A PRACTICAL GUIDE TO IMPACT THRESHOLD OF A CONFOUNDING VARIABLE (ITCV) AND ROBUSTNESS OF INFERENCE TO REPLACEMENT (RIR)

Ken Frank*
Michigan State University

Guan Saw*
Claremont Graduate University

Qinyun Lin
University of Gothenburg

Ran Xu
University of Connecticut

Joshua Rosenberg
University of Tennessee, Knoxville

Spiro Maroulis
Arizona State University

Bret Staudt Willet
Florida State University

**Equal first authors*

April 2025

--- Draft Copy ---

If you notice any errors in this report, please contact us at guan.saw@cgu.edu.
Your feedback is greatly appreciated

CONTENTS

SUGGESTED CITATION

Frank, K. A.*, Saw, G. K.*, Lin, Q., Xu, R., Rosenberg, J. M., Maroulis, S. J., & Willet, B. S. (2025). *A practical guide to impact threshold of a confouding variable (ITCV) and robustness of inference to replacement (RIR)*. Michigan State University. * co first authors.

ACKNOWLEDGMENTS

The authors thank Xuesen Cheng, Jihoon Choi, Yunhe Cui, Sarah Narvaiz, Dallin Overstreet, Wei Wang, Xukun Xiang, and Gaofei Zhang for providing outstanding research assistance.

This practical guide was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D220022 to Michigan State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

1. INTRODUCTION[1]

Social science research has the potential to inform and influence policy or public action (Bulmer, 2015; Burawoy, 2005; Weiss & Bucuvalas, 1980) with rigorous empirical evidence (National Research Council [NRC], 2012). The need for rigorous research evidence leads to the generation and communication of causal questions and results relevant to critical real-world issues (Frank et al., 2023; Schneider et al., 2007). Causal research evidence, even those produced from properly designed and implemented experimental or quasi-experimental studies, can be ambiguous and difficult to interpret.

Debates about the theoretical and methodological foundations for causal inferences in the social sciences date back to the 1900s (e.g., Rubin, 1974; Thorndike and Woodworth, 1901; see Oakley, 1998 for review). Concerns about inferences from non-experimental studies without randomized controlled trials, are particularly pronounced. Consider research on the impacts of eviction on individuals (e.g., economic hardship, health) and communities (e.g., homelessness rates, voter turnout) in which social scientists have employed an array of identification strategies, including coarsened exact matching, instrumental variable, and fixed effects, to address confounding factors and selection biases associated with being evicted (e.g., Hoke & Boen, 2021; Schwartz et al., 2022; Slee & Desmond, 2023; Treglia et al., 2023). The controversy primarily concerns the internal validity of the results: despite controlling for both observed and unobserved heterogeneity, how can we be certain that evicted individuals are being compared with similar others?

In quantitative research, sensitivity analysis represents one of the key approaches for quantifying and communicating uncertainty of estimated findings. Inspired by Cornfield et al.'s (1959) work on identifying alternative factors that could account for the estimated effect of smoking on lung cancer, sensitivity analyses have become a useful and widely used approach to assess and facilitate the discussion about the robustness of estimated effects in the fields of 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).

The document is intended to serve as a practical guide to implementing the Impact Threshold of a Confounding Variable (ITCV; Frank, 2000; Frank et al., 2013), a sensitivity analysis approach based on omitted variables, and Robustness of Inference to Replacement (RIR; Frank et al., 2013; Frank et al., 2021), a sensitivity analysis approach based on replacement of cases. As a first motivating example, we apply the ITCV to Desmond and Kimbro's (2015, p. 311) estimated effect of a recent eviction on women's material hardship using propensity score matching, which was 1.02 standard

[1] Part of this chapter is modified from Frank et al.'s (2023) paper published in *Social Science Research*.

deviations (standard error = 0.29; t = 3.52; p < 0.01; sample size = 122; number of covariates = 41). For this example, the ITCV generates the statement "To nullify the inference of the estimated effect of a recent eviction, an omitted variable would have to be correlated at 0.439 with eviction and with material hardship" (Frank, 2000). As a second motivating example, we apply the RIR to Yeager et al.'s (2019, p. 366) estimated effect of growth mindset intervention in a randomized controlled trial (RCT) on core course GPA among lower-achieving adolescents, which was 0.10 grade points (95% confidence interval = 0.04, 0.16; standard error = 0.03; sample size = 6,320; t = 3.51; p = 0.001). For this example, the RIR generates the statement "To nullify the inference of the estimated effect of growth mindset intervention, 41.19% of the cases (or 2,603 students) would have to be replaced with counterfactual cases with zero effect of the treatment" (Frank et al., 2013).

