Social Science Research 110 (2023) 102815

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science☆

Kenneth A. Frank [a,*], Qinyun Lin [b], Ran Xu [c], Spiro Maroulis [d], Anna Mueller [e]

[a] Michigan State University, USA
[b] University of Chicago, USA
[c] University of Connecticut, USA
[d] Arizona State University, USA
[e] University of Indiana, USA

ARTICLE INFO

Keywords:
Sensitivity analysis
Causal inference
Pragmatic sociology

ABSTRACT

Social scientists seeking to inform policy or public action must carefully consider how to identify effects and express inferences because actions based on invalid inferences may not yield the intended results. Recognizing the complexities and uncertainties of social science, we seek to inform inevitable debates about causal inferences by quantifying the conditions necessary to change an inference. Specifically, we review existing sensitivity analyses within the omitted variables and potential outcomes frameworks. We then present the Impact Threshold for a Confounding Variable (ITCV) based on omitted variables in the linear model and the Robustness of Inference to Replacement (RIR) based on the potential outcomes framework. We extend each approach to include benchmarks and to fully account for sampling variability represented by standard errors as well as bias. We exhort social scientists wishing to inform policy and practice to quantify the robustness of their inferences after utilizing the best available data and methods to draw an initial causal inference.

1. Introduction

If social science is to inform policy or public action (Burawoy, 2005), the social scientist must attend carefully to the basis for making causal inferences (e.g., Schneider et al., 2007). In Holland's (1986) terms, manipulations based on incorrect inferences will not yield the intended results. But study findings can be ambiguous. Indeed, debate about the general bases for causal inferences in the social sciences dates back to the 1900s (e.g., Rubin, 1974; Thorndike and Woodworth, 1901; see Oakley, 1998 for review).

Concerns about inferences from quasi-experimental studies come immediately to the fore. Consider research on the effects of kindergarten retention (not promoted to the first grade) on achievement in which social scientists have employed a variety of identification strategies to reduce the bias associated with selection into retention (e.g., Burkam et al., 2007; Eide and Showalter, 2001; Hong and Raudenbush, 2005). The controversy primarily concerns the internal validity of the results; despite controlling for background characteristics, how can we be certain that students who were retained are being compared with similar others?

Sensitivity analyses are one of the key tools for expressing uncertainty of findings from quantitative studies. Initiated with Cornfield

☆ This work was supported by grant R305D220022 from the US Institute for Education Sciences.
* Corresponding author.
E-mail address: kenfrank@msu.edu (K.A. Frank).

https://doi.org/10.1016/j.ssresearch.2022.102815
Received 24 May 2022; Received in revised form 14 October 2022; Accepted 14 October 2022
Available online 17 November 2022
0049-089X/

et al.'s (1959) characterization of alternative factors that could account for the estimated effect of smoking on lung cancer, sensitivity analyses have a deep history in health and medicine (e.g., Baer et al., 2021; Brumback et al., 2004; Dorie et al., 2016; Frank et al., 2021a,b; Gastwirth et al., 1998; Lash et al., 2009; Robins, Rotnitzky and Scharfstein, 2000; Rosenbaum and Rubin, 1983a,b; Scharfstein et al., 2021; Vanderweele and Arah, 2011; Vanderweele and Ding, 2017; Walsh et al., 2014; Walter et al., 2020) and have developed in economics (Altonji, Elder and Taber, 2005; Imbens 2003; Oster, 2019), political science (Acharya, Blackwell and Sen 2016; Blackwell, 2014; Neumayer and Plümper, 2017; Plümper and Traunmüller, 2020), psychology (e.g., Fritz et al., 2016; Imai et al., 2010a,b; Lin et al., 2022; Liu and Wang, 2020; Mauro, 1990), sociology (Diprete and Gangl, 2004; Frank, 2000; Frank and Min, 2007); education (Carnegie, Harada and Hill, 2016; Frank et al., 2013a,b; Rosenbaum, 1986), machine learning (Chernozhukov et al., 2021; Jesson et al., 2021; Kallus et al., 2019) and statistics (Cinelli and Haslett, 2020; Copas and Li, 1997; Franks, D'Amour and Feller, 2019; Hong, Yang and Qin, 2021a; Hong et al., 2018; Hosman, Hansen and Holland, 2010).[1] The purpose of this paper is to make sensitivity analyses more useful for social science research and practice. We orient our discussion around two frameworks for sensitivity analysis, one that foregrounds omitted variables, and one that foregrounds potential outcomes.

As an example of the first framework foregrounding omitted variables in the linear model, we generate statements such as "to invalidate an inference of an effect, an omitted variable would have to be correlated at _ with the predictor of interest and with the outcome" (Frank, 2000). This is known as the Impact Threshold of a Confounding Variable (ITCV). As an example of the second framework foregrounding potential outcomes, we generate statements such as "to invalidate the inference, _% of the cases would have to be replaced with counterfactual cases with zero effect of the treatment" (Frank et al., 2013a,b). This is known as the Robustness of Inference to Replacement (RIR) – see also Frank et al. (2021b).

We present a full discussion of sensitivity based on omitted variables and replacement of cases approaches. This includes closed form expressions, applications, benchmarks, conditions to preserve standard errors, and in the discussion a comparison of the frameworks to each other and their use in conjunction with p-values and confidence intervals. Our goal is to provide a set of tools, representative of a large set of techniques, that are broadly applicable. As a set, the techniques presented here provide a more precise language for researchers to debate the strength of the evidence relative to concerns about potential violations of the assumptions of the inference.

It is important to note that sensitivity analyses cannot substitute for the substantive issues of a debate about an inference in terms of the design of the study, relevant alternative explanations, and measures. But sensitivity analyses do inform debates about causal inferences by quantifying the conditions necessary to change inferences. For example, instead of debating whether or not a variable has been omitted from an observational study (at least one most certainly has), one can debate about the properties required of the omitted variable necessary to change the inference. In this sense, sensitivity analyses provide accessible terms for stakeholders to weigh concerns about study design against the evidence supporting an inference. This is especially relevant for broadening discourse to include stakeholders from underrepresented groups who may hold less formal authority in a given policy context.

We emphasize here that **sensitivity analysis should not be used to buttress an otherwise weak model.** Sensitivity analysis begins only after the analyst has presented and defended the strongest model or set of models for causal inference possible, given the data. The sensitivity analyses presented here are also not a substitute for examining the robustness of results to alternative specifications of models, estimation procedures, and measures using observed data and corresponding estimation techniques. Instead, the sensitivity analyses should be thought of as complementary techniques that examine the robustness of inferences to unobserved and hypothetical conditions that cannot be directly addressed with observed data.

We also note that although in some of our derivations below we use a p-value as a threshold for inference, we do not take a specific level (e.g., .05) as hard and fast for making decisions. In fact, the very point of sensitivity analysis is to represent how much an estimate is above or below a threshold for making an inference, stimulating discussion around the uncertainty of inference. Ultimately our goal is to facilitate intelligent discourse (e.g., Black and Donald, 2001; Boltanski and Thévenot, 2006; Habermas, 1987; Weiss, 1977) that can inform action, contributing to *pragmatic sociology*, in which multiple stakeholders, from different backgrounds and with different interests, can assess when there is enough evidence to act.

2. Background

2.1. Control for confounders

The primary challenge to a causal inference is usually made in terms of a confounder (Frank, 2000; Pearl, 2009; Wooldridge, 2010), defined as related to both the predictor of interest and the outcome, and causally prior to both. If a confounder is not accounted for then some or all of the association between the predictor of interest and the outcome could be attributed to the omitted variable. Correspondingly, any inference about an effect of the predictor of interest on the outcome may be invalid.

Consider the model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z, \tag{1}$$

where Y is the outcome (or dependent variable), X is the predictor of interest (or independent variable), and Z is an observed

[1] See the reviews in Carnegie, Harada and Li (2017), Chernozhukov et al. (2021); Franks et al. (2019); Lash et al. (2009); Neumayer and Plümper (2017); Rosenbaum (2002, chapter 4) or for a software review see Kawabata et al. (2022).

confounder. For example, Y might be achievement, X kindergarten retention, and Z socioeconomic status which is related to both kindergarten retention and achievement and is causally prior to both. Any observed association between kindergarten retention and achievement that does not account for the confounder of socioeconomic status may lead to an overestimate and invalid inference of the effect of kindergarten retention on achievement.

Most estimation techniques adjust for observed confounders in some way. Ordinary least squares estimates of models such as (1) adjust for confounding as a function of the correlation of the omitted variable with the predictor of interest and with the outcome. Propensity scores account for the relationship with the predictor of interest through matching, weighting, or stratification (Hong, 2015; Morgan and Winship, 2007; Murnane and Willett, 2010; Rosenbaum and Rubin, 1983). New techniques either select from a range of possible models (Belloni et al., 2016) or average over sets of models (Frank et al., 2013a,b; Young and Holsteen, 2017). Instrumental variables leverage an alternative measure of the predictor of interest that is unrelated to confounders to reduce bias in estimation (e.g., Heckman, 2005; Wooldridge, 2010).[2] While these techniques offer advancements over ordinary least squares estimates of conventional linear models, they are functions of observed covariates only (propensity scores and model averaging)[3] and instrumental variables are based on strong assumptions (e.g., the exclusion restriction and strong instruments – Wooldridge, 2010, page 102) that are difficult to satisfy in practice (Busenbark et al., 2021).

2.2. The need for sensitivity analysis

While the procedures for controlling for observed confounds or averaging across models have contributed greatly to our ability to estimate effects of specific parameters of interest, they do not pre-empt discourse about an inferred effect because they are functions of only the observed variables. Thus, there may be unobserved variables omitted from even the most sophisticated models that might account for any estimated effect. Anticipating that concerns about such omitted variables can never be resolved completely in an observational study, we turn to techniques for informing the discourse about such concerns by quantifying the conditions necessary to change an inference.

Sensitivity analyses do not alter estimates or inferences; they inform our interpretation and ability to communicate the robustness of those inferences. For example, sensitivity analyses turn the challenge "But there may be an omitted variable that biases the estimated effect of kindergarten retention on achievement" to statements about omitted variables such as "An omitted varable would have to be correlated at 0.36 with retention and with achievement to change the inference of an effect of retention on achievement." Or sensitivity analyses can conceptualize bias in terms of changes in the data, producing statements such as "to invalidate the inference of a negative effect of kindergarten retention on achievement one would have to replace 85% of the cases with counterfactual cases for which there was no treatment effect."

The fundamental challenge is to express sensitivity in parsimonious and accessible terms that are nonetheless functions of the dual relationships associated with the confounder – the relationship with the predictor of interest and with the outcome. Early efforts treated these components separately. For example, for matched cases, Rosenbaum (2002, p. 114) shows that, "to attribute the higher rate of death from lung cancer to an unobserved covariate *u* 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" (emphasis added).

Advances in sensitivity analysis have extended early tabular forms (Mauro, 1990) to heat maps (Franks et al., 2019) and line and contour plots showing the combinations of the dual components of confounding that reduce an estimated effect to a specific (or multiple) thresholds (Carnegie et al., 2016; Cinelli and Hazlett, 2020; Dorie et al., 2016; Imbens, 2003; Middleton et al., 2016). While such visualizations are more efficiently informative than comprehensive tables, they still require interpreters of inferences to engage multiple conditions that could alter an inference. Any discourse requires simultaneously navigating up and down dual measures of association, decreasing the chances of coming to a shared understanding of the conditions that will change the inference. Therefore, in Section 3 we will present two approaches that express sensitivity as a single function of the two components of confounding, reducing ambiguity in the discourse about inferences.

