}$ for a significantly positive $\hat{\delta}$ or $\delta < \delta^{\#}$ for a significantly negative $\hat{\delta}$. Drawing on this interpretation, the PIV is defined as the probability that an inference cannot be invalidated for the ideal sample $\mathbf{D^{id}}$. Specifically, the PIV is defined as follows for a significantly positive $\hat{\delta}$

$$P(\delta > \delta^{\#} | \mathbf{D^{id}}). \tag{1}$$

Likewise, the PIV is defined as follows for a significantly negative $\hat{\delta}$

$$P(\delta > \delta^{\#} | \mathbf{D^{id}}). \tag{2}$$

It's noteworthy that the PIV in equations (1) and (2) are actually the simplified version of $P(\delta > \delta^{\#} | \hat{\delta} > \delta^{\#}, \mathbf{D^{id}})$ and $P\,(\delta < \delta^{\#} | \hat{\delta} < \delta^{\#}, \mathbf{D^{id}})$, respectively. Given the ideal sample must contain the observed sample, we can ignore the condition $\hat{\delta} > \delta^{\#}$ or $\hat{\delta} < \delta^{\#}$ as they should be conveyed by the ideal sample $\mathbf{D^{id}}$. The PIV essentially is the probability of rejecting the null hypothesis again for the ideal sample, given the same null hypothesis has been rejected for the observed sample, when the counterfactual outcomes has been taken into consideration. By definition, the PIV is the statistical power of retesting the null hypothesis: $\delta = 0$ versus the alternative hypothesis: $\delta = \hat{\delta}^{id}$ ($\hat{\delta}^{id} \neq 0$) based on the ideal sample. When such hypothesis testing is based on either normal or student T distribution, the PIV has the following relationship with the T ratio $T = \frac{\hat{\delta}^{id}}{SE_{\hat{\delta}^{id}}}$:

For a significantly positive $\hat{\delta}$ and a critical value C, we have

$$probit(PIV) = T - C. \tag{3}$$

For a significantly negative $\hat{\delta}$ and a critical value C, we have

$$probit(PIV) = C - T. \tag{4}$$

Note here that equations (3) and (4) will only be approximately true for studies with small sample sizes and typically C is chosen based on the level of significance. For example, C would be 1.96 if $\hat{\delta}$ is significantly positive and the level of significance is 0.05.

The Relationship Between the PIV and the Mean Counterfactual Outcomes

If the treated outcome and the control outcome are independent and roughly normally distributed, the distribution of δ based on the ideal sample would be as follows given their variances are $\hat{\sigma}_t^2, \hat{\sigma}_c^2$:

$$\delta|\mathbf{D^{id}} \sim N(\bar{Y}_t^{id} - \bar{Y}_c^{id}, \phi_t + \phi_c). \tag{5}$$

where,

$$\begin{aligned} \bar{Y}_t^{id} &= (1 - \pi)\bar{Y}_t^{un} + \pi\bar{Y}_t^{ob}, \\ \phi_t &= \frac{\hat{\sigma}_t^2}{n^{ob}}, \\ \bar{Y}_c^{id} &= \pi\bar{Y}_c^{un} + (1 - \pi)\bar{Y}_c^{ob}, \\ \phi_c &= \frac{\hat{\sigma}_c^2}{n^{ob}}. \end{aligned} \tag{6}$$

Here we need to conceptualize $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$. For example, for Hong and Raudenbush (2005), $\bar{Y}_t^{un}$ is the mean reading score of the promoted students had they been retained in the kindergarten, and $\bar{Y}_c^{un}$ is the mean reading score of the retained students had they been promoted to the first grade. The mean outcome in the ideal treated (or control) sample is the weighted average between the mean outcome in the unobserved treated (or control) sample and the mean outcome in the observed treated (or control) sample, while the weight is defined by π. The simple estimator of average treatment effect, that is, $\hat{\delta}^{id}$, equals $\bar{Y}_t^{id} - \bar{Y}_c^{id}$.