To set the stage for the rest of the practical guide, we will now briefly summarize the most important points about the ITCV and RIR:

- Sensitivity analyses are intended to inform dialogue about causal inferences, not to establish or nullify existing inferences.
- Sensitivity analysis in general and ITCV and RIR in particular should be used after the analyst has conducted the strongest model or set of models for causal inference, given the experimental or non-experimental data.
- There is no fixed "good cut-off" for ITCV and RIR across studies and fields. Using a fixed cut-off would pre-empt discourse about an inference instead of promoting discourse.
- We encourage analysts (1) to benchmark either the ITCV or RIR using observed covariates, and/or (2) to quantify the RIR for similar studies in the field (see examples in Frank et al., 2013)
- ITCV and RIR could be calculated and interpreted along with other analyses that estimate the sensitivity of estimates to observed covariates or alternative specifications.
- The ITCV and RIR also complement reporting of effect sizes and confidence intervals as well as *p*-values.

2. IMPACT THRESHOLD OF A CONFOUNDING VARIABLE (ITCV)[2]

2.1 Overview of the Impact Threshold of a Confounding Variable (ITCV)

The history of sensitivity analysis based on 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 & Hazlett, 2020). Nearly all draw on expressions of associations between the omitted variable and the focal predictor 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 & Glynn 2021; Hong & 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 Cochrane (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 & Hazlett, 2020; Dorie et al., 2016, have extended this to contour plots). See similar implications for binary outcomes (Rosenbaum, 2002; Harding, 2003; Vanderweele & Arah, 2011) and in a propensity score framework (Hirano & Imbens, 2001; Hong et al., 2021) and mediation framework (Imai et al., 2010).

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 Desmond and Kimbro's (2015) study of the effects of eviction on economic hardship. In Figure 1, the relationship of interest is between a recent eviction (X) and material hardship (Y), which could be impacted by an omitted confounding variable (e.g., loss of income, tenant-landlord disputes). Frank (2000) defines the impact of the confounding variable as $impact = r_{eviction \cdot cv} r_{hardship \cdot cv}$; the two components of confounding are resolved into a single term by taking the product.

[2] This chapter is modified from Frank et al.'s (2023) paper published in *Social Science Research*.

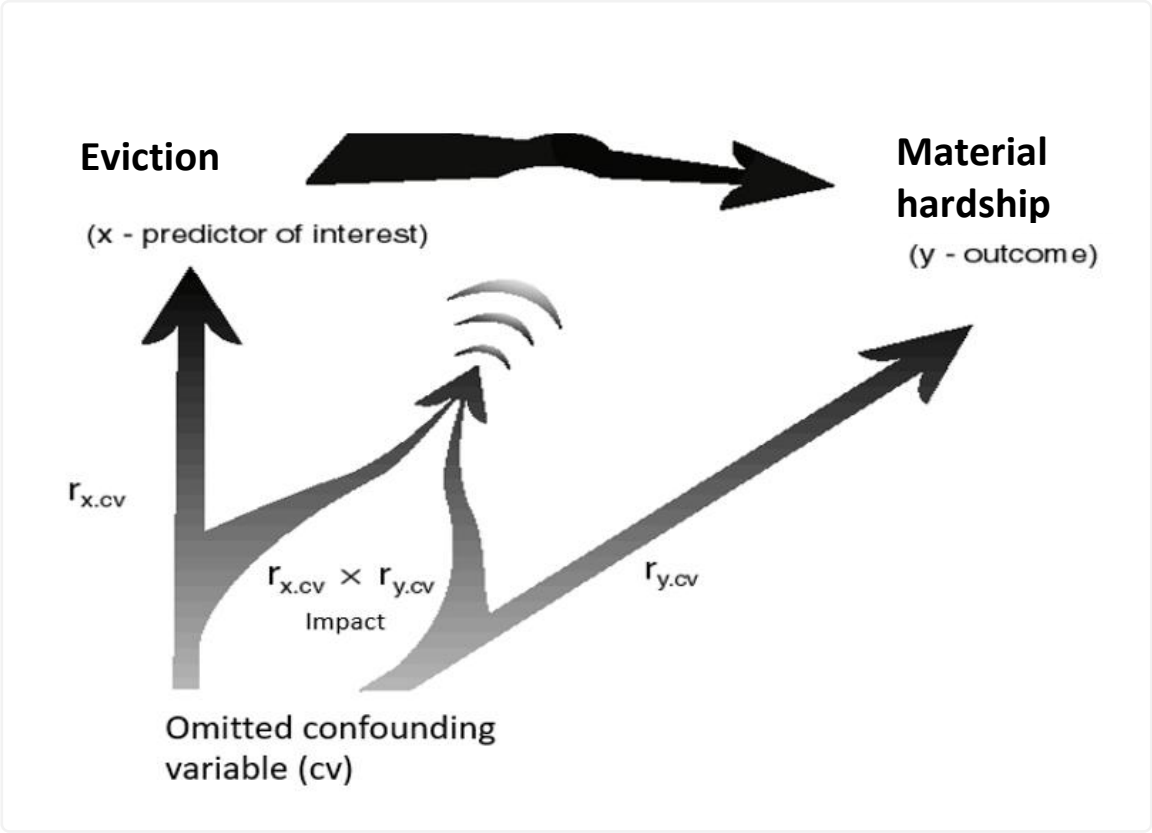


Figure 1. The impact of a confounding variable.