2.3. Relationship of sensitivity analyses to sampling variability

Sampling variability is also a critical component of statistical, and causal inference. Sampling variability explicitly recognizes that specific results can change from sample to sample even when an underlying cause is present. For example, one may choose not to infer an effect unless the estimated effect is unlikely (e.g., $p < .05$) to occur by chance alone in samples from a population in which the null hypothesis is true. Implied by this statement is that sampling variability may be used to define a threshold for making an inference. For example, the inference is made when the ratio of an estimated effect to its standard error is greater than a critical value corresponding to a particular p-value (e.g., a $t_{critical}$ value of 1.96 for a p-value of .05).

In spite of the many critiques of p-values, and more generally of the null hypothesis significance testing paradigm (e.g., Harrington et al., 2019; Kraemer, 2019; Trafimow and Marks, 2015), p-values remain central to discourse about scientific findings (Goldfarb and King, 2016; McCann and Schwab, 2020). The standard error is critical in this discourse because it quantifies variability that can occur

[2] For introductions to instrumental variables techniques, see Murnane and Willett (2011), Chapters 10 and 11; Morgan and Winship (2007), Chapter 9; Wooldridge (2010), Chapter 5.

[3] Propensity score analyses also assume that there is a region of common support – that some who received the treatment had a relatively high propensity to receive the control and vice versa (e.g., Morgan and Winship, 2007, Chapter 4).

simply by chance sampling from a single population, as in a randomized controlled trial, without direct implication of a confounder (see Deaton and Cartwright, 2018, for an alternative interpretation). Furthermore, changes in standard errors such as resulting from inclusion of covariates can be interpreted in terms of changes in the sample size through the concept of efficiency. For example, Raftery (1995, page 124) notes an increase of a standard error by 47% (from 0.066 to 0.098) is equivalent to "throwing away more than half the data – $(0.066/.098)^2 = 0.45$."

Sensitivity analysis should not only account for the dual aspects of an omitted confounder, but also for role of sampling variability, through the standard error, in inferences. Stated differently, sensitivity measures should account not only for how alternative conditions can affect an estimate but also its precision through the standard error. In this paper we will attend to sensitivity with respect to estimates and standard errors, including new results deriving the conditions that can change an estimate to a specific threshold while holding the standard error constant.

3. Sensitivity analyses foregrounding omitted variables: Impact Threshold for a confounding variable

The history of sensitivity analysis based on the linear model includes Mauro's (1990) tables and several contemporary techniques including *bias masking* (Middleton et al., 2016), simulation approaches (e.g., Carnegie et al., 2016), and the *robustness value* based on expressions of R^2 (Cinelli and Hazlett, 2020). Nearly all draw on expressions of associations between the omitted variable and the predictor of interest and between the omitted variable and the outcome, with the primary challenge being to generate a single expression that is a function of both associations. These are the two associations that generate changes in the estimated effect for the predictor of interest that might be used to characterize the importance of a covariate (An and Glynn, 2021; Hong and Raudenbush, 2005; Oster, 2019).

One expression of the dual associations of the confounder is the *product* of the two associations: (association of omitted variable with the predictor of interest) x (association of the omitted variable with the outcome). Examples of this type of product can be traced back to Cochran (1938) if not earlier (e.g., to Fisher 1936) as well as to expressions for omitted variable bias in econometrics (Wooldridge, 2010). Through the product each component of confounding is important in proportion to the size of the other; relationships with the outcome are important for variables strongly related to the predictor of interest, and vice versa.

The functional form of the product is implied by other sensitivity analyses. The curvature in Imbens, (2003) line plot implies that small increases in the partial R^2 between omitted variables and assignment reduce an estimate by the targeted amount for large values of the partial R^2 with the outcome, and vice versa (others such as Carnegie et al., 2016; Cinelli and Hazlett, 2020; Dorie et al., 2016, have extended this to contour plots). See similar implications for binary outcomes (Rosenbaum, 2002; Harding, 2003) and in a propensity score framework (Hirano and Imbens, 2001; Hong, Yang and Qin, 2021b) and mediation framework (Imai, Keele and Yamamoto, 2010a).

It is intuitive then to express sensitivity analysis in terms of the product: (association of omitted variable with the predictor of interest) x (association of the omitted variable with the outcome). Specifically, Frank (2000) quantifies the sensitivity of an inference in terms of the product of two correlations: $r_{x\cdot cv}r_{\ y\cdot cv}$, where $r_{x\cdot cv}$ is the sample correlation between the predictor of interest (X) and the confounding variable (CV) and $r_{\ y\cdot cv}$ is the sample correlation between the outcome (Y) and the confounding variable. Consider Hong and Raudenbush's (2005) study of the effects of kindergarten retention on achievement. In Fig. 1, the relationship of interest is between kindergarten retention (X) and reading achievement (Y), which could be impacted by an omitted confounding variable (e.g., motivation). Frank (2000) defines that impact of the confounding variable as *impact* $= r_{\text{retention}\cdot cv}r_{\text{achievement}\cdot cv}$; the two components of confounding are resolved into a single term by taking the product.

Frank (2000) turns the expression *impact* $= r_{x\cdot cv}r_{\ y\cdot cv}$ into sensitivity analysis by showing how large the impact of an omitted variable must be to invalidate an inference. Drawing on Fig. 1, consider the model:

$$\text{Achievement} = \beta_0 + \beta_1 \text{Retention}, \tag{2}$$

and assume $\widehat{\beta}_1$ is statistically significant, rejecting the null hypothesis that $\beta_1 = 0$. But there may be a skeptic who challenges the

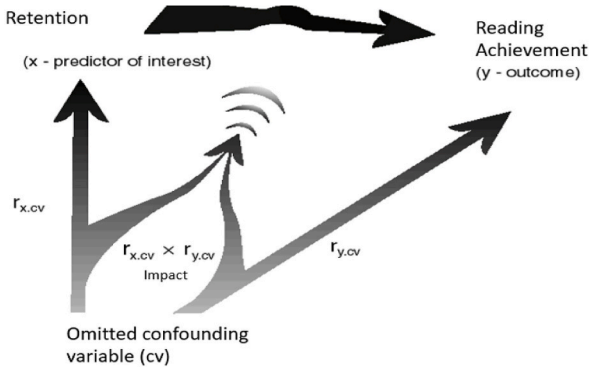

Fig. 1. The impact of a confounding variable.

inference. The skeptic might be a reviewer acting in the name of good science, or the skeptic might be someone who resists the policy implications of rejecting the null hypothesis to protect existing policy. Correspondingly, the skeptic may challenge the inference based on the existence of an omitted variable, that, if included in the model would alter the inference for β_1. Consider the model:

$$\text{Achievement} = \beta_0 + \beta_1 \text{Retention} + \beta_2 \textit{Motivation}, \tag{3}$$

for which *Motivation* is unmeasured. A skeptic might challenge the inference from (1) for which $\widehat{\beta}_1$ is statistically significant by claiming $\widehat{\beta}_1$ would not be statistically significant in (2) upon including the omitted confounding variable, *Motivation*.

Either in response to, or in anticipation of, such debates, researchers routinely employ controls for observed confounds through estimation techniques (e.g., regression analysis, propensity scores, regression discontinuity, instrumental variables). But what if $\widehat{\beta}_1$ is still statistically significant even after controlling for observed confounds? Concerns about omitted variables might persist because it may be difficult to exhaustively account for all confounders with observed variables. The question then is, can the evidence be strong enough relative to a threshold for inference to inform action, even if there are potentially omitted variables?

To respond to this question, Frank (2000) quantifies how strong the impact of an omitted variable must be to invalidate an inference. Specifically, Frank (2000) expresses $\widehat{\beta}_1$ as a function of correlations associated with an omitted variable:

$$\widehat{\beta}_{1|CV} = \frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X} \frac{r_{X\cdot Y} - r_{Y\cdot CV} r_{X\cdot CV}}{1 - r^2_{X\cdot CV}} \tag{4a}$$

where $r_{x\cdot y}$ is the sample correlation between X and Y. Note how the product $r_{x\cdot cv} r_{y\cdot cv}$ appears in the numerator of (4a). Importantly, Frank (2000) also expresses how an omitted variable can affect the standard error representing sampling variability and used for inference:

$$\begin{aligned} se\left(\widehat{\beta}_{1|CV}\right) &= \frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X} \times \sqrt{\frac{1 - R^2_{Y\cdot X}}{n - q - 1} \times \frac{1}{1 - r^2_{X\cdot CV}}} \\ &= \frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X} \times \sqrt{\frac{\frac{1 - \left(r^2_{X\cdot Y} + r^2_{Y\cdot CV} - 2 r_{X\cdot Y} r_{Y\cdot CV} r_{X\cdot CV}\right)}{1 - r^2_{X\cdot CV}}}{n - q - 1} \times \frac{1}{1 - r^2_{X\cdot CV}}} \end{aligned} \tag{4b}$$

where $\widehat{\sigma}$ is a sample variance, n is the sample size, and q the number of covariates. As can be observed, the product $r_{x\cdot cv} r_{y\cdot cv}$ appears in both the expression for $\widehat{\beta}_1$ as well as its standard error.

A straightforward way (Frank et al., 2008) to calculate the impact necessary to invalidate and inference leverages the fact that there is a one-to-one correspondence between a t ratio as in (5) and a partial correlation. Specifically,

$$r_{x\cdot y|cv} = \frac{t_{x\cdot y|cv}}{\sqrt{df + t_{x\cdot y|cv}}}, \text{ where } t_{x\cdot y|cv} = \frac{\widehat{\beta}_{1|CV,Z}}{se\left(\widehat{\beta}_{1|CV,Z}\right)}. \tag{5}$$

The partial correlation, $r_{x\cdot y|cv}$ can be expressed as (Cohen, West and Aitkin, 2014):

$$r_{x\cdot y\ cv} = \frac{r_{x\cdot y} - r_{x\cdot cv} \times r_{y\cdot cv}}{\sqrt{1 - r^2_{y\cdot cv}}\sqrt{1 - r^2_{x\cdot cv}}}. \tag{6}$$

Similar to Frank (2000) for (4a) and (4b), Xu et al. (2019) show the maximum for (6) occurs for $r_{x\cdot cv} = r_{y\cdot cv}$. That is, the smallest possible product that could reduce $r_{x\cdot cv|z}$ below a threshold occurs when $r_{x\cdot cv} = r_{y\cdot cv}$. Thus, making the assumption that $r_{x\cdot cv} = r_{y\cdot cv}$ favors the challenger of the inference, consistent with a conservative stance for making inferences.[4]

If $r_{x\cdot cv} = r_{y\cdot cv}$ then $\textit{impact} = r_{x\cdot cv} \times r_{y\cdot cv} = r^2_{x\cdot cv} = r^2_{y\cdot cv}$. Substituting *impact* for $r_{x\cdot cv} \times r_{y\cdot cv}, r^2_{x\cdot cv}$, and $r^2_{y\cdot cv}$ in (6) yields:

$$r_{xy\ cv} = \frac{r_{x\cdot y} - \textit{impact}}{1 - |\textit{impact}|}. \tag{7}$$

Setting $r_{x\cdot y|cv}$ to be less than or equal to any threshold value, $r^{\#}$, and solving for *impact* yields:

$$\text{Im}\textit{pact} \geq \frac{r_{x\cdot y} - r^{\#}}{1 - |r^{\#}|}. \tag{8}$$

Thus, the partial correlation $r_{x\cdot y|cv}$ will fall below the threshold value of $r^{\#}$ if the *impact* of an omitted confounder is greater than $\frac{r_{x\cdot y} - r^{\#}}{1 - |r^{\#}|}$, which defines the Impact Threshold for a Confounding Variable (ITCV).

The closed form expression in (8) supports intuition about sensitivity. Specifically, sensitivity about an inference is based on the

[4] Frank et al. (2021) observed that estimates change most when $r_{x\cdot cv} = r_{y\cdot cv}$ and Cinelli and Hazlett also assume $r_{x\cdot cv} = r_{y\cdot cv}$ in generating their robustness value but they do not provide a justification, and in their modeling framework the assumption does not necessarily maximize or minimize the estimated effect.

difference between the estimated effect and the threshold for inference ($r_{x\cdot cv}$ -$r^{\#}$). This difference is then scaled relative to the threshold in the denominator (1-$r^{\#}$). An inference based on a given difference is less robust if the threshold is small, as would be the case for large sample sizes.