It's remarkable that results equations (5) and (6) can be derived in a either frequentist fashion or Bayesian fashion (see derivations in Online Appendix, which can be found at http://smr.sagepub.com/supplemental/), and therefore, it has both frequentist and Bayesian interpretations (Li 2018). In frequentist world, the unobserved sample is part of the ideal sample so that $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$ will shape the distribution of δ as well as the final inference that are built on the ideal sample. In Bayesian world, the prior is conceived to be built on the unobserved sample and the likelihood is built on the observed sample, which is consistent with the literature stating that prior can be treated as a function of the data of particular interest (Diaconis and Ylvisaker 1979, 1985; Frank and Min 2007; Hoff 2009; Pearl and Mackenzie 2018). Strictly speaking, $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$ are the prior parameters in the Bayesian world rather than the sample statistics that are sufficient for the distribution of δ in the frequentist world.

Results equations (5) and (6) show that the distribution of δ conditional on $\mathbf{D^{id}}$ is determined by $\bar{Y}_t^{un}$, $\bar{Y}_c^{un}$ based on the unobserved sample as well as by $\bar{Y}_t^{ob}$, $\bar{Y}_c^{ob}$, n^{ob} based on the observed sample. It further indicates that the probit link of the PIV is a function of $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$, conditional on the observed sample statistics $\pi, n^{ob}, \bar{Y}_t^{ob}, \bar{Y}_c^{ob}, \hat{\sigma}_t^2, \hat{\sigma}_c^2$ and the decision threshold $\delta^{\#}$ for rejecting the null hypothesis. We formalize this relationship as follows:

For a significant positive $\hat{\delta}$, we have

$$probit(PIV) = \frac{\sqrt{n^{ob}}}{\sqrt{\hat{\sigma}_t^2 + \hat{\sigma}_c^2}}\left[(1-\pi)\bar{Y}_t^{un} - \pi\bar{Y}_c^{un} + \left(\bar{Y}_t^{ob} + \bar{Y}_c^{ob}\right)\cdot\pi - \bar{Y}_c^{ob} - \delta^{\#}\right]. \tag{7}$$

For a significant negative $\hat{\delta}$, we have

$$probit(PIV) = \frac{\sqrt{n^{ob}}}{\sqrt{\hat{\sigma}_t^2 + \hat{\sigma}_c^2}}\left[\pi\bar{Y}_c^{un} - (1-\pi)\bar{Y}_t^{un} - \left(\bar{Y}_t^{ob} + \bar{Y}_c^{ob}\right)\cdot\pi + \bar{Y}_c^{ob} + \delta^{\#}\right]. \tag{8}$$

Note that the decision threshold $\delta^{\#}$ could be either a fixed value that is pragmatically set based on transaction cost/policy implication/literature review (Frank et al. 2013) or a statistical threshold that is a product between the critical value and the standard error. When $\delta^{\#}$ is a statistical threshold, it equals $C \times SE_{\hat{\delta}^{id}}$, where the critical value C is chosen based on the level of significance. $SE_{\hat{\delta}^{id}}$, which refers to the standard error of the simple estimator of average treatment effect based on the ideal sample, is computed as follows:

$$SE_{\hat{\delta}^{id}} = \sqrt{\phi_t + \phi_c} = \sqrt{\frac{\hat{\sigma}_t^2 + \hat{\sigma}_c^2}{n^{ob}}}. \tag{9}$$

Resultantly, the probit functions in equation (7) becomes

$$probit(PIV) = \frac{\sqrt{n^{ob}}}{\sqrt{\hat{\sigma}_t^2 + \hat{\sigma}_c^2}}\left[(1-\pi)\bar{Y}_t^{un} - \pi\bar{Y}_c^{un} + \left(\bar{Y}_t^{ob} + \bar{Y}_c^{ob}\right)\cdot\pi - \bar{Y}_c^{ob}\right] - C. \tag{10}$$

Likewise, the probit function in equation (8) becomes

$$probit(PIV) = \frac{\sqrt{n^{\mathrm{ob}}}}{\sqrt{\hat{\sigma}_t^2 + \hat{\sigma}_c^2}}\left[\pi\bar{Y}_c^{un} - (1-\pi)\bar{Y}_t^{un} - \left(\bar{Y}_t^{ob} + \bar{Y}_c^{ob}\right)\cdot\pi + \bar{Y}_c^{ob}\right] + C. \tag{11}$$

Drawing on the results above, one can bound the PIV based on a belief about $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$. For example, if one believe that the mean reading score of the retained students had they been promoted to the first grade (i.e., $\bar{Y}_c^{un}$) equals 45.2 and the upper bound of the mean reading score of the promoted students had they been retained instead (i.e., $\bar{Y}_t^{un}$) is 45.78 (their observed mean reading score), the lower bound of the PIV of Hong and Raudenbush (2005) would be 0.77. If one changes his belief to be $\bar{Y}_t^{un} \leq 45.78$ and $\bar{Y}_c^{un} \geq 44.77$, the lower bound of the PIV of Hong and Raudenbush (2005) would be 0.73 instead.