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 nullify an inference. Drawing on Figure 1, consider the model:

$$Hardship=\beta_0+\beta_1 Eviction, \quad (1)$$

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

$$Hardship=\beta_0+\beta_1 Eviction+\beta_2 IncomeLoss, \quad (2)$$

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

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 $\hat{\beta}_1$ 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 nullify an inference. Specifically, Frank (2000) expresses $\hat{\beta}_{1|CV}$ as a function of correlations associated with an omitted variable:

$$\hat{\beta}_{1|CV} = \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\frac{r_{X\cdot Y} - r_{Y\cdot CV}r_{X\cdot CV}}{1-r^2_{X\cdot CV}}, \quad (3a)$$

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 (3a). Importantly, Frank (2000) also expresses how an omitted variable can affect the standard error representing sampling variability and used for inference:

$$se(\hat{\beta}_{1|CV}) = \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \times \sqrt{\frac{1-R^2_{Y\cdot X}}{n-q-1} \times \frac{1}{1-r^2_{X\cdot CV}}}$$
$$= \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \times \sqrt{\frac{\frac{1-(r^2_{X\cdot Y}+r^2_{Y\cdot CV}-2r_{X\cdot Y}r_{Y\cdot CV}r_{X\cdot CV})}{1-r^2_{X\cdot CV}}}{n-q-1} \times \frac{1}{1-r^2_{X\cdot CV}}}, \quad (3b)$$

where $\hat{\sigma}$ 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 $\hat{\beta}_1$ as well as its standard error.

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

$$r_{x\cdot y|cv} = \frac{t_{x\cdot y|cv}}{\sqrt{df + t_{x\cdot y|cv}}}, \text{ where } t_{x\cdot y|cv} = \frac{\hat{\beta}_{1|CV,Z}}{se(\hat{\beta}_{1|CV,Z})}. \quad (4)$$

The partial correlation, $r_{x\cdot y|cv}$ can be expressed as (Cohen et al., 2014):

$$r_{x\cdot y|cv} = \frac{r_{x\cdot y} - r_{x\cdot cv} \times r_{y\cdot cv}}{\sqrt{1-r_{y\cdot cv}^2}\sqrt{1-r_{x\cdot cv}^2}}. \quad (5)$$

Similar to Frank (2000) for (3a) and (3b), Xu et al. (2019) show the maximum for (5) 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. [3]

$$\text{If } r_{x\cdot cv}=r_{y\cdot cv} \text{ then } impact = r_{x\cdot cv} \times r_{y\cdot cv} = r_{x\cdot cv}^2 = r_{y\cdot cv}^2.$$

Substituting *impact* for $r_{x\cdot cv} \times r_{y\cdot cv}, r_{x\cdot cv}^2$, and $r_{y\cdot cv}^2$ in (5) yields:

$$r_{xy|cv} = \frac{r_{x\cdot y} - impact}{1-|impact|}. \quad (6)$$

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

$$\text{Impact} \geq \frac{r_{x\cdot y} - r^{\#}}{1-|r^{\#}|}. \quad (7)$$

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 (7) supports intuition about sensitivity. Specifically, sensitivity about an inference is based on the 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.

[3] 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.

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,

$$r^{\#} = \frac{t_{critical}}{\sqrt{df + t^2_{critical}}}, \quad (8)$$

where *df* is the degrees of freedom used to test $\hat{\beta}_1$. Because the inference for the regression coefficient in (2) is identical to that for the partial correlation in (5), the expressions in (7) and (8) 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 $r^{\#}$, $\hat{\beta}_1$ would not be statistically different from zero if the omitted variable were included in the model.

*** A further reading list on ITCV can be found on KonFound-It! Website resources page.

2.2 Application of the ITCV: A Step-by-Step Guide

The calculations of ITCV can be performed with,

(1) a Shiny app KonFound-it! At https://konfound-project.shinyapps.io/konfound-it/,
(2) Konfound commands in R software,
(3) Konfound commands in Stata software, or
(4) a Konfound-it! Spreadsheet (in Microsoft Excel)

To compute the ITCV for an estimated effect in a linear model (i.e., regression), a researcher will need the four following values from the data or estimated model: (1) estimated coefficient for the predictor of interest, (2) standard error, (3) sample size, and (4) number of covariates.

As an example of how to apply the ITCV, consider Desmond and Kimbro's (2015) estimated effect of a recent eviction on women's material hardship using propensity score matching, which was 1.02 standard deviation (standard error = 0.29; t = 3.52; p < 0.01; sample size = 122; number of covariates = 41; see the abstract, Table 2, and results interpretation of the paper below).