A second advantage of working in terms of the partial correlation $r_{x\cdot y|cv}$ is that the threshold for statistical significance can be directly calculated. Although $r^{\#}$ can represent any specified threshold, a threshold for statistical significance is defined as

$$r^{\#} = \frac{t_{critical}}{\sqrt{df + t^2_{critical}}} \tag{9}$$

where *df* is the degrees of freedom used to test $\widehat{\beta}_1$. Because the inference for the regression coefficient in (3) is identical to that for the partial correlation in (6), the expressions in (8) and (9) directly account for changes in the estimated effect and its standard error. Correspondingly, when the impact of an omitted variable is greater than the ITCV defined by $r^{\#}$, $\widehat{\beta}_1$ would not be statistically different from zero if the omitted variable were included in the model.

3.1. Application of the ITCV to the inference of an effect of kindergarten retention on achievement

We apply the ITCV to quantify the robustness of an inferred negative effect of kindergarten retention on achievement. Kindergarten retention is a large-scale phenomenon, with the US Department of Health and Human Services (2000) estimating that 8% of second graders (more than five hundred thousand) were a year behind their expected grade level as a result of not being promoted, known as retention, in kindergarten or first grade (see also Alexander et al., 2003, p. 1). As Hong and Raudenbush (2005, page 205) note in their abstract, "grade retention might harm high-risk students by limiting their learning opportunities." Furthermore, a disproportionate percentage of those retained are from low socioeconomic backgrounds and/or are racial minorities (Alexander et al. chap. 5). As Alexander et al. wrote: "next to dropout, failing a grade is probably the most ubiquitous and vexing issue facing school people today" (p. 2). Furthermore, even if a child were retained in anticipation of positive psychological benefits, the state may not realize its investment in a retained child's learning that should lead to a more productive, healthier, and engaged life.

Given the prevalence and importance of retention, there have been considerable studies and syntheses of retention effects (e.g., Alexander et al., 2003; Holmes, 1989; Jimerson, 2001; Karweit, 1992; Shepard and Smith, 1989). Yet none of these studies has been conclusive, as there has been extensive debate regarding the effects of retention, especially regarding which covariates must be conditioned on (e.g., Alexander et al., 2003; Shepard, Smith and Marion, 1998).

In the absence of random assignment which is not ethical for kindergarten retention (Alexander et al., 2003), researchers turn to observational studies that employ statistical controls to estimate effects of retention on achievement. Of the studies of retention effects (Burkam et al., 2007; Jimerson, 2001; Shepard and Smith, 1989), we focus on Hong and Raudenbush's (2005) analysis of nationally representative data in the Early Childhood Longitudinal Study (ECLS) which included extensive measures of student background, emotional disposition, motivation, and pretests.

Hong and Raudenbush (2005) used the measures described above in a propensity score model to define a "retained counterfactual" group representing what would have happened to the students who were retained if they instead had been promoted (e.g., Rubin 1974; Holland 1986). Hong and Raudenbush estimated that the "retained observed" group scored nine points lower on reading achievement than the "retained counterfactual" group at the end of first grade. The estimated effect was about two thirds of a standard deviation on the test, almost half a year's expected growth (Hong and Raudenbush, p. 220), and was statistically significant ($p \leq .001$, with standard error of 0.68, and t-ratio of −13.67). Ultimately, Hong and Raudenbush concluded that retention reduces achievement: "children who were retained would have learned more had they been promoted" (page 200).

Hong and Raudenbush (2005) relied on statistical controls with covariates to approximate equivalence between the retained and promoted groups. But they may not have conditioned for some confounding factor unmeasured in the data set, such as an aspect of a child's cognitive ability, emotional disposition, or motivation. Instead of abandoning the process of inference altogether, we quantify the conditions necessary to change the inference to inform debate about the inference.

Applying the analysis in this section, Hong and Raudenbush's observed t-ratio of −13.67 can be converted to a correlation using (5). Specifically, using sample size of 7639 (for 7168 promoted students, 471 retained students), 221 covariates used to estimate the effects, and assuming 1 extra covariate to be added to the model:

$$r = \frac{t}{\sqrt{(n - q - 1) + t^2}} = \frac{-13.25}{\sqrt{(7639 - 221 - 2) + (13.25)^2}} = -.152.$$

Using equation (9) the degrees of freedom[5] of 7416 yields a threshold correlation $r^{\#}$:

$$r^{\#} = \frac{t_{critical}}{\sqrt{(n - q - 1) + t^2_{critical}}} = \frac{-1.96}{\sqrt{(7639 - 221 - 2) + 1.96^2}} = -.023.$$

From (7),

[5] Df of 7416 adjusts the sample of 7639 for the 221 covariates used in the propensity model.

$$ITCV = \frac{r_{x \cdot y} - r^{\#}}{1 - |r^{\#}|} = \frac{-.152 - -.023}{1 - |.023|} = -.132.$$

Correspondingly, from (7)

$$r_{\mathrm{x} \cdot y\ cv} = \frac{-.152 - (\ -.132)}{1 - |\ -.132|} = -.023 = r^{\#}.$$

An omitted confounding variable would have to have an impact of $r_{\mathrm{x \bullet cv}}$ x $r_{\mathrm{y \bullet cv}} = -0.132$, with component correlations of $.132^{1/2} = 0.36$ (taking opposite signs) to result in a partial correlation of −0.023 (associated with a p-value of .05). Correspondingly, if an omitted variable had an impact greater in magnitude than 0.132 the estimated effect of kindergarten retention on achievement would not be statistically significant ($p > .05$). This calculation accounts for how the omitted variable would change both the estimated effect *and* its standard error as in (4) through (9). All calculations can also be performed with the app at http://konfound-it.com or the R or Stata Konfound commands (see the app for details, or Xu et al., [2019] for the Konfound Stata command and Lin et al., 2022 for the Konfound command in R).

For an observational study it is certainly quite likely that at least one confounding variable was omitted from the model. But the ITCV quantifies what the properties of that omitted variable must be to change the inference. Specifically, an omitted variable would have to be correlated with achievement at .36 and with retention at −0.36 (signs are interchangeable) to invalidate the inference (accounting for sampling variability) of an effect of kindergarten retention on achievement. In this way, the ITCV informs comparisons of the weight of evidence against concerns about an incorrectly specified model, ultimately informing discourse about inference that can be the basis of public action.

3.2. Extensions of the Impact Threshold for a confounding variable (ITCV)

3.2.1. Benchmarks for the ITCV based on correlations associated with observed variables

While the ITCV quantifies the exact hypothetical conditions necessary to change an inference, it can be useful to evaluate the ITCV by comparing with the impacts of observed covariates (e.g., Frank, 2000; Rosenbaum, 1986). To begin, for models such as (3) that already include observed covariate, z, the ITCV should be adjusted (Frank, 2000):

$$ITCV|_z = \frac{r_{x \cdot y|z} - r^{\#}}{1 - |r^{\#}|} = r_{\mathrm{y \cdot cv|z}} r_{\mathrm{x \cdot cv|z}} = \frac{ITCV}{\sqrt{\left(1 - r^2_{\mathrm{y \cdot z}}\right)\left(1 - r^2_{\mathrm{x \cdot z}}\right)}} \tag{10}$$

Then the ITCV|z can be expressed relative to the impact for an observed benchmark covariate z[6]:

$$ITCV(benchmark) = \frac{ITCV|z}{r_{\mathrm{y \cdot z}} r_{\mathrm{x \cdot z}}} = \frac{r_{\mathrm{y \cdot cv}} r_{\mathrm{x \cdot cv}}}{r_{\mathrm{y \cdot z}} r_{\mathrm{x \cdot z}}} \tag{11}$$

The ITCV(benchmark) is the ratio of the unobserved impact necessary to change the inference relative to the observed impact of the covariate z. ITCV(benchmark) > 1 indicates that to invalidate the inference for β_1, the impact of an unobserved covariate would have to be greater than the impact of the observed covariate (Altonji et al., 2005; Oster, 2019). In the kindergarten retention example, the impact of "student approaches to learning (SAL)" (the strongest covariate identified by Hong and Raudenbush 2006) is $r_{\mathrm{SAL,\ retention}} r_{\mathrm{SAL,\ achievement}} = (-0.1849)\ (0.4442) = -0.08$. Correspondingly, the impact of an omitted variable would have to be 65% stronger than the impact of the strongest covariate to change the inference $-0.132/(-0.08) = 1.65$.

Consider the model to now include a vector of observed covariates, Z:

$$\text{achievement} = \beta_0 + \beta_1 \text{retention} + B\text{'Z} \tag{12}$$

Frank (2000) shows that the expression in (11) can be generalized for the model in (12):

$$ITCV(benchmark) = \frac{r_{\mathrm{y \cdot cv}} r_{\mathrm{x \cdot cv}}}{R_{\mathrm{y \cdot Z}} R_{\mathrm{x \cdot Z}}}\ , \tag{13}$$

where the terms $R^2_{\mathrm{y \cdot Z}}$ and $R^2_{\mathrm{x \cdot Z}}$ can be obtained directly from the overall R^2 from (13) and $\widehat{\beta}_1$, $se(\widehat{\beta}_1)$, $\widehat{\sigma}_X$ and $\widehat{\sigma}_Y$. The derivation assumes all the individual impacts for the variables in Z take the same sign (e.g., they all reduce the estimated effect). Thus, one can benchmark using one, some, or all covariates in Z. One can also benchmark conditional on other variables in the model such as pretests which have been shown to dramatically reduce the bias (by 60%–90%) in observational studies when compared with randomized controlled trials (e.g., Shadish et al., 2008; Steiner et al., 2010, 2011; see review in Wong et al. 2017).

3.2.2. Assigning $\widehat{\beta}_1$ to a threshold while preserving the standard error

While the ITCV accounts for changes in an estimate and standard error if a confounding variable were added to a model, it can be

[6] Altonji et al. (2005) and Oster (2019) quantify sensitivity to a confounder in terms of the ratio of selection on observables to selection on unobservables $r_{\mathrm{x \cdot cv}}/r_{\mathrm{y \cdot z}}$, with their specification of the maximum R^2 from (2) implying a value of $r_{\mathrm{y \cdot cv}}$ (Frank et al., 2022).

critiqued as applying only to statistical significance and not generally to evaluation of estimates against thresholds other than statistical significance from zero (e.g., Cinelli and Haslett, 2020). Here we note that sensitivity analyses that focus only on the estimated effect due to an omitted variable (e.g., Carnegie et al., 2016; Dorie et al., 2016; Imbens, 2003) likely make an implicit assumption that the standard error will not change with inclusion of an omitted variable.

In online technical appendix A, we show how to choose correlations associated with Z to modify an estimated effect while holding the standard error constant. For the example of kindergarten retention on achievement, $\widehat{\beta}_{1|Z}$ = -9.01(in absolute value) with a standard error of 0.68. We then set $\widehat{\beta}_{1|cv} = \beta^{\#}$ = -2, which represents about .1 of the growth in a given year and an effect size of about 0.15, near the norm for educational research (Kraft, 2020). The result in online technical appendix A shows that including a confounding variable (cv) in the model for which $r_{x\bullet cv|Z} = 0.28$ and $r_{y\bullet cv|Z} = -0.43$ (with interchangeable signs), will yield $\widehat{\beta}_{1|cv,Z}$ at the threshold of −2 with standard error of 0.68. This shows how we can consider the impact of the omitted variable on the estimated effect while preserving the precision of the original estimate. While $r_{x\bullet cv|Z}$ of 0.28 and $r_{y\bullet cv|Z}$ of −0.43 represent sensitivity in two quantities note that the product yields an impact of $r_{x\bullet cv|Z}\, r_{y\bullet cv|Z}$ =(0.28) (-0.43) = -0.120 which is 50% larger than the benchmark impact of .08 for *SAL*.