Example: The Effect of Kindergarten Retention on Reading Achievement

Overview

Alexander, Entwisle, and Dauber (2003) established kindergarten retention as a widespread phenomenon in the United States and with profound impacts for both promoted children and retained children, and therefore, it has long been a controversial issue. To address such controversy, Hong and Raudenbush (2005) conducted an analysis that combined a multilevel model controlling for logits of propensity scores and propensity score strata to evaluate the effects of kindergarten retention policy and actual kindergarten retention on students' academic achievement. They used a nationally representative sample that contained about 7,639 students and 1,070 schools. Drawing on this design, Hong and Raudenbush (2005) estimated the effect of kindergarten retention on students' reading achievement as -9.01 with standard error of 0.68, which amounted to a significant effect whose size is about 0.67. In light of this considerable effect, Hong and Raudenbush (2005) concluded that "children who were retained would have learned more had they been promoted" and therefore "kindergarten retention treatment leaves most retainees even further behind" [page 220].

Nevertheless, the internal validity of Hong and Raudenbush (2005) is subject to debate because propensity score analysis is built on the assumption of unconfoundedness, which implies all confounding variables are able to be observed and controlled in the causal model. However, as argued by Frank et al. (2013), some confounding variables may not be fully measured and controlled, incurring selection bias in the estimate. In cases such that an omitted variable was negatively correlated with kindergarten retention and positively correlated with reading achievement, the negative effect of kindergarten retention could be biased, and thus their inference would be invalidated if such a variable were taken into account.

To address the concern about the internal validity of Hong and Raudenbush's inference, we propose an analytical procedure that employs the PIV and its relationship with the mean counterfactual outcomes. This analytical procedure comprises six steps: (1) get the observed sample statistics, (2) choose critical value C,[2] (3) obtain the relationship between the PIV and the mean counterfactual outcomes, (4) state belief about the mean counterfactual outcomes, (5) bound the PIV, (6) conclusion.

Quantifying the Robustness of the Inference of Hong and Raudenbush (2005)

1. Get the observed sample statistics: The required observed sample statistics are as follows: $\bar{Y}_t^{ob} = 36.77$, $\bar{Y}_c^{ob} = 45.78$, $\hat{\sigma}_t^2 = 143.26$, $\hat{\sigma}_c^2 = 138.83$, $n^{ob} = 7,639, \pi = 0.0617$ (Frank et al. 2013). In this context, $\bar{Y}_t^{ob}$ refers to the observed mean reading score of the retained

students, and $\bar{Y}_c^{ob}$ refers to the observed mean reading score of the promoted students.

2. Choose critical value *C*: Since Hong and Raudenbush (2005) reported the effect of kindergarten retention was significantly negative, we choose *C* as -1.96 which means $\delta^{\#} = -1.96 \times SE_{\hat{\delta}_{id}}$.
3. Obtain the relationship between the PIV and the mean counterfactual outcomes: Once the observed sample statistics and *C* are plugged into the probit model equation (11), the probit model for Hong and Raudenbush can be explicitly written as

   $$probit(PIV) = 0.32\bar{Y}_c^{un} - 4.883\bar{Y}_t^{un} + 209.77. \tag{12}$$

   where, in this context, $\bar{Y}_t^{un}$ refers to the mean counterfactual reading score of the promoted students had they been retained instead, and $\bar{Y}_c^{un}$ refers to the mean counterfactual reading score of the retained students had they been promoted instead.
4. State belief about the mean counterfactual outcomes: This step asks one to state and bound his belief about $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$. To best illustrate this procedure, we form two beliefs about the mean counterfactual outcomes.