Eviction's Fallout: Housing, Hardship, and Health

Matthew Desmond, *Harvard University*
Rachel Tolbert Kimbro, *Rice University*

Millions of families across the United States are evicted each year. Yet, we know next to nothing about the impact eviction has on their lives. Focusing on low-income urban mothers, a population at high risk of eviction, this study is among the first to examine rigorously the consequences of involuntary displacement from housing. Applying two methods of propensity score analyses to data from a national survey, we find that eviction has negative effects on mothers in multiple domains. Compared to matched mothers who were not evicted, mothers who were evicted in the previous year experienced more material hardship, were more likely to suffer from depression, reported worse health for themselves and their children, and reported more parenting stress. Some evidence suggests that at least two years after their eviction, mothers still experienced significantly higher rates of material hardship and depression than peers.

Table 2. Effects of a Recent Eviction (child age 4–5) on Maternal and Child Wellbeing Outcomes at Child Age 5

	Propensity score weighting ($N = 2,676$)		Propensity score matching ($N = 122$)		
	Model 1	Model 2	Model 3	Model 4	Model 5
	No shocks	With shocks	No shocks	With shocks	Regression adjusted, with shocks
Outcome	Coefficient		ATT		
Material hardship	0.99*** (0.16)	0.96*** (0.16)	1.06*** (0.23)	1.03*** (0.24)	1.02** (0.29)
Poverty ratio	–0.35** (0.11)	–0.30** (0.11)	–0.38 (0.31)	–0.34 (0.31)	–0.35 (0.33)
Parenting stress	1.19** (0.39)	1.18** (0.38)	1.42* (0.64)	1.45* (0.68)	1.41† (0.73)

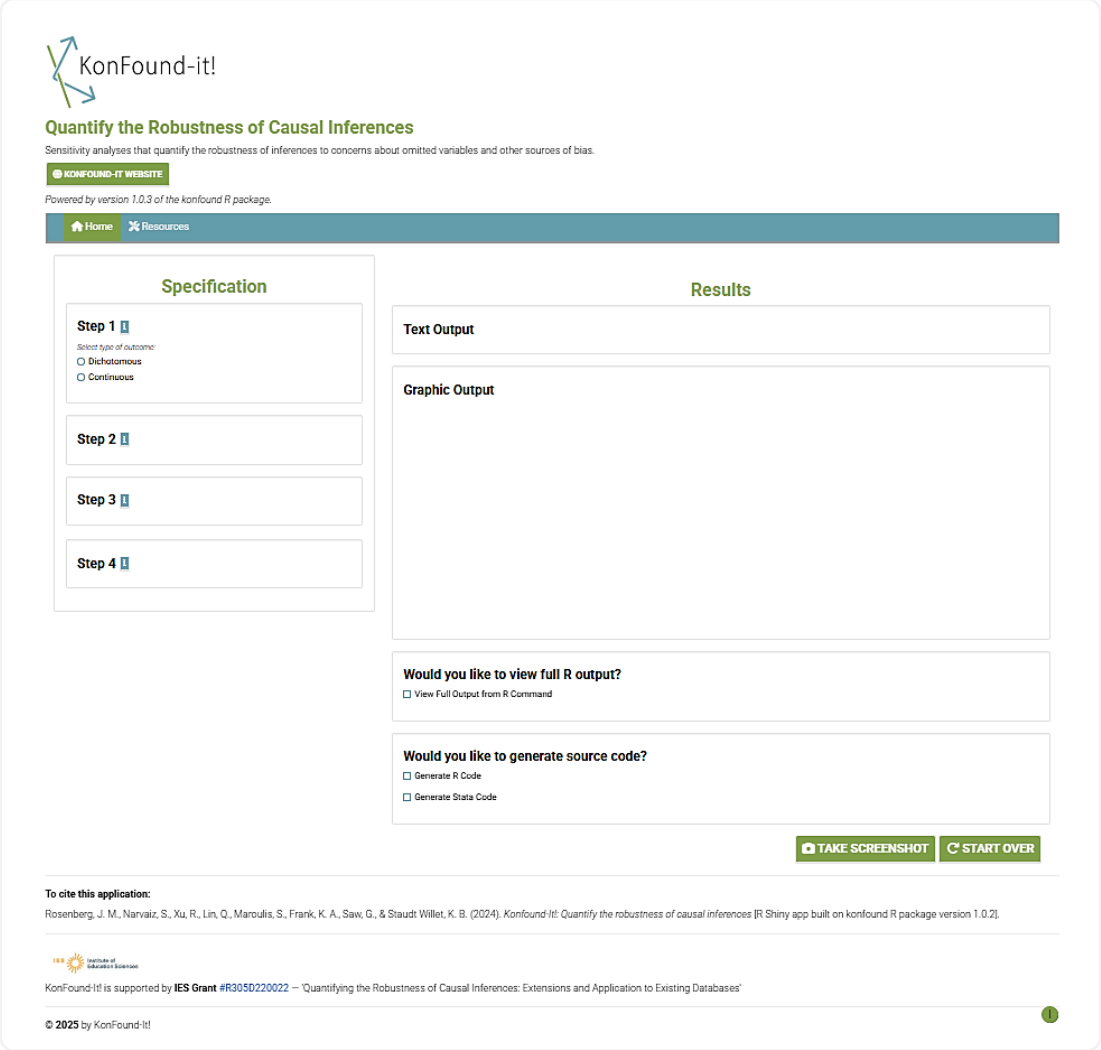
Effects of a Recent Eviction

We turn first to results estimating the effect of a recent eviction on the wellbeing of mothers and children when the focal child is 5 (see table 2). Across all models, there is a large and robust relationship between a recent eviction and material hardship. Regardless of the estimation technique, respondents who experienced an eviction in the past year report around one standard deviation higher material hardship. We found eviction to be associated with reductions in the income-to-poverty ratio, although this relationship becomes insignificant in ATT models