4. Sensitivity analysis foregrounding potential outcomes: robustness of inference to replacement (RIR)

An alternative to expressing confounding in terms of the dual components $r_{x\bullet cv}\, r_{y\bullet cv}$ is to express differences between treatment and control groups on potential outcomes, some of which might be due to a confounder related both to treatment assignment and to the outcome. The potential outcomes framework is best understood through the counterfactual sequence: I had a headache; I took an aspirin; the headache went away. Is it because I took the aspirin? One will never know because we do not know what I would have experienced if I had not taken the aspirin. One of the potential outcomes I could have experienced by either taking or not taking an aspirin will be counter to fact, termed the counterfactual within Rubin's Causal Model – RCM (for a history and review of RCM see Holland, 1986; or Morgan and Winship, 2007, chapter 2). In the example in this paper, it is impossible to observe a single student who is simultaneously retained in kindergarten and promoted into the first grade.

Formally expressing the counterfactual shows how potential outcomes can be applied to represent bias from non-random assignment to treatments and thus can be utilized for sensitivity analysis. Define the potential outcome y_i^t as the value on the dependent variable (e.g., reading achievement) that would be observed if unit *i* were exposed to the treatment (e.g., being retained in kindergarten); and define Y_i^c as the value on the dependent variable that would be observed if unit *i* were in the control condition and therefore not exposed to the treatment (e.g., being promoted to the first grade). [i] If SUTVA (Rubin, 1986, 1990) holds – that there are no spillover effects of treatments from one unit to another – then the causal mechanisms are independent across units, and the effect of the treatment on a single unit can be defined as

$$\delta_i = Y_i^t - Y_i^c. \tag{14}$$

The problems of bias due to non-random assignment to treatment are addressed by defining causality for a single unit– there is no concern about confounding because the unit assigned to the treatment is identical to the unit assigned to the control. Similarly, there is no concern about sampling error because the model refers only to the single unit *i.* Of course, the potential outcomes framework does not eliminate the problems of bias due to non-random assignment to treatments or sampling error. Instead, it recasts these sources of bias in terms of missing data (Holland, 1986), because for each unit, one potential outcome is missing.

The potential outcomes framework has been leveraged by multiple approaches to sensitivity analysis from closed form calculations based on matches (e.g., Rosenbaum and Rubin, 1983a,b) to graphical representations (e.g., Cinelli and Hazlett, 2020; Imbens, 2003), to computation of the properties of covariates (e.g., Kallus, Mao and Zhou, 2019; Jesson et al., 2021) to simulation-based techniques that generate full distributions of potential outcomes (e.g., Blackwell, 2014; Brumback et al., 2004; Dorie et al., 2016; Franks et al., 2019). The key is to recognize that there is evidence of confounding if those assigned to the treatment would have done better in the control condition than those assigned to the control. For example, Blackwell (2014) defines confounding in terms of the expected differences between potential outcomes in the absence of a treatment effect. To explore sensitivity, Blackwell (2014) then replaces observed outcomes with outcomes adjusted for a given level of confounding and re-estimates the treatment effect. In a sense, the replaced cases achieve the conditional independence assumption associated with no unobserved confounding in the potential outcomes framework (Rosenbaum and Rubin, 1983a,b).

Here we leverage the RIR (Frank et al., 2013a,b; Frank et al., 2021a,b) to generate a compact expression of robustness based on the potential outcomes framework. The starting point for the RIR is when an analyst makes an inference (from a strong design) when the empirical evidence exceeds a threshold. As is commonly the case in academic research, the threshold can be defined by statistical significance – the threshold is an estimate just large enough to be interpreted as unlikely (e.g., $p < .05$) to occur by chance alone (for a given a null hypothesis). However, the threshold could also be generally defined as the point at which evidence from a study would make one indifferent to the inference. For example, the threshold could be the effect size where the *benefits* of a policy intervention outweigh its *costs* for either an individual or community (Kraft, 2020).

Regardless of the specific threshold, one can compare an estimate with a threshold to represent how much bias there must be to invalidate, or undo, the inference. The more the estimate exceeds the threshold, the more robust the inference with respect to that threshold.

Consider the idealized example in Fig. 2. Here, the estimated treatment effect is 6. If the standard error were 2, then the estimate would be statistically significant from zero if the estimate were greater than $t_{critical}$ x standard error; the estimate must be greater than

$1.96 \times 2 = 3.92$ (where 1.96 represents the critical value of a t-distribution for probability level of 0.05 for a two-tailed test and sample size of 60). Thus, we draw the threshold for inference at 3.92, or about a value of 4 shown on the graph in Fig. 2.

Frank et al.'s, (2013b) interpretation of Fig. 2 is that because one-third of the estimated effect of 6 exceeds the threshold of 4, one-third of the estimate would have to be due to bias to change the inference. One could interpret this purely in terms of omitted variables (e.g., An and Glynn, 2021); an omitted variable would have to reduce the estimated effect by the one-third to invalidate the inference. But this would not account for the corresponding change in standard error upon including the omitted variable, returning to the ITCV.

Frank et al. (2013) also demonstrate that one can interpret the % bias to invalidate an inference in terms of replacing observed cases with counterfactual cases where a null hypothesis of zero treatment effect held. Specifically, one would need to replace 1/3 of the observed cases with cases for which the treatment had no effect to reduce the estimated effect of 6 below the threshold for inference of 4. The larger the proportion to replace, the more robust the inference.

Formally, to calculate the changes in the data necessary to modify an estimated effect to a specific value, we follow Frank et al. (2021) to define the estimated effect from observed and unobserved data as ($\overline{\widehat{\delta}}$) as a function of the observed estimated effect ($\widehat{\delta}_o$) and the hypothesized effect in the unobserved (e.g., counterfactual) replacement data (δ_u). See Cronbach (1982) or Frank and Min (2007) for details. Assuming the proportion of units receiving the treatment is the same in the observed and unobserved data, an expression for $\overline{\widehat{\delta}}$ is:

$$\overline{\widehat{\delta}} = (1-\pi)\widehat{\delta}_O + \pi\delta_U. \tag{15}$$

where π is the proportion of observed cases replaced by unobserved cases. For example, cases can be replaced by their counterfactual counterparts (e.g., treatment cases replaced by counterfactual controls) in which there is no treatment effect. Therefore, $\overline{\widehat{\delta}}$ is a mixture, according to π, of $\widehat{\delta}_o$ and δ_u.

To determine the conditions necessary to change an inference, first assume a null hypothesis of zero effect holds exactly in the unobserved data: $\delta_u = 0$ (Cinelli and Hazlett, 2020; Frank et al., 2013a,b; VanderWeele and Ding, 2017). For example, $\delta_u = 0$ holds exactly if the unobserved data are generated from a null hypothesis of zero effect and there is no sampling variability because there is no covariate imbalance (the approach can also be extended to include countervailing effects in the replacement data – Frank et al., 2013b). Or, $\delta_u = 0$ holds if there are no variables confounded with treatment and outcome. Then set $\overline{\widehat{\delta}} = \delta^{\#}$ where $\delta^{\#}$ defines the threshold for making an inference (such as an estimate associated with an effect size of specific clinical significance – Angst et al., 2017) or with a specific p-value (e.g., 0.05) and finally solving for π yields:

$$\pi = 1 - \frac{\delta^{\#}}{\widehat{\delta}_O} = \text{Robustness of Inference to Replacement (RIR)}. \tag{16}$$

The closed form expression in (16) allows one to calculate what proportion of the cases (π) in the observed sample would have to be replaced with counterfactual zero effect cases to reduce the combined estimate ($\overline{\widehat{\delta}}$) below the threshold ($\delta^{\#}$) for making an inference (Frank et al., 2013a,b). For instance, in the simple example in Fig. 2 where $\widehat{\delta}_o = 6$ and $\delta^{\#} = 4$, $\pi = 1–4/6 = 1/3$, implying that to change the inference, 1/3 of the observed cases would have to be replaced with counterfactual cases in which there was no effect of the treatment.

4.1. Application of the RIR to the inference of an effect of kindergarten retention on achievement[7]

Using the RIR, our question is not whether Hong and Raudenbush's estimated effect of retention was biased because of variables omitted from their analysis. It almost certainly was. Our question instead is "How much bias must there have been to invalidate Hong and Raudenbush's inference?" Using statistical significance as a threshold for Hong and Raudenbush's sample of 7639 (471 retained students and 7168 promoted students, page 215 of Hong and Raudenbush), and standard error of 0.68, $\delta^{\#} = \text{se}\,(\widehat{\delta})\ \text{x}\ t_{\text{critical, df = 7600}} = 0.68\ \text{x}\ (-1.96) = -1.33$; an estimated effect weaker than -1.33 would not be statistically significant assuming the standard error remains at 0.68. Given the estimated effect of -9, to invalidate the inference bias must have accounted for 85.2% of the estimated effect 1-(-1.33/-9) = 0.852. Verifying from (15):

$$\overline{\widehat{\delta}} = (1-\pi)\widehat{\delta}_O + \pi\delta_U$$
$$\Rightarrow (1-.852)(-9) + (.852)(0) = 1.33 = \delta^{\#}$$

Drawing on the potential outcomes framework, to invalidate Hong and Raudenbush's inference of a negative effect of kindergarten retention on achievement one would have to replace 85% of the cases with counterfactual cases for which there was no effect of retention, conditional on the covariates in the model including background characteristics, school membership, and pretests.

Fig. 3 (adapted from Fig. 4 in Frank et al., 2013a,b) shows the replacement process. The dashed lines indicate the observed distribution of test scores for those who were retained (on the left) and promoted (on the right). The black bars represent cases that were

[7] This section is adapted from Frank et al. (2013).

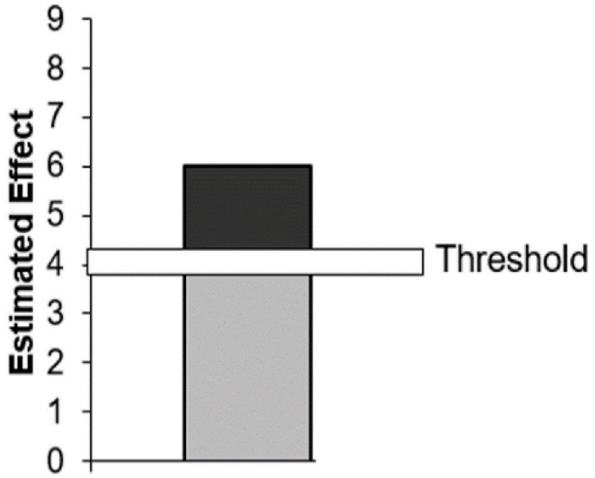

Fig. 2. Estimated effect relative to a threshold for inference.

not replaced. The gray bars represent *counterfactual* data used to replace 85% of the observed cases. The resulting distributions including the black and gray bars would have difference of −1.25 with p = .064.

Our analysis appeals to the intuition of those who consider what would have happened to the promoted children if they had been retained, as these are exactly the potential outcomes on which our analysis is based. Consider test scores of a set of children who were retained that are considerably lower (9 points) than others who were candidates for retention but who were in fact promoted. No doubt some of the difference is due to advantages the comparable others had before being promoted. But now to believe that retention did not have an effect one must believe that 85% of those comparable others would have held their advantages, already conditioned on measured covariates, whether or not they had been retained. Although interpretations will vary, our framework allows us to interpret Hong and Raudenbush's inference in terms of the ensemble of factors that might differentiate retained students from comparable promoted students. In this sense the RIR quantifies the robustness of the inference in terms of the experiences of promoted and retained students as might be observed by educators in their daily practice.