   4.1. The first belief: Given the inference of Hong and Raudenbush (2005) mostly informed the mean counterfactual reading score of the retained students (i.e., $\bar{Y}_c^{un}$), we decide to bound $\bar{Y}_t^{un}$ and assume $\bar{Y}_c^{un} = 45.2$. We choose this value because it is the grand sample mean so that $\bar{Y}_t^{un} - \bar{Y}_c^{un}$ measures the degree to which the counterfactual reading scores deviate from the null hypothesis: $\delta = 0$. The probit model equation (12) is thus simplified as follows:

   $$probit(PIV) = 224.28 - 4.883\bar{Y}_t^{un}. \tag{13}$$

   In this case, one need to ask himself "what could the mean reading score of the promoted students had they been retained instead (i.e., $\bar{Y}_t^{un}$) possibly be" when the mean reading score of the retained students had they been promoted instead (i.e., $\bar{Y}_c^{un}$) is assumed to be 45.2. It might be illuminating to reflect on the counterfactual outcomes based on the belief about the average retention effects for the retained students and for the promoted students, identified by $\bar{Y}_t^{ob} - \bar{Y}_c^{un}$ and $\bar{Y}_t^{un} - \bar{Y}_c^{ob}$, respectively. For example, given the average retention effect for the retained students is strongly negative (36.77–45.2 = -8.43), it is reasonable to think the average retention effect for the promoted

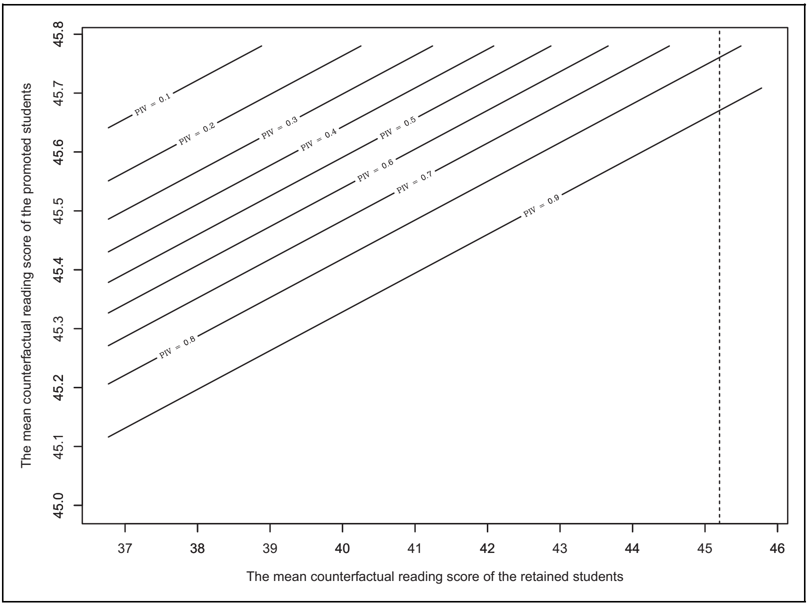

Figure 2. The contour plot of the PIV in the plausible region ($\bar{Y}_t^{un}$: the Y axis; $\bar{Y}_c^{un}$: the X axis). The plausible region is defined based on the belief that the average retention effect for the promoted students should not be positive, and the average retention effect for the retained students was overestimated, which means both $\bar{Y}_t^{un}$ and $\bar{Y}_c^{un}$ are smaller than 45.78. The vertical dashed line corresponds to our first belief where $\bar{Y}_c^{un} = 45.2$.

students should be at least smaller than 0, as supported by literature in recent years. This leads to the upper bound for $\bar{Y}_t^{un}$ as 45.78.

4.2. The second belief: First of all, we believe that the average retention effects for the promoted students and for the retained students should be both negative and the average retention effect for the retained students, which was originally estimated as -9 by Hong and Raudenbush, was overestimated. Therefore, the plausible region is defined based on the bounded belief that $\bar{Y}_t^{un} \leq 45.78$ and $36.77 \leq \bar{Y}_c^{un} \leq 45.78$. Figure 2 is used to illustrate the plausible region. Furthermore, we strengthen this belief by assuming $\bar{Y}_t^{un} \leq 45.2$, which means the mean reading score of the promoted students had they been retained instead cannot exceed the grand sample mean.