2.2.1 Computing ITCV with KonFound-it! Shiny App

2.2.1.1 Access

To use the KonFound-it! Shiny app, go to https://konfound-project.shinyapps.io/konfound-it/. As of the release of this practical guide, the KonFound-it! Shiny app is built with version 1.0.3 of the konfound R package, which will be updated over time.



2.2.1.2 Calculating ITCV

To calculate the ITCV, follow the steps illustrated below:

Step 1

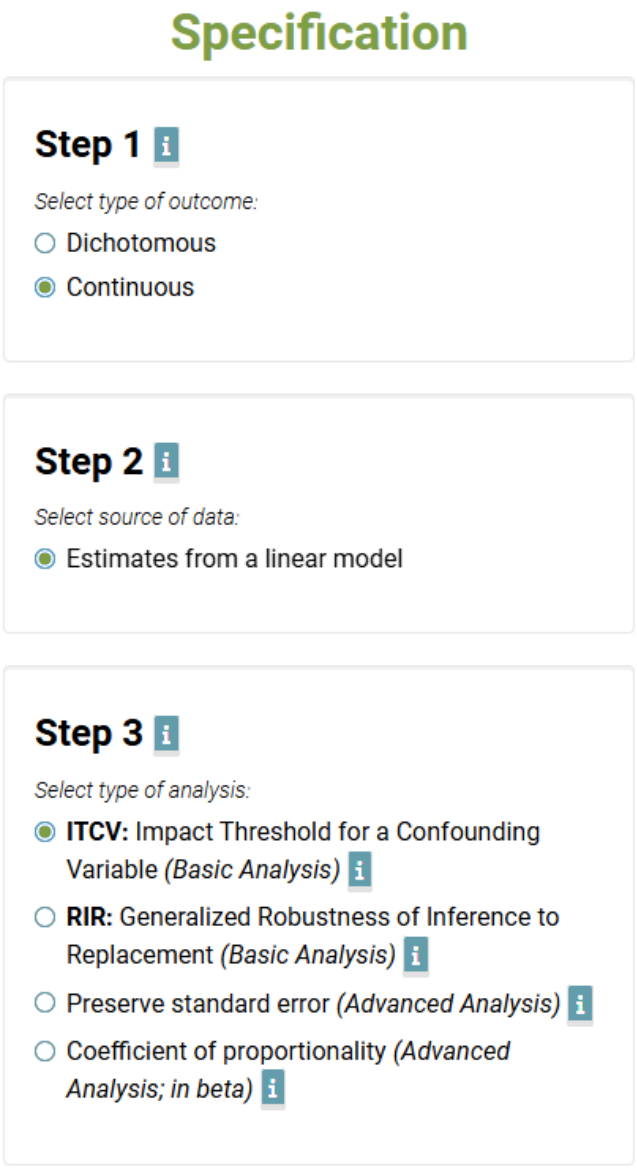
Choose the option of "Continuous" as the type of outcome

Step 2

The option of "Estimates from a linear model" will be automatically chosen

Step 3

Choose the option of "ITCV: Impact Threshold for a Confounding Variable *(Basic Analysis)*"



Step 4.1

Enter the coefficient for the predictor of interest

Step 4.2

Enter the standard error of the estimated effect

Step 4.3

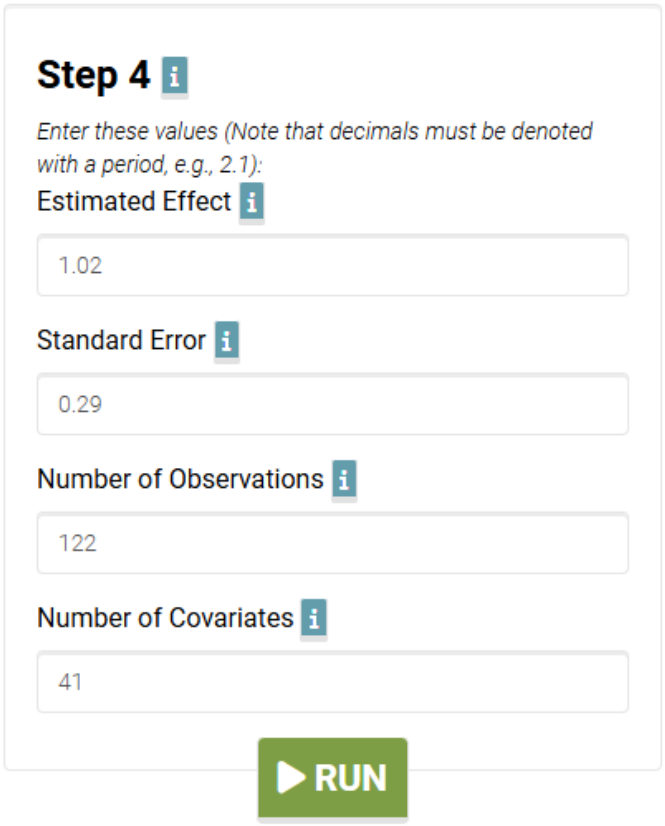
Enter the number of observations (or sample size) of the estimated model