4.2. Extensions of the robustness of inference to replacement (RIR)

4.2.1. Benchmarks for the RIR

As with the ITCV, it can be valuable to interpret the hypothetical conditions associated with the RIR relative to observed data. Once controlling for pre-tests, background characteristics (including mothers' education, poverty, two parent home, and gender) reduced the estimated effect of kindergarten retention on achievement by about 1%.[8] Therefore, once controlling for pretests, omitted factors would have to be approximately 85 times more powerful than background characteristics to invalidate the inference of an effect of kindergarten retention on achievement. We ask: What omitted variable could be 85 times more important than background characteristics, including mothers' education, in accounting for the relationship between kindergarten retention and achievement? The question does not change the initial inference but informs the discourse about that inference.

4.2.2. Conditions necessary to preserve the standard error

Many of the simulation approaches to sensitivity based on the potential outcomes framework account for sampling variability (e.g., Blackwell, 2014; Brumback et al., 2004; Dorie et al., 2016; Franks et al., 2020). While these approaches may simulate the sampling exercise, they introduce a stochastic element into sensitivity that may make discourse more challenging by requiring interpretation over a range of results. As an alternative, we can consider the conditions that ensure that the standard error does not change when cases are replaced. This ensures that changes in the data are due to replacement and not in terms of the precision which as noted earlier has implications for the sample size (Raftery, 1995).

Online technical appendix B shows that cases can be replaced without changing the precision of the estimate as reflected in the standard error. This can be accomplished with a small adaptation of the RIR and with a careful choice of S_y^{unobs}, the standard deviation of Y in the replacement data. In the example of kindergarten retention on achievement, the RIR is 84% which reflects how the standard deviation of Y in the combined data will change when some data are replaced. Correspondingly, the standard deviation in the replacement cases, S_y^{unobs}, is 14.4. That is, if 84% of the cases are replaced with data for which $S_y^{unobs} = 14.4$, then $\widehat{\delta}$ = -1.33 and se $(\widehat{\delta})$ remains at 0.68 as originally reported, with a p-value of .05.

[8] Results available upon request.

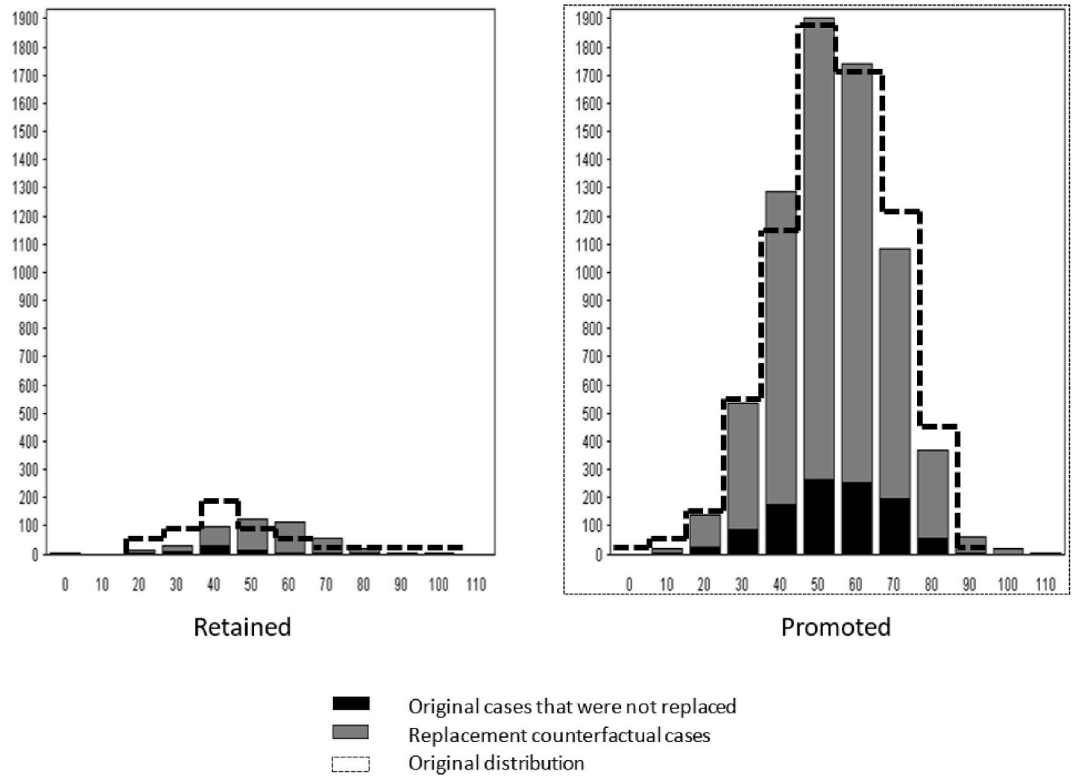

Fig. 3. Example replacement of cases with counterfactual data to invalidate inference of an effect of kindergarten retention on achievement.

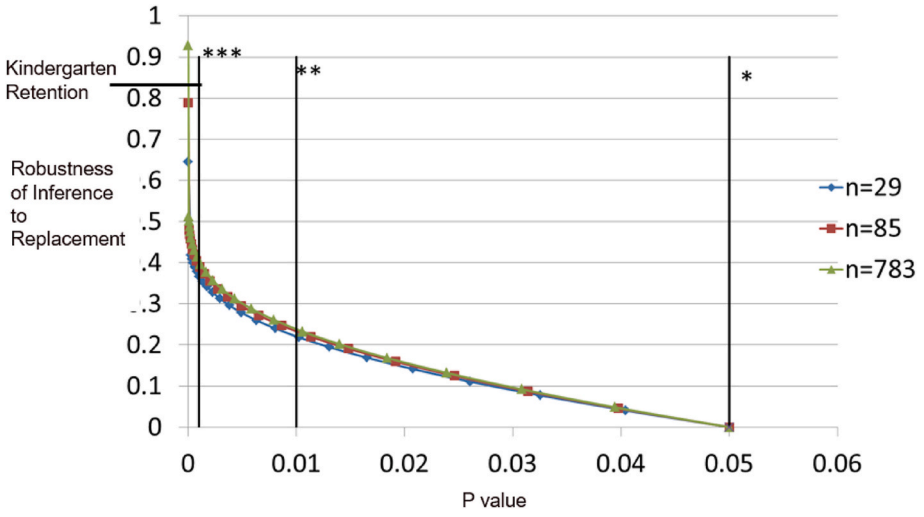

Fig. 4. Robustness of Inference to Replacement (RIR) vs p-value.

4.2.3. Relationship of RIR to p-values and confidence intervals

At its core, we emphasize that sensitivity analyses can inform conversation about the robustness of an inference, helping a broader community weigh in on the link between evidence and practice. RIR can be used to enhance and refine the standard ways researchers communicate the confidence of their findings in terms of P-values and confidence intervals. Consider the language of "highly" or "marginally" statistically significant often used in conjunction with p-values, or the language of an interval "close to zero.". Then examine the comparison of RIR and the p-value depicted in Fig. 4 (assuming a p-value of .05 is used as a basis of an inference, sample sizes chosen to represent .8 power for small, medium and large effects in the social sciences). The steep slope on the left indicates that

the RIR conveys differences in uncertainty even when very small p-values are almost indistinguishable and difficult to interpret (in fact the % bias increases logarithmically with the decline in p-value). Specifically, while it is difficult to conceptualize the difference between p of .01 and p of .001 on the horizontal axis, it is more direct to understand that the inference would change if 25% versus 85% (as in the example of kindergarten retention) of the cases needed to be replaced on the vertical axis.

Confidence intervals are often suggested as alternatives to p-values for expressing the uncertainty of findings that account for sampling variability (e.g., Wilkinson et al., 1999). Typically, researchers are guided to express robustness by assessing how close the boundary of a confidence interval is to zero, ultimately leading to such statements as "rejected overwhelmingly" (Romer, 2020). But the basis for such decisions is usually a comparison of the distance from the lower bound of the confidence interval to zero. This is determined by two quantities: the estimated effect, $\widehat{\delta}$, which determines the location of the center of the interval, and $se(\widehat{\delta})\ t_{critical}$ which determines the width of the interval, and therefore the lower bound conditional on $\widehat{\delta}$. These are exactly the quantities used to define the RIR, which intuitively represents the % bias necessary to invalidate the inference.

4.2.4. Contrast and comparison of the ITCV and RIR

While the ITCV and the RIR draw on different frameworks, they have elements that contrast and allow direct comparison. First, the ITCV is expressed in terms of variables, such that it represents general relationships among constructs that may appeal to those who think in terms of underlying theory. In contrast, the RIR is expressed in terms of cases, which may appeal to those such as clinicians familiar with the discrete observations or individuals. In a sense, then, the RIR quantifies robustness by considering replacing whole cases, whereas the ITCV quantifies robustness by adjusting each case as a function of the value of an omitted variable.

To fully compare the ITCV and RIR, we note how the general linear model approximates the counterfactual condition. In model (2), the variable *retention* takes a value of 1 if the student was retained, 0 if promoted. Therefore, for a given student i who was retained (retention = 1) the counterfactual is defined by rentention = 0. We can then predict the outcome for student i under the counterfactual by substituting retention = 0 instead of retention = 1 into the model:

Observed for student i who was retained (retention = 1)

$$\begin{aligned}&\text{achievement}_i|\text{retention}=1\\&=\beta_0+\beta_1(1)_i+\beta_2\text{motivation}_i+e_i=\beta_0+\beta_1+\beta_2\text{motivation}_i+e_i\ .\end{aligned}\tag{17a}$$

Counterfactual for student i who was retained (retention = 0)

$$\begin{aligned}&\text{achievement}_i|\ \text{retention}=0\\&=\beta_0+\beta_1+\beta_2\text{motivation}_i+e_i=\beta_0+\beta_2\text{motivation}_i+e_i\ .\end{aligned}\tag{17b}$$

Note that β_0, β_2, *motivation* and the error (e_i) in (17b), are exactly as in the observed condition in (17a). This is based on an interpretation of the counterfactual as an alternate universe identical to the observed except for the single change in treatment assignment. Then the difference between (17a) and (17b) is:

$$[\beta_0+\beta_1+\beta_2\text{motivation}_i+e_i]-[\beta_0+\beta_2\text{motivation}_i+e_i]=\beta_1.$$

Thus, for any case specific case β_1 is defined by the difference between the observed outcome conditional on motivation and the counterfactual outcome assuming the same error term, covariates, and estimates of β_0 and β_2. One can decompose differences in expected outcomes between the treatment and control groups into a component based on differences in mean values on the covariates and a component based on differences in estimated coefficients when the covariates are used to predict the outcome separately for the treatment and control groups (Blinder, 1973; Oaxaca, 1973).

While the counterfactual can be approximated by a linear model, it is important to note that because the RIR is rooted in cases it is essentially non-parametric and can be applied to evaluation of any relationship between a predictor of interest and an outcome. Stated another way, whereas the ITCV is most appropriate for conventional linear modeling and NHST, the RIR is appropriate for virtually any kind of estimation technique. As a result, the RIR can readily be adapted to iterative estimation procedures that are not closed form (including multilevel models or generalized estimation equations). For example, Frank et al. (2021) apply the RIR to inferences for dichotomous outcomes, extending the Fragility Index (e.g., Walsh et al., 2014) by accounting for the prevalence of positive and negative outcomes (the technique is extended to logistic regression at http://konfound-it.com, see Lin et al., 2022). This complements Vanderweele and Ding (2017) by using a threshold that is a function of sampling variability. For multilevel models that can be expressed as weighted least squares, one can conceptualize RIR as replacing one observed level two unit (e.g., a school) with a final weight of w with a hypothetical unit that also is associated with a weight of w. We also note that the RIR can be directly applied to concerns about sampling variability as in external validity (Frank and Min, 2007; Frank et al., 2013a,b) such as for a randomized controlled trial.