5. Bound the PIV: For the first belief, the lower bound for the PIV is 0.77 given $\bar{Y}_t^{un} \leq 45.78$ and $\bar{Y}_c^{un} = 45.2$. This means, given our belief that the mean reading score of the retained students had they been promoted instead is 45.2 and the mean reading score of the promoted students had they been retained instead is at most 45.78, the chance that Hong and Raudenbush's inference is robust for internal validity is at least 77%. For the second belief, the lower bound of the PIV is 0.8 given $\bar{Y}_t^{un} \leq 45.2$ and $36.77 \leq \bar{Y}_c^{un} \leq 45.78$. This means, given our belief that the mean reading score of the retained students had they been promoted instead is at most 45.2 and the average retention effect for the retained students was negative but overestimated, the chance that Hong and Raudenbush's inference is robust for internal validity is at least 80%.
6. Conclusion: To facilitate the decision-making process, one can use a threshold about the PIV such that an inference is deemed robust for internal validity whenever the PIV exceeds this threshold. Since the PIV is the statistical power of retesting the null hypothesis: $\delta = 0$ based on the ideal sample, one can use PIV $= 0.8$ as the threshold which is often used for strong statistical power (Cohen 1988, 1992). Therefore, the two beliefs we formed in the fourth step would lead to the conclusion that Hong and Raudenbush's inference is robust for internal validity. We caution readers that this conclusion might not be hold if one has a different belief and/or a different threshold for the PIV.

There are two key observations in Figure 2: First, in general, the PIV will be more sensitive to $\bar{Y}_t^{un}$ than $\bar{Y}_c^{un}$, which is probably due to the fact that the promoted students predominated the observed sample. This indicates the inference of Hong and Raudenbush is likely to be robust as long as kindergarten retention is believed to have stronger-than-minimal negative impact on the promoted students. Second, even if the kindergarten retention has minimal negative impact on the promoted students, the inference of Hong and Raudenbush (2005) would still be robust for internal validity as long as the average retention effect for the retained students was just slightly overestimated. For example, the lower bound of the PIV is 0.73 when the average retention effect for the retained students was believed to be at least -8 ($\bar{Y}_c^{un} \geq 44.77$), which is one point smaller in size than the original estimate. However, it would be risky to claim that the inference of Hong and Raudenbush (2005) is robust if kindergarten retention has a minimal impact on the promoted students and a significantly overestimated negative impact on the