Step 4.4

Enter the number of covariates included in the model other than the predictor of interest

Step 4.5

Click “RUN”



2.2.1.3 Output and Interpretation

Text Output

Presents a statement interpreting the calculated ITCV

Graphic Output

Presents an illustration that displays the (1) correlation between an omitted confounding variable and the predictor of interest, (2) correlation between an omitted confounding variable and the outcome, and (3) product of the two correlations

Results

Text Output

Impact Threshold for a Confounding Variable (ITCV):

The minimum impact of an omitted variable needed to nullify an inference for a null hypothesis of 0 (nu) is based on correlations of 0.437 with the outcome and 0.437 with the predictor of interest (conditioning on all observed covariates in the model; signs are interchangeable if they are different).
This is based on a threshold effect of 0.219 for statistical significance (alpha = 0.05).

Correspondingly, the impact of an omitted variable (Frank 2000) must be 0.437 × 0.437 = 0.191 to nullify the inference.

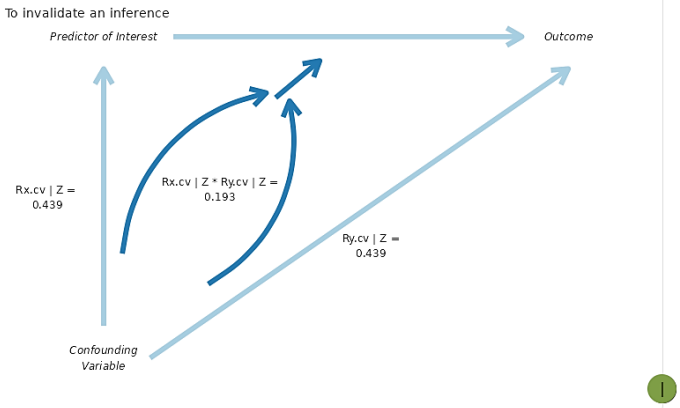
See Frank (2000) for a description of the method.

Citation:
Frank, K. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research, 29*(2), 147-194.

Accuracy of results increases with the number of decimals reported.
The ITCV analysis was originally derived for OLS standard errors. If your standard errors are not OLS-based, interpret the ITCV with caution.
This analysis assumes the use of default parameters. For greater flexibility, use the R or Stata versions of the konfound package, beginning with the advanced code provided below on this page.

Calculated with konfound R package version 1.0.3

Graphic Output



Suggested Interpretation:

A suggested statement for interpreting the calculated ITCV of Desmond and Kimbro's (2015) estimated effect of a recent eviction on women's material hardship reads: "to nullify the inference of the estimated effect of a recent eviction on material hardship (coefficient = 1.02; standard error = 0.29; $p < 0.01$; sample size = 122; number of covariates = 41), an omitted variable would have to be correlated at 0.439 with the eviction and with the material hardship" (Frank, 2000).

*** Other published empirical examples with ITCV interpretation can be found on KonFound-It! Website resources page.

2.2.1.4 Generating R and Stata codes

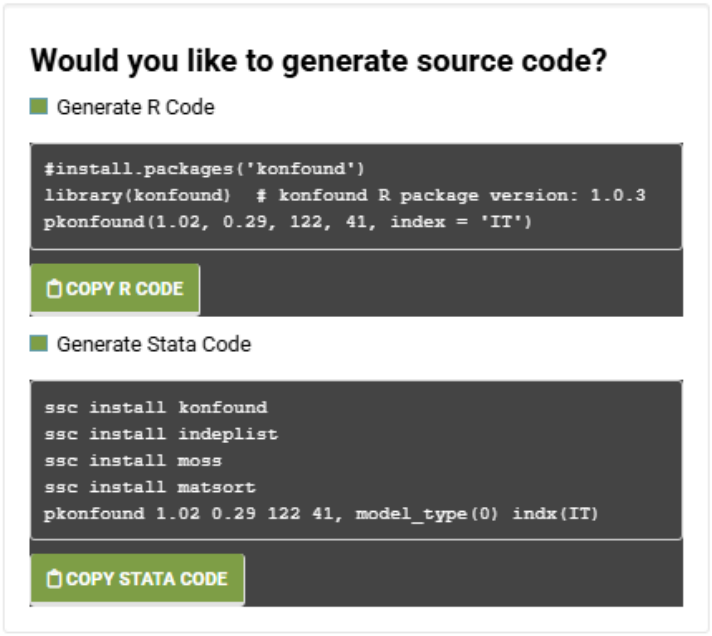
To generate R and Stata codes:

Generate R Code

To generate R codes, check the box on the left of “Generate R Code”

Generate Stata Code

To generate Stata codes, check the box on the left of “Generate Stata Code”



2.2.2 Computing ITCV with R Software

2.2.2.1 Installation

To install the latest CRAN version of konfound (April 2025):

```
install.packages("konfound")
```

To install the development version from GitHub (including new features possibly in beta mode):

```
install.packages("devtools")
devtools::install_github("jrosen48/konfound")
```

2.2.2.2 Calculating ITCV

To calculate the ITCV by manually entering results using long-form code:

```
library(konfound)
pkonfound(est_eff = 1.02,
          std_err = 0.29,
          n_obs = 122,
          n_covariates = 41,
          index = 'IT')
```

To calculate the ITCV by manually entering results using short-form code:

```
pkonfound(1.02, 0.29, 122, 41, index = 'IT')
```

2.2.2.3 Output and Interpretation

R output (the same for both long- and short-form approach):

```
Impact Threshold for a Confounding Variable (ITCV):
The minimum impact of an omitted variable to nullify an inference
for a null hypothesis of an effect of 0 (nu) is based on a
correlation of 0.437 with the outcome and 0.437 with the predictor
of interest (conditioning on all observed covariates in the model;
signs are interchangeable if they are different). This is based on
a threshold effect of 0.219 for statistical significance (alpha =
0.05).

Correspondingly the impact of an omitted variable (as defined in
Frank 2000) must be 0.437 X 0.437 = 0.191 to nullify an inference
for a null hypothesis of an effect of 0 (nu).

For calculation of unconditional ITCV using pkonfound(),
additionally include the R2, sdx, and sdy as input.
```

See Frank (2000) for a description of the method.

Citation:
Frank, K. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research, 29*(2), 147-194.

Accuracy of results increases with the number of decimals reported.

The ITCV analysis was originally derived for OLS standard errors. If the standard errors reported in the table were not based on OLS, some caution should be used to interpret the ITCV.

2.2.3 Computing ITCV with Stata Software

2.2.3.1 Installation

To install the Stata konfound command:

```
ssc install konfound
ssc install indeplist
ssc install moss
ssc install matsort
```

2.2.3.2 Calculating ITCV

To calculate ITCV by manually entering results:

```
pkonfound 1.02 0.29 122 41, indx("IT")
```

2.2.3.3 Output and Interpretation

Stata output:

```
Impact Threshold for a Confounding Variable (ITCV):

The minimum impact of an omitted variable to nullify the inference for a
null hypothesis of an effect of 0 (nu) is based on a correlation of 0.437
with the outcome and 0.437 with the predictor of interest (conditioning
on all observed covariates in the model; signs are interchangeable if they
are different). This is based on a threshold effect of 0.219 for
statistical significance (alpha = 0.050).

Correspondingly the impact of an omitted variable (as defined in Frank
2000) must be 0.437 X 0.437 = 0.191 to nullify the inference for a null
hypothesis of an effect of 0 (nu).
For calculation of unconditional ITCV, include the rs (for R2), sdx and
sdy as input and include 'return list' following the pkonfound command.

See Frank (2000) for a description of the method.
Citation:
Frank, K. (2000). Impact of a confounding variable on the inference of a
regression coefficient. Sociological Methods and Research, 29 (2), 147-
194.

Accuracy of results increases with the number of decimals reported.

The ITCV analysis was originally derived for OLS standard errors. If the
standard errors reported in the table were not based on OLS, some caution
should be used to interpret the ITCV.
```

2.2.4 Computing ITCV with Konfound-it! Spreadsheet

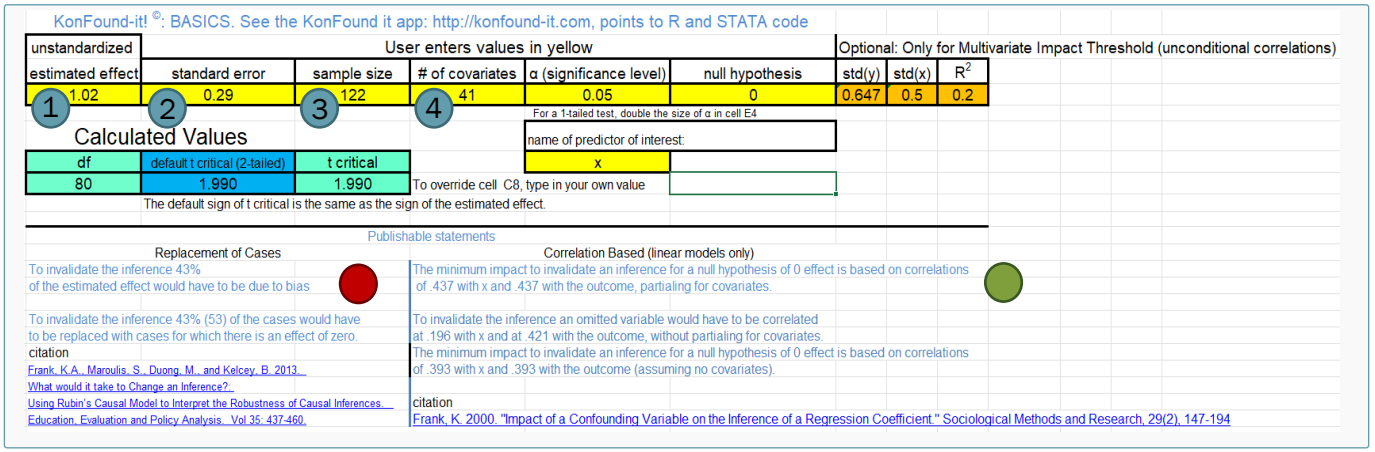
2.2.4.1 Download

Go to KonFound-it! Website Resources page and download the KonFound-it! spreadsheet for calculating indices.

2.2.4.2 Calculating ITCV, Output, and Interpretation

Follow the steps illustrated below to calculate the ITCV and obtain the output and interpretation:

1. Enter the estimated effect
2. Enter the standard error
3. Enter the sample size
4. Enter the number of covariates



IGNORE this set of output for the ITCV analysis

USE this set of output for the ITCV analysis

2.2.5 Calculating ITCV for Models Fitted in R

2.2.5.1 Dataset

This example uses the `concord1` dataset built into the `konfound` package. See the description of the dataset and procedure in Narvaiz et al. (2024).

2.2.5.2 Fitting a Linear Model in R

Below is the code to fit a linear model using the `concord1` variables:

```
m <- lm(water81 ~ water80 + income + educat + retire + peop80, data = concord1)
```

2.2.5.3 Calculating ITCV

To calculate ITCV for `peop80`:

```
library(konfound)

konfound(m, peop80, index = "IT")
```

2.2.5.4 Output and Interpretation

R output:

```
Impact Threshold for a Confounding Variable (ITCV):

The Unconditional ITCV:
The minimum impact of an omitted variable to nullify an inference
for a null hypothesis of an effect of 0 (nu) is based on a
correlation of 0.319 with the outcome and 0.424 with the
predictor of interest (BEFORE conditioning on observed
covariates; signs are interchangeable if they are different).
This is based on a threshold effect of 0.088 for statistical
significance (alpha = 0.05).

Correspondingly the UNCONDITIONAL impact of an omitted variable
(as defined in Frank 2000) must be 0.319 X 0.424 = 0.135 to
nullify an inference for a null hypothesis of an effect of 0
(nu).

Conditional ITCV:
The minimum impact of an omitted variable to nullify an inference
for a null hypothesis of an effect of 0 (nu) is based on a
correlation of 0.519 with the outcome and 0.519 with the
predictor of interest (conditioning on all observed covariates in
the model; signs are interchangeable if they are different). This
is based on a threshold effect of 0.088 for statistical
significance (alpha = 0.05).

Correspondingly the impact of an omitted variable (as defined in
Frank 2000) must be 0.519 X 0.519 = 0.269 to nullify an inference
for a null hypothesis of an effect of 0 (nu).
```

Interpretation of Benchmark Correlations for ITCV:
Benchmark correlation product ('benchmark_corr_product') is Rxz*Ryz = 0.2082, showing the association strength of all observed covariates Z with X and Y.

The ratio ('itcv_ratio_to_benchmark') is unconditional ITCV/Benchmark = 0.1350/0.2082 = 0.6481, indicating the robustness of inference.

The larger the ratio the stronger must be the unobserved impact relative to the impact of all observed covariates to nullify the inference. The larger the ratio the more robust the inference.

If Z includes pretests or fixed effects, the benchmark may be inflated, making the ratio unusually small. Interpret robustness cautiously in such cases.

See Frank (2000) for a description of the method.

Citation:
Frank, K. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research, 29*(2), 147-194.

Accuracy of results increases with the number of decimals reported.

The ITCV analysis was originally derived for OLS standard errors. If the standard errors reported in the table were not based on OLS, some caution should be used to interpret the ITCV.NULL

2.2.6 Calculating ITCV for Models Fitted in Stata

2.2.6.1 Dataset

This example uses the `concord1` dataset which can be downloaded by using the following Stata code (see the description of the dataset and procedure in Ran et al. [2019]):

```
use https://stats.idre.ucla.edu/stat/stata/examples/rwg/concord1, clear
```

2.2.6.2 Fitting a Linear Model in Stata

Below is the code to fit a linear model using the `concord1` variables:

```
regress water81 water80 income educat retire peop80
```

2.2.6.3 Calculating ITCV

To calculate ITCV for `peop80`:

```
konfound peop80, indx(