As noted in the presentation of the ITCV and RIR, each can be considered an evaluation of an estimated effect versus a threshold for inference: $r_{y.x}-r^{\#}$ (when expressed as a correlation metric). The ITCV evaluated this comparison relative to $1-|r^{\#}|$ while the RIR evaluates relative to $r_{y.x}$. Thus, the ITCV includes an extra penality for large samples; large samples produce small values of $r^{\#}$ and therefore a larger denominator and smaller value of the ITCV. On the other hand, the RIR includes a penalty for large estimated effects – the larger the estimated effect the smaller the proportion bias must be to invalidate the inference. Thus the preference for the ITCV versus the RIR depends in part on whether one conceptualizes robustness as scaled relative to the threshold for making the inference or the size of the estimated effect. This has important implications such as applied to BIG data which have large n or replicability experiments which tend to have small n.

5. Discussion

Sensitivity analyses are key for informing discourse about inferences. Sensitivity analyses are especially useful when they express the robustness of an inference in a single quantity representing the dual component of confounding as well as when they account for sampling variability. In this paper we have reviewed two frameworks for accomplishing these goals. First, we presented the Impact Threshold for a Confounding Variable (ITCV) which foregrounds omitted variables in the linear model. This captures the dual elements of confounding as the product of the two correlations associated with potentially omitted variables, and accounts for sampling variability by defining the threshold for inference in terms of the partial correlation (e.g., as associated with a p-value of .05). The second was the Robustness of Inference to Replacement (RIR) which foregrounds the potential outcomes of individuals. This captures the two components of confounding in terms of differences in potential outcome means between the counterfactual treatment and control groups that can be caused by an omitted variable associated with both treatment assignment and outcome. In our example, using the ITCV, an omitted variable would have to be correlated at 0.36 with kindergarten retention and achievement (taking opposite signs) to invalidate Hong and Raudenbush's (2005) inference of a negative effect of retention on achievement. Using the RIR, to invalidate the inference 85% of the cases would have to be replaced with counterfactual cases in which retention had no effect. We note that the ITCV and RIR are representative of sensitivity techniques based on omitted variables or the potential outcomes framework as reviewed in the introduction. To complete the frameworks, we extended the ITCV and RIR to identify the conditions that changed estimates to specific values without increasing or decreasing sampling variability as represented by the standard error.

5.1. Best practices for sensitivity analysis

Application of sensitivity analysis is relatively new, and therefore we provide the following guides to practice. First, sensitivity analyses assume that models have already been developed appropriately and are inclusive of alternative explanations represented by observed variables –**Stated plainly, the first best practice is that sensitivity analysis should only be conducted after a researcher has fully specified a model including all relevant and available measures.** This especially applies to leveraging longitudinal data, as including pretests (lagged dependent variables) in a model can reduce bias by 60–90% versus comparable randomized experiments (e.g., Shadish et al., 2008; see the review in Wong et al., 2017). See also recent work by Belloni et al. (2016), Young and Holsteen (2017) and Young (2018) regarding model selection. **It is misleading to apply sensitivity analyses to models that have not already been rigorously vetted given the available data**. But once the best models have been estimated and adjudicated, there may still be concerns about potential bias through omitted variables or sampling variability. Sensitivity analyses such as the ITCV and RIR can then inform discussions about inferences by quantifying the terms of uncertainty about inferences rooted in the underlying frameworks applied to data analysis (omitted variables and potential outcomes).

Second, the ITCV and RIR may best support rigorous science when used with conventional, transparent methods that can be interpreted by a range of consumers and producers of research. Use of sensitivity analysis allows one to plainly state the adjustments employed in a model used while acknowledging that there may still be other factors unaccounted for. That is, sensitivity analyses focus scientific discourse on the factors that were and were not controlled for rather than on estimation techniques that may be more difficult to track. Ultimately, the transparency of the models and interpretation among researchers can transfer to discourse among a broad range of stakeholders in a policy.

It may be tempting to offer a threshold for a sensitivity index itself such as an ITCV greater than _ or an RIR greater than _ indicates a robust inference. But adding a threshold value for a sensitivity index invites one to then compare how much the index exceeds its own threshold (just barely or by a lot?), which can lead to an infinite spiral of derivative comparisons to thresholds. We believe it is best to make inferences by comparing estimated effects to thresholds that are meaningful in a given literature, and then by quantifying the discourse about the robustness of the inference. These comparisons can then be further evaluated against benchmarks defined by observed covariates. The sensitivity analyses should inform, not preempt by virtue of exceeding a threshold, the discourse among researchers, consumers, and stakeholders invested in an inference.

5.2. The context for sensitivity analyses

We have presented sensitivity analyses in the context of discourse about causal related to policy. But it is important to recognize that policy can be informed by purely descriptive analyses. Policies about voter fraud, sexually transmitted infections, or births and population growth can be informed by purely descriptive analyses.[9] The observational studies we have focused on here occupy a middle ground between the purely descriptive and the more complex models that would account for feedback loops (Maroulis, 2016) and implementation processes (e.g., Frank, Xu and Penuel, 2018). We contend that we can learn from such studies, if interpreted in context and understanding of limitations.

Key to interpretation of any study is the ability for different stakeholders across a range of backgrounds and interests to engage in the discourse about the interpretation, and the sensitivity analyses we present here provide an accessible language for broad discourse. This can apply even to randomized control trials (RCTs) when there are concerns about the external validity of volunteer or convenient samples or imbalance for small samples (e.g., Frank and Min, 2007; Frank et al., 2013a,b; Tipton, 2014). Sensitivity analyses allow

[9] We thank an anonymous reviewer for this observation.

interpreters of research to make inferences based on the strength of the evidence while acknowledging that there may be alternative factors at work or heterogeneity of effects in different populations. This supports pragmatic utilization of a mixture of study designs instead of reliance exclusively on RCTS that may be difficult to replicate when sample sizes are small (Moody, Keister & Ramos-Flor, 2022).

5.3. Conclusion

Sensitivity analyses are vital to the social sciences because the inferences made are about people, and corresponding action will directly affect people. Therefore, inferences will typically be debated by researchers and stakeholders who have different experiences, interests in related practices and policies. In the example used throughout this paper, inferences about kindergarten retention will be debated by family members' seeking the best experiences for their children, teachers who directly educate children, and by schools, districts and the state responsible for educating sets of children. Sensitivity analyses provide a precise language for discourse about such an inference, ultimately supporting better, more informed, decision-making. This may be especially crucial when, as a society, we must weigh the expected benefits and harms of action against the consequences of inaction.

Online Technical Appendix A: Derivation of Correlations Associated with Omitted Variable to Assign $\widehat{\beta}_{1|cv}$ to a Threshold while Preserving the Standard Error

One can make explicit the assumption that the standard error remains unchanged while producing the desired value of $\widehat{\beta}_1$ if an omitted variable were added to a model. To do so, set the expression for $\widehat{\beta}_{1|CV,Z}$ equal to a threshold value, $\beta^{\#}$ and set the standard error for $\widehat{\beta}_{1|CV,Z}$ to be equal to the standard error for $\widehat{\beta}_{1|Z}$:

$$\widehat{\beta}_{1|CV,Z}=\frac{\widehat{\sigma}_{Y|Z}}{\widehat{\sigma}_{X|Z}}\frac{r_{X\cdot Y|Z}-r_{Y\cdot CV|Z}r_{X\cdot CV|Z}}{1-r^2_{X\cdot CV|Z}}=\beta^{\#}, \text{and} \tag{A1}$$

$$\begin{aligned} se\left(\widehat{\beta}_{1|CV,Z}\right)&=\frac{\widehat{\sigma}_{Y|Z}}{\widehat{\sigma}_{X|Z}}\times\sqrt{\frac{1-R^2_{Y\cdot X|Z}}{n-q-1}\times\frac{1}{1-r^2_{X|Z\cdot CV}}} \\ &=\frac{\widehat{\sigma}_{Y|Z}}{\widehat{\sigma}_{X|Z}}\times\sqrt{\frac{1-\left(r^2_{X\cdot Y|Z}+r^2_{Y\cdot CV|Z}-2r_{X\cdot Y|Z}r_{Y\cdot CV|Z}r_{X\cdot CV|Z}\right)}{1-r^2_{X|Z\cdot CV|Z}}\times\frac{1}{1-r^2_{X\cdot CV|Z}}}=se\left(\widetilde{\beta}_{1|Z}\right)=\sqrt{\frac{1-r^2_{X\cdot Y|Z}}{n-q-1}}.. \end{aligned} \tag{A2}$$

The first line in (A1) represents the assignment of $\widehat{\beta}_{1|cv}$ to a specific value, $\beta^{\#}$. The equality on the right hand side of (A2) then reflects the assumption that the standard error remains unchanged when a confound is added to the model: $se(\widehat{\beta}_{1|CV,Z})=se(\widehat{\beta}_{1|Z})$. The expressions in (A1) and (A2) establish a system of two equations and two unknowns: $r_{X\cdot CV}$ and $r_{Y\cdot CV}$. Then using Mathematica (because the expression is very complex producing three solutions, but two that typically have imaginary roots) to obtain a preliminary solution for $r_{X\cdot CV}$ (after which one can directly solve for $r_{Y\cdot CV}$ from either A1 or A2) yields three solutions of which we include the one that produces results with $-1<rxcv<1$ for our example (other solutions available upon request).

Solution:

$$\begin{aligned} rxcv^2=&-\frac{2-b^2+2b\bullet ryx-2ryx^2}{3(-1+ryx^2)}+\frac{\left(1-i\sqrt{3}\right)\left(-\left(2-b^2+2b\bullet ryx-2ryx^2\right)^2+6(-1+ryx^2)\left(b^2-2b\bullet ryx+ryx^2\right)\right)}{T} \\ &-\frac{1+i\sqrt{3}}{6\times 2^{1/3}(-1+ryx^2)}S^{1/3}\Bigg\} \end{aligned}$$

$$rycv=\frac{ryx-b\bullet(1-rxcv^2)}{rxcv}$$

where

$$\begin{aligned} T=&3\times 2^{2/3}\left(-1+ryx^2\right)\Big(-16+15b^2+6b^4+2b^6-30b\bullet ryx-24b^3\bullet ryx-12b^5\bullet ryx+39ryx^2+12b^2\bullet ryx^2+18b^4\bullet ryx^2 \\ &+24b\bullet ryx^3+8b^3\bullet ryx^3-30ryx^4-27b^2\bullet ryx^4+6b\bullet ryx^5+7ryx^6 \\ &+\sqrt{\Big(\left(-16+15b^2+6b^4+2b^6-30b\bullet ryx-24b^3\bullet ryx-12b^5\bullet ryx+39ryx^2+12b^2\bullet ryx^2+18b^4\bullet ryx^2+24b\bullet ryx^3+8b^3\bullet ryx^3\right.} \\ &\left.-30ryx^4-27b^2\bullet ryx^4+6b\bullet ryx^5+7ryx^6\right)^2+4\Big(-\left(2-b^2+2b\bullet ryx-2ryx^2\right)^2+6\left(-1+ryx^2\right)\left(b^2-2b\bullet ryx+ryx^2\right)\Big)^3\Big)\Big)^{1/3} \end{aligned}$$

$$S = -16 + 15b^2 + 6b^4 + 2b^6 - 30b \bullet ryx - 24b^3 \bullet ryx - 12b^5 \bullet ryx + 39ryx^2 + 12b^2 \bullet ryx^2 + 18b^4 \bullet ryx^2 + 24b \bullet ryx^3 + 8b^3 \bullet ryx^3 - 30ryx^4 - 27b^2 \bullet ryx^4 + 6b \bullet ryx^5 + 7ryx^6 + \sqrt{\Big(\big(-16 + 15b^2 + 6b^4 + 2b^6 - 30b \bullet ryx - 24b^3 \bullet ryx - 12b^5 \bullet ryx + 39ryx^2 + 12b^2 \bullet ryx^2 + 18b^4 \bullet ryx^2 + 24b \bullet ryx^3 + 8b^3 \bullet ryx^3 - 30ryx^4 - 27b^2 \bullet ryx^4 + 6b \bullet ryx^5 + 7ryx^6\big)^2 + 4\Big(-\big(2 - b^2 + 2b \bullet ryx - 2ryx^2\big)^2 + 6\big(-1 + ryx^2\big)\big(b^2 - 2b \bullet ryx + ryx^2\big)\Big)^3\Big)}$$

where b represents the threshold. For models that already contain covariates z, the threshold must be modified:

$$\widetilde{\beta} = \beta^{\#} \frac{\widehat{\sigma}_{x|z}}{\widehat{\sigma}_{y|z}}.$$

These analyses can be run in R using the followign commands (please check the Github page as we are actively updating the konfound package):

Install.packages("devtools")

Devtools:install_github("jrosen48/konfound")

Library(konfound).