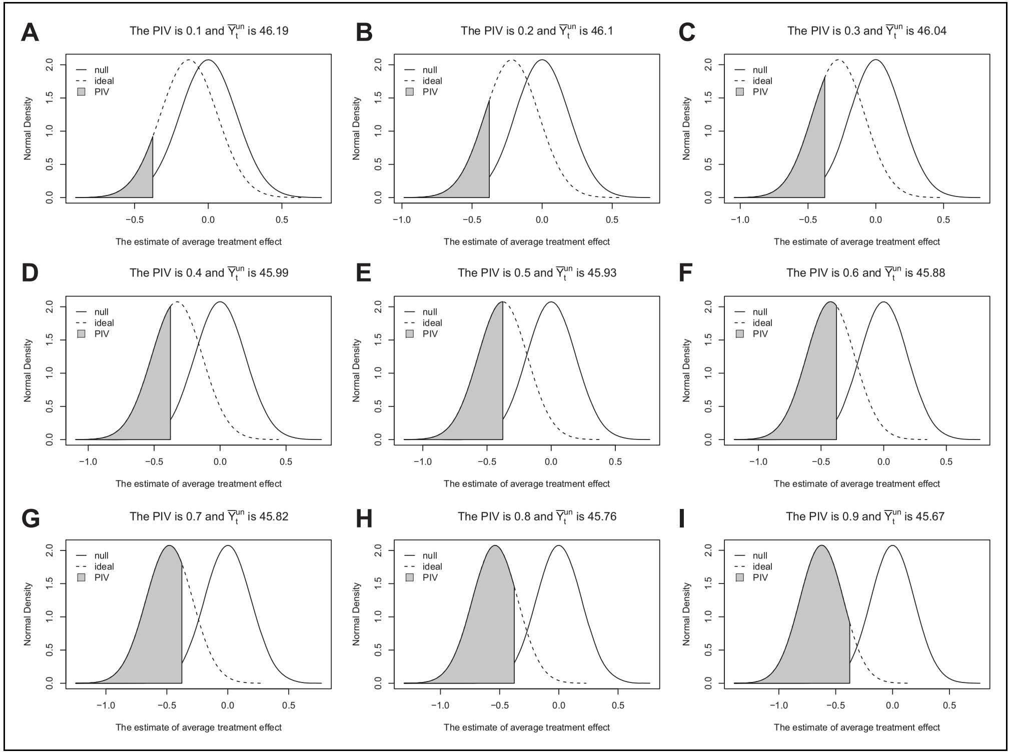

Figure 3. The relationship between the PI and retesting hypothesis in the ideal sample for Hong and Raudenbush (2005), assuming $\bar{Y}_c^{un} = 45.2$. The solid curve represents the null hypothesis: $\delta = 0$, and the dashed curve represents the alternative hypothesis: $\delta = \hat{\delta}^{id}$. The gray shaded area symbolizes the PIV for Hong and Raudenbush. The title of each graph describes the value of the PIV and the value of $\bar{Y}_t^{un}$ corresponds to it.

retained students. For example, the lower bound of the PIV in this case would drop below 0.64 for $\bar{Y}_c^{un} \leq 44$ and $\bar{Y}_t^{un} \leq 45.78$.

By definition, the PIV is the statistical power of retesting the null hypothesis: $\delta = 0$ versus the alternative hypothesis: $\delta = \hat{\delta}^{id}$ ($\hat{\delta}^{id} \neq 0$), had the counterfactual outcomes became observable. This is illustrated by Figure 3 made by fixing $\bar{Y}_c^{un} = 45.2$. It is clear that, as $\bar{Y}_t^{un}$ decreases, the estimate of average treatment effect in the ideal sample (i.e., $\hat{\delta}^{id} = \bar{Y}_t^{id} - \bar{Y}_c^{id}$, see (equation 6)) will be more extremely negative and resultantly the two distributions will be further apart. The PIV will then grow larger as those two distributions overlap less. Figure 3 demonstrates how the PIV is equivalent to the statistical power when retesting the null hypothesis as if the counterfactual outcomes were available.

Conclusion

Focusing on the mean counterfactual outcomes for treated and controlled subjects, we began by defining the unobserved sample as the collection of counterfactual outcomes and the ideal sample as the collection of all the potential outcomes of the observed sample. It's worth emphasizing that the ideal sample is sufficient for securing internal validity, and based on the ideal sample the null hypothesis is thought to be tested against the alternative hypothesis. The PIV is thus defined in this scenario as the probability of rejecting the same null hypothesis again in the ideal sample given it has been rejected in the observed sample. This study recasts the assessment of internal validity as the task of bounding the PIV for an inference based on a bounded belief about the mean counterfactual outcomes.

This article makes three main contributions to the field: First, it promotes counterfactual reasoning by prompting one to conceptualize the mean counterfactual outcomes and form bounded belief about them. Counterfactual reasoning is a necessary step of causal reasoning as it takes one to an imaginary world of what could have happened, thanks to human strength in thinking about cause (Pearl and Mackenzie 2018). Through counterfactual reasoning, causal inference really boils down to comparing the means as one explores all potential outcomes (Imbens and Rubin 2015). The PIV informs internal validity by quantifying the likelihood of an inference would still hold under all different scenarios of counterfactual reasoning. Second, the PIV has an intuitive interpretation. It is the statistical power of retesting the hypothesis $H_0 : \delta = \delta_0$ versus $H_a : \delta = \hat{\delta}^{id}$ in the ideal sample. Therefore, the PIV is pragmatic as it informs how the mean counterfactual outcomes (and thus internal validity) influence the validity of a decision. Third, the modeling framework for the PIV is simple enough for empirical researchers and has both frequentist and Bayesian flavors.

Future work should focus on extending this model in two aspects: First, future work should revise the current model for subpopulations that are either non-normal or heterogeneous in nature as the normality assumption is unlikely to hold in this case. Second, built on the framework which informs how counterfactuals affect the NHST through the PIV, future work needs to delve deeper into why counterfactuals change, which may due to missing confounders, the violation of Stable Unit Treatment Value Assumption (SUTVA), or measurement error.

Authors' Note

Tenglong Li is now affiliated with Department of Biostatistics, Boston University.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Tenglong Li https://orcid.org/0000-0002-5243-1254

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Our framework can be easily modified for one-sided alternative hypothesis.
2. Here we assume one use a statistical threshold, but a decision threshold could be a nonstatistical one. See section "The Relationship Between the PIV and the Mean Counterfactual Outcomes."

References

Alexander, Karl L., Doris R. Entwisle, and Susan L. Dauber. 2003. *On the Success of Failure: A Reassessment of the Effects of Retention in the Primary School Grades*. New York: Cambridge University Press.

Allen, Chiharu S., Qi Chen, Victor L. Willson, and Jan N. Hughes. 2009. “Quality of Research Design Moderates Effects of Grade Retention on Achievement: A Meta-analytic, Multilevel Analysis.” *Educational Evaluation and Policy Analysis* 31(4): 480-99.

Boos, Dennis D. and Leonard A. Stefanski. 2011. “P-value Precision and Reproducibility.” *The American Statistician* 65(4):213-21.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, Jacob. 1992. “A Power Primer.” *Psychological Bulletin* 112(1):155-59.

Copas, John B. and H. G. Li. 1997. “Inference for Non-Random Samples.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59(1):55-95.

Diaconis, Persi and Donald Ylvisaker. 1979. “Conjugate Priors for Exponential Families.” *The Annals of Statistics* 7:269-81.

Diaconis, Persi and Donald Ylvisaker. 1985. “Quantifying Prior Opinion.” *Bayesian Statistics* 2:133-56.

Doros, Gheorghe and Andrew B. Geier. 2005. “Probability of Replication Revisited: Comment on ‘An Alternative to Null-Hypothesis Significance Tests.’” *Psychological Science* 16(12):1005-6.

Frank, Kenneth A. 2000. “Impact of a Confounding Variable on a Regression Coefficient.” *Sociological Methods & Research* 29(2):147-94.

Frank, Kenneth A. and Kyung-Seok Min. 2007. “Indices of Robustness for Sample Representation.” *Sociological Methodology* 37:349-92.

Frank, Kenneth A., Spiro J. Maroulis, Minh Q. Duong, and Benjamin M. Kelcey. 2013. “What Would it Take to Change an Inference? Using Rubin’s Causal Model to Interpret the Robustness of Causal Inferences.” *Education Evaluation and Policy Analysis* 35:437-60.

Greenwald, Anthony G., Richard Gonzalez, Richard J. Harris, and Donald Guthrie. 1996. “Effect Sizes and P Values: What Should Be Reported and What Should be Replicated?” *Psychophysiology* 33(2):175-83.

Heckman, James J. 2005. “The Scientific Model of Causality.” *Sociological Methodology* 35(1):1-97.

Hoff, Peter D. 2009. *A First Course in Bayesian Statistical Methods*. New York: Springer Science & Business Media.

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81(396):945-60.

Hong, Guanglei. 2010. “Marginal Mean Weighting Through Stratification: Adjustment for Selection Bias in Multilevel Data.” *Journal of Educational and Behavioral Statistics* 35(5):499-531.

Hong, Guanglei and Stephen W. Raudenbush. 2005. “Effects of Kindergarten Retention Policy on Children’s Cognitive Growth in Reading and Mathematics.” *Educational Evaluation and Policy Analysis* 27:205-224.

Hosman, Carrie A., Ben B. Hansen, and Paul W. Holland. 2010. “The Sensitivity of Linear Regression Coefficients’ Confidence Limits to the Omission of a Confounder.” *The Annals of Applied Statistics* 4(2):849-70.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. “Misunderstandings Between Experimentalists and Observationalists About Causal Inference.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(2):481-502.

Imbens, Guido W. 2004. “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review.” *The Review of Economics and Statistics* 86:4-29.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Iverson, Geoffrey J., Eric-Jan Wagenmakers, and Michael D. Lee. 2010. “A Model-Averaging Approach to Replication: The Case of p_{rep}.” *Psychological Methods* 15(2):172-81.

Killeen, Peter R. 2005. “An Alternative to Null-Hypothesis Significance Tests.” *Psychological Science* 16(5):345-53.

Li, T. (2018). *“The Bayesian Paradigm of Robustness Indices of Causal Inferences.”* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Lin, Danyu Y., Bruce M. Psaty, and Richard A. Kronmal. 1998. “Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies.” *Biometrics* 54(3):948-63.

Manski, Charles F. 1990. “Nonparametric Bounds on Treatment Effects.” *The American Economic Review* 80(2):319.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, Charles F. and Daniel S. Nagin. 1998. “Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism.” *Sociological Methodology* 28(1):99-137.

Masten, Matthew A. and Alexandre Poirier. 2018. “Identification of Treatment Effects Under Conditional Partial Independence.” *Econometrica* 86(1):317-51.

McCandless, Lawrence C. and Paul Gustafson. 2017. “A Comparison of Bayesian and Monte Carlo Sensitivity Analysis for Unmeasured Confounding.” *Statistics in Medicine* 36(18):2887-901.

McCandless, Lawrence C., Paul Gustafson, and Adrian Levy. 2007. “Bayesian Sensitivity Analysis for Unmeasured Confounding in Observational Studies.” *Statistics in Medicine* 26(11):2331-347.

McCandless, Lawrence C., Paul Gustafson, Adrian R. Levy, and Sylvia Richardson. 2012. "Hierarchical Priors for Bias Parameters in Bayesian Sensitivity Analysis for Unmeasured Confounding." *Statistics in Medicine* 31(4):383-96.

Murnane, Richard J. and John B. Willett. 2011. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. New York: Oxford University Press.

Pearl, Judea and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Posavac, Emil J. 2002. "Using P Values to Estimate the Probability of a Statistically Significant Replication." *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences* 1(2):101-12.

Robins, James M., Andrea Rotnitzky, and Daniel O. Scharfstein. 2000. "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models." Pp. 1-94 in *Statistical Models in Epidemiology*, *the Environment, and Clinical Trials* edited by M. E. Halloran and D. Berry. New York: Springer.

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

Rosenbaum, Paul R. 1987. "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies." *Biometrika* 74(1):13-26.

Rosenbaum, Paul R. 1991. "Sensitivity Analysis for Matched Case-Control Studies." *Biometrics* 47(1): 87-100.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York, NY: Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983a. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome." *Journal of the Royal Statistical Society, Series B (Methodological)* 45:212-18.

Rosenbaum, Paul R. and Donald B. Rubin. 1983b. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.

Rubin, Donald B. 2004. "Teaching Statistical Inference for Causal Effects in Experiments and Observational Studies." *Journal of Educational and Behavioral Statistics* 29(3):343-67.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100(469): 322-31.

Rubin, Donald B. 2007. "The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels With the Design of Randomized Trials." *Statistics in Medicine* 26(1):20-36.

Rubin, Donald B. 2008. “For Objective Causal Inference, Design Trumps Analysis.” *The Annals of Applied Statistics* 2(3):808-40.

Schafer, Joseph L. and Joseph Kang. 2008. “Average Causal Effects From Nonrandomized Studies: A Practical Guide and Simulated Example.” *Psychological Methods* 13(4):279.

Schneider, Barbara, Martin Carnoy, Jeremy Kilpatrick, William H. Schmidt, and Richard J. Shavelson. 2007. *Estimating Causal Effects Using Experimental and Observational Design*. Washington, D.C.: American Educational & Research Association.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin.

Shao, Jun and Shein-Chung Chow. 2002. “Reproducibility Probability in Clinical Trials.” *Statistics in Medicine* 21(12):1727-742.

Sobel, Michael E. 1996. “An Introduction to Causal Inference.” *Sociological Methods & Research* 24(3):353-79.

VanderWeele, Tyler J. 2008. “Sensitivity Analysis: Distributional Assumptions and Confounding Assumptions.” *Biometrics* 64(2):645-49.

Author Biographies

Tenglong Li is postdoctoral associate in the department of biostatistics at Boston University. He is also a lecturer at Northeastern University. His research interests include Bayesian inference, causal inference and computational statistics.

Ken Frank is MSU foundation professor of Sociometrics at Michigan State University. His research interests include social networks and causal inference.