Pkonfound(est_eff = 0.125, std_err = 0.050, n_obs = 6300, sdx = 0.217, sdy = 0.991, R2 = 0.251, eff_thr = 0, FR2max = 0.61, index = "PSE", to_return = "raw_output")

Online Technical Appendix B: Deriving sd_y^{unob} to preserve $se(\widehat{\delta})$ as in the Robustness of Inference to Replacement

To derive the sd_y^{unob} that preserves $se(\widehat{\delta})$ first note:

$$se(\widehat{\delta}) = \frac{sd_{y|x}}{\sqrt{SSX}} = \frac{sd_{y|x}}{sd_x\sqrt{df}}. \quad \Rightarrow sd_{y|x} = se(\widehat{\delta}) sd_x \sqrt{df} \tag{B1}$$

Therefore, $sd_{y|x}$ can be calculated from reported quantities. Next,

$$sd_{y|x} = sd_y^{combined}\sqrt{1-(r^{\#})^2} \quad \Rightarrow sd_y^{combined} = \frac{sd_{y|x}}{\sqrt{1-(r^{\#})^2}}, \tag{B2}$$

where

$$r^{\#} = \frac{t_{critical}}{\sqrt{t_{critical}^2 + df}}. \tag{B3}$$

Therefore, $sd_y^{combined}$ can also be calculated from $sd_{y|x}$ and known quantities $t_{critical}$ and df.

Before proceeding, note that because $d_y^{combined} \neq sd_y$, we must define π as:

$$\pi = 1 - \frac{r^{\#}\frac{sd_y^{combined}}{sd_y}}{r_{xy}}. \tag{B4}$$

Finally, assume (these assumptions can be relaxed using online appendix B in Frank and Min, 2007):

1) the variance and mean of the predictor are the same in the replacement cases as in the observed cases: $s_x^{2unobs} = s_x^{2unobs}$; $\bar{x}^{unobs} = \bar{x}^{unobs}$. This assumption does not pertain to the relationship between X and Y used for inference. The assumption holds, for example, if 30 treatment and 20 control cases removed from the observed data are replaced with 30 treatment and 20 control cases with different values on Y.
2) the variance in the sample of observed cases not replaced is s_y^{2obs} (as would be expected under for random sampling); and
3) $\bar{y}^{unobs} = \bar{y}^{obs}$ (that is, that the source of data does not affect the overall mean of Y).

Then the variance in the combined data ($sd_y^{2combined}$) including those original cases not replaced as well as the replacement data is:

$$sd_y^{2combined} = (1-\pi)s_y^{2obs} + \pi s_y^{2unob} \quad \Rightarrow sd_y^{unob} = \sqrt{\frac{sd_y^{2combined} - (1-\pi)s_y^{2obs}}{\pi}}. \tag{B5}$$

Therefore, sd_y^{unob} can be calculated from (B5) given the values of $sd_y^{2combined}$ and the sample values of s_y^{2obs} and π.

A combined expression for sd_y^{unob} is:

$$sd_y^{unob} = \sqrt{\frac{\left(\frac{se(\widehat{\beta})sd_x\sqrt{df}}{\sqrt{1-(r^{\#})^2}}\right)^2 - (1-\pi)s_y^{2obs}}{\pi}}. \quad \text{(B6)}$$

These results may be less stable when the number of cases to replace is less than 10.

In the example of Hong and Raudenbush's (2005) inference of an effect kindergarten retention on achievement:

$$sd_{y|x} = se(\widehat{\delta})sd_x\sqrt{df} = (.68)(.24)\sqrt{7637} = 14.26,$$

$$r^{\#} = \frac{t_{critical}}{\sqrt{t_{critical}^2 + df}} = \frac{-1.96}{\sqrt{-1.96^2 + 7637}} = -.022$$

and

$$sd_y^{combined} = \frac{sd_{y|x}}{\sqrt{1-(r^{\#})^2}} = \frac{14.26}{\sqrt{1-(-.022)^2}} = 14.27.$$

Then

$$\pi = 1 - \frac{r^{\#}\frac{sd_y^{combined}}{sd_y}}{r_{xy}} = 1 - \frac{-.022\frac{14.26}{13.50}}{-.15} = .84,$$

where.

$r_{xy} = \frac{t}{\sqrt{t^2+df}} = \frac{-13.25}{\sqrt{-13.25^2+7637}} = -.15$, based on

$T = sd_{y|x} = \frac{\widehat{\delta}}{se(\delta)} = \frac{-9.01}{.68} = -13.25$ where $\widehat{\delta}$ and $se(\widehat{\delta})$ are obtained from Hong and Raudenbush (2005) as reported in the main text.

Then

$$sd_y^{unob} = \sqrt{\frac{sd_y^{2combined} - (1-\pi)s_y^{2obs}}{\pi}} = \sqrt{\frac{14.27^2 - (1-.84)13.5^2}{13.5}} = 14.40.$$

Therefore, if 84% of the cases are replaced with cases for which $sd_y^{unob} = 14.40$ the estimated effect will be −1.33 with standard error of 0.68 associated with p = .05 (R code available upon request).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ssresearch.2022.102815.

References

Acharya, A., Blackwell, M., Sen, M., 2016. Explaining causal findings without bias: detecting and assessing direct effects. Am. Polit. Sci. Rev. 110 (3), 512–529.

Alexander, K., Entwisle, D.R., Dauber, S.L., 2003. On the Success of Failure: a Reassessment of the Effects of Retention in the Primary School Grades. Cambridge University Press, Cambridge, UK.

Altonji, J.G., Elder, T.E., Taber, C.R., 2005. Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. J. Polit. Econ. 113 (1), 151–184.

An, W., Glynn, N., A., 2021. Treatment effect deviation as an alternative to blinder–oaxaca decomposition for studying social inequality. Socio. Methods Res. 50 (3), 1006–1033.

Angst, F., Aeschlimann, A., Angst, J., 2017. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J. Clin. Epidemiol. 82, 128–136.

Baer, B.R., Gaudino, M., Charlson, M., Fremes, S.E., Wells, M.T., 2021. Fragility indices for only sufficiently likely modifications. Proc. Natl. Acad. Sci. USA 118 (49).

Belloni, A., Chernozhukov, V., Wei, Y., 2016. Post-selection inference for generalized linear models with many controls. J. Bus. Econ. Stat. 34 (4), 606–619.

Black, N., Donald, A., 2001. Evidence based policy: proceed with careCommentary: research must be taken seriously. BMJ 323 (7307), 275–279.

Blackwell, M., 2014. A selection bias approach to sensitivity analysis for causal effects. Polit. Anal. 22 (2), 169–182.

Blinder, A.S., 1973. Wage discrimination: reduced form and structural estimates. J. Hum. Resour. 436–455.

Boltanski, L., Thévenot, L., 2006. On Justification: Economies of Worth, vol. 27. Princeton University Press.

Brumback, B.A., Hernán, M.A., Haneuse, S.J., Robins, J.M., 2004. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stat. Med. 23 (5), 749–767.

Burkam, D.T., LoGerfo, L., Ready, D., Lee, V.E., 2007. The differential effects of repeating kindergarten. J. Educ. Stud. Placed A. T. Risk 12 (2), 103–136.

Busenbark, J.R., Frank, K.A., Maroulis, S.J., Xu, R., Lin, Q., 2021. Quantifying the robustness of inferences for strategic management in urgent times: the impact threshold of a confounding variable and robustness of inference to replacement. Res. Methodol. Strat. Manag. 13, 127–154.

Burawoy, M., 2005. For public sociology. Am. Socio. Rev. 70 (1), 4–28.

Carnegie, N.B., Harada, M., Hill, J.L., 2016. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. Journal of Research on Educational Effectiveness 9 (3), 395–420.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., Syrgkanis, V., 2021. Omitted Variable Bias in Machine Learned Causal Models arXiv preprint arXiv:2112.13398.

Cinelli, C., Hazlett, C., 2020. Making sense of sensitivity: extending omitted variable bias. J. Roy. Stat. Soc. B 82 (1), 39–67.

Cochran, W.G., 1938. The omission or addition of an independent variate in multiple linear regression. J. Roy. Stat. Soc. Suppl. 5 (2), 171–176.
Cohen, P., West, S.G., Aiken, L.S., 2014. Applied Multiple Regression/correlation Analysis for the Behavioral Sciences. Psychology press.
Copas, J.B., Li, H.G., 1997. Inference for non-random samples. J. Roy. Stat. Soc. B 59 (1), 55–95.
Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B., Wynder, E.L., 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. J. Natl. Cancer Inst. 22 (1), 173–203.
Cronbach, L.J., 1982. Designing Evaluations of Educational and Social Programs. Jossey-Bass, San Francisco.
Deaton, A., Cartwright, N., 2018. Understanding and misunderstanding randomized controlled trials. Soc. Sci. Med. 210, 2–21.
Diprete, T., Gangl, M., 2004. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. Socio. Methodol. 34.
Dorie, V., Harada, M., Carnegie, N.B., Hill, J., 2016. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. Stat. Med. 35 (20), 3453–3470.
Eide, E.R., Showalter, M.H., 2001. The effect of grade retention on educational and labor market outcomes. Econ. Educ. Rev. 20 (6), 563–576.
Fisher, R.A., 1936. Statistical Methods for Research Workers, eleventh ed. OliverandBoyd, Edinburgh. sixth ed.
Frank, K.A., 2000. Impact of a confounding variable on the inference of a regression coefficient. Socio. Methods Res. 29 (2), 147–194.
Frank, K.A., Dai, S., Jess, N., Lin, H.C., Lin, W., Liu, Y., Maestrales, S., Searle, E., Tait, J., 2022. Exact Calculation of Coefficient of Proportionality Including Evaluation of Oster's δ*, Corresponding Bounds, and Alternatives. Presented at the Society for Research on Educational Effectiveness, Washington, D.C.. Sept 2022.
Frank, K.A., *Lin, Q., *Maroulis, S., *Mueller, A.S., Xu, R., Rosenberg, J.M., et al., 2021a. Response to "three comments on the RIR method". J. Clin. Epidemiol. S0895–S4356.
Frank, K.A., Maroulis, S., Duong, M., Kelcey, B., 2013a. What would it take to change an inference?: using Rubin's causal model to interpret the robustness of causal inferences. Educ. Eval. Pol. Anal. 35, 437–460.
Frank, K.A., Min, K., 2007. Indices of robustness for sample representation. Socio. Methodol. 37, 349–392 (* co first authors.
Frank, K.A., Muller, C., Mueller, A.S., 2013b. The embeddedness of adolescent friendship nominations: the formation of social capital in emergent network structures. Am. J. Sociol. 119 (1), 216–253.
Frank, K.A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., McCrory, R., 2008. Does NBPTS certification affect the number of colleagues a teacher helps with instructional matters? Educ. Eval. Pol. Anal. 30 (1), 3–30.
Frank, K.A., *Lin, Q., *Maroulis, S., *Mueller, A.S., Xu, R., Rosenberg, J.M., et al., 2021b. Hypothetical case replacement can be used to quantify the robustness of trial results. J. Clin. Epidemiol. 134, 150–159. *authors listed alphabetically.
Frank, K.A., *Xu, Ran, Penuel, W.P., 2018. Implementation of evidence based practice in human service organizations: implications from agent-based models. J. Pol. Anal. Manag. 37 (4), 4867–4895 (*Co-equal first authors.
Franks, A., D'Amour, A., Feller, A., 2019. Flexible sensitivity analysis for observational studies without observable implications. J. Am. Stat. Assoc.
Fritz, M.S., Kenny, D.A., MacKinnon, D.P., 2016. The combined effects of measurement error and omitting confounders in the single-mediator model. Multivariate Behav. Res. 51 (5), 681–697. https://doi.org/10.1080/00273171.2016.1224154.
Gastwirth, J.L., Krieger, A.M., Rosenbaum, P.R., 1998. Dual and simultaneous sensitivity analysis for matched pairs. Biometrika 85 (4), 907–920.
Goldfarb, B., King, A.A., 2016. Scientific apophenia in strategic management research: significance tests and mistaken inference. Strat. Manag. J. 37 (1), 167–176.
Harrington, D., D'Agostino, R.B., Gatsonis, C., Hogan, J.W., Hunter, D.J., Normand, S.-L.T., et al., 2019. New guidelines for statistical reporting in the journal. N. Engl. J. Med. 381, 285–286. https://doi.org/10.1056/NEJMe1906559.
Habermas, J., 1987. Knowledge and Human Interests. Polity Press, Cambridge, United Kingdom.
Harding, D.J., 2003. Counterfactual models of neighborhood effects: the effect of neighborhood poverty on dropping out and teenage pregnancy. Am. J. Sociol. 109 (3), 676–719.
Heckman, J.J., 2005. The scientific model of causality. Socio. Methodol. 35 (1), 1–97.
Hirano, K., Imbens, G.W., 2001. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv. Outcome Res. Methodol. 2 (3), 259–278.
Holland, P.W., 1986. Statistics and causal inference. J. Am. Stat. Assoc. 81, 945–970.
Holmes, C.T., 1989. Grade level retention effects: a meta-analysis of research studies. In: Shepard, L.A., Smith, M.L. (Eds.), Flunking Grades. Falmer Press, New York, pp. 16–33.
Hong, G., 2015. Causality in a Social World: Moderation, Meditation and Spill-Over. John Wiley & Sons, Ltd.
Hong, G., Yang, F., Qin, X., 2021a. Did you conduct a sensitivity analysis? A new weighting-based approach for evaluations of the average treatment effect for the treated. J. Roy. Stat. Soc. 184 (1), 227–254.
Hong, G., Qin, X., Yang, F., 2018. Weighting-based sensitivity analysis in causal mediation studies. J. Educ. Behav. Stat. 43 (1), 32–56.
Hong, G., Yang, F., Qin, X., 2021b. Post-Treatment Confounding in Causal Mediation Studies: A Cutting-Edge Problem and a Novel Solution via Sensitivity Analysis arXiv preprint arXiv:2107.11014.
Hong, G., Raudenbush, S.W., 2005. Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. Educ. Eval. Pol. Anal. 27 (3), 205–224.
Hosman, C.A., Hansen, B.B., Holland, P.W., 2010. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. Ann. Appl. Stat. 4 (2), 849–870.
Imbens, G., 2003. Sensitivity to exogeneity assumptions in program evaluation. Am. Econ. Rev. 93 (2), 126–132.
Imai, K., Keele, L., Tingley, D., 2010a. A general approach to causal mediation analysis. Psychol. Methods 15 (4), 309–334. https://doi.org/10.1037/a0020761.
Imai, K., Keele, L., Yamamoto, T., 2010b. Identification, inference and sensitivity analysis for causal mediation effects. Stat. Sci. 25 (1), 51–71.
Jesson, A., Mindermann, S., Gal, Y., Shalit, U., 2021. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In: International Conference on *Machine Learning*. PMLR, pp. 4829–4838.
Jimerson, S., 2001. Meta-analysis of grade retention research: implications for practice in the 21st century. Sch. Psychol. Rev. 30 (3), 420–437.
Kallus, N., Mao, X., Zhou, A., 2019. Interval estimation of individual-level causal effects under unobserved confounding. In: The 22nd International Conference *on Artificial Intelligence and Statistics*. PMLR, pp. 2281–2290.
Karweit, N.L., 1992. Retention policy. In: Alkin, M. (Ed.), Encyclopedia of Educational Research. Macmillan, New York, pp. 114–118.
Kawabata, E., Tilling, K., Groenwold, R.H., Hughes, R.A., 2022. Quantitative Bias Analysis in Practice: Review of Software for Regression with Unmeasured Confounding. medRxiv. https://www.medrxiv.org/content/medrxiv/early/2022/02/21/2022.02.15.22270975.full.pdf.
Kraemer, H.C., 2019. Is it time to ban the P value? JAMA Psychiatr. 76 (12), 1219–1220.
Kraft, M.A., 2020. Interpreting effect sizes of education interventions. Educ. Res. 49 (4), 241–253.
Lash, Timothy L., Fox, Matthew P., Fink, Aliza K., 2009. Applying Quantitative Bias Analysis to Epidemiologic Data. Springer, New York.
Lin, Q., *Maroulis, S., *Rosenberg, J.M., Frank, K.A., Xu, R., Mueller, A.S., Dietz, T., 2022. Robustness of inference to replacement using the konfound R package https://10.13140/RG.2.2.29372.72329. *Co-equal first authors.
Liu, X., Wang, L., 2020. The impact of meaasurement error and omitting confounders on statistical inference of mediation effects and tools for sensitivity analysis. In: Psychological Methods. Advanced online publication. https://doi.org/10.1037/met0000345.
Maroulis, S., 2016. Interpreting school choice treatment effects:Results and implications from computational experiments. J. Artif. Soc. Soc. Simulat. 19 (1), 7.
Mauro, R., 1990. Understanding LOVE (left out variables error): a method for estimating the effects of omitted variables. Psychol. Bull. 108 (2), 314.
McCann, B.T., Schwab, A., 2020. Bayesian analysis in strategic management research: time to update your priors. Strategic Management Review.
Middleton, J.A., Scott, M.A., Diakow, R., Hill, J.L., 2016. Bias amplification and bias unmasking. Polit. Anal. 24 (3), 307–323.
Moody, J.W., Keister, L.A., Ramos, M.C., 2022. Reproducibility in the social sciences. Annu. Rev. Sociol. 48.
Morgan, S.L., Winship, C., 2007. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, Cambridge.

Murnane, R.J., Willett, J.B., 2010. Methods Matter: Improving Causal Inference in Educational and Social Science Research. Oxford University Press.
Neumayer, E., Plümper, T., 2017. Robustness Tests for Quantitative Research. Cambridge University Press.
Oakley, A., 1998. Experimentation and social interventions: a forgotten but important history. Br. Med. J. 317 (7176), 1239–1242.
Oaxaca, R., 1973. Male-female wage differentials in urban labor markets. Int. Econ. Rev. 693–709.
Oster, E., 2019. Unobservable selection and coefficient stability: theory and evidence. J. Bus. Econ. Stat. 37 (2), 187–204.
Pearl, J., 2009. Causality. Cambridge university press.
Plümper, T., Traunmüller, R., 2020. The sensitivity of sensitivity analysis. Political Science Research and Methods 8 (1), 149–159.
Raftery, A.E., 1995. Bayesian model selection in social research. Socio. Methodol. 111–163.
Robins, J.M., Rotnitzky, A., Scharfstein, D.O., 2000. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Statistical Models in Epidemiology, the Environment, and Clinical Trials. Springer, New York, NY, pp. 1–94.
Romer, D., 2020. In praise of confidence intervals. AEA Papers and Proceedings 110, 55–60.
Rosenbaum, P., 2002. Observational Studies. Springer, New York.
Rosenbaum, P.R., Rubin, D.B., 1983a. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J. Roy. Stat. Soc. B 45 (2), 212–218.
Rosenbaum, P.R., Rubin, D.B., 1983b. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.
Rosenbaum, P.R., 1986. Dropping out of high school in the United States: an observational study. J. Educ. Stat. 11 (3), 207–224.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and non_randomized studies. J. Educ. Psychol. 66, 688–701.
Rubin, D.B., 1986. Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference. J. Am. Stat. Assoc. 83, 396.
Rubin, D.B., 1990. Formal modes of statistical inference for causal effects. J. Stat. Plann. Inference 25, 279–292.
Scharfstein, D.O., Nabi, R., Kennedy, E.H., Huang, M.Y., Bonvini, M., Smid, M., 2021. Semiparametric Sensitivity Analysis: Unmeasured Confounding in Observational Studies arXiv preprint arXiv:2104.08300.
Schneider, Carnoy, B.M., Kilpatrick, J., Schmidt, W.H., Shavelson, R.J., 2007. Estimating Causal Effects Using Experimental and Observational Designs. AERA, publisher's information and FREE download at. http://www.aera.net/publications/Default.aspx?menu_id=46&id=3360.
Shadish, W.R., Clark, M.H., Steiner, P.M., 2008. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. J. Am. Stat. Assoc. 103 (484), 1334–1344.
Shepard, L.A., Smith, M.L., 1989. Flunking Grades. Falmer Press, New York.
Shepard, L.A., Smith, M.L., Marion, S.F., 1998. On the success of failure: a rejoinder to Alexander. Psychol. Sch. 35, 404–406.
Steiner, Peter M., Cook, Thomas D., William, R. Shadish, 2011. On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. J. Educ. Behav. Stat.
Steiner, Peter M., Cook, Thomas D., Shadish, William R., Clark, M.H., 2010. The importance of covariate selection in controlling for selection bias in observational studies. Psychol. Methods 15 (3), 250–267.
Tipton, E., 2014. How generalizable is your experiment? An index for comparing experimental samples and populations. J. Educ. Behav. Stat. 39 (6), 478–501.
Trafimow, D., Marks, M., 2015. Editorial. Basic Appl. Soc. Psychol. 37 (1), 1–2.
Thorndike, E.L., Woodworth, R.S., 1901. The influence of improvement in one mental function upon the efficiency of other functions. Psychol. Rev. 8, 247–261, 384-95, 553-261.
US Department of Health and Human Services, 2000. Trends in the Well-Being of America's Children and Youth. Washington, DC.
VanderWeele, T.J., Arah, O.A., 2011. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology 22 (1), 42–52.
VanderWeele, T.J., Ding, P., 2017. Sensitivity analysis in observational research: introducing the E-value. Ann. Intern. Med. 167 (4), 268–274.
Walsh, M., Srinathan, S.K., McAuley, D.F., Mrkobrada, M., Levine, O., Ribic, C., et al., 2014. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J. Clin. Epidemiol. 67, 622–628.
Walter, S.D., Thabane, L., Briel, M., 2020. The fragility of trial results involves more than statistical significance alone. J. Clin. Epidemiol. 124, 34–41.
Weiss, C.H., 1977. Research for policy's sake: the enlightenment function of social research. Pol. Anal. 531–545.
Wilkinson, L. and Task Force on Statistical inference (1999). Statistical methods in psychology journals: guidelines and explanations. Am. Psychol., 54, 594-604.
Wong, V.C., Valentine, J.C., Miller-Bains, K., 2017. Empirical performance of covariates in education observational studies. Journal of Research on Educational Effectiveness 10 (1), 207–236.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT press.
Xu, R., Frank, K.A., Maroulis, S.J., Rosenberg, J.M., 2019. konfound: command to quantify robustness of causal inferences. STATA J. 19 (3), 523–550.
Young, C., Holsteen, K., 2017. Model uncertainty and robustness: a computational framework for multimodel analysis. Socio. Methods Res. 46 (1), 3–40.
Young, C., 2018. Model uncertainty and the crisis in science. Socius 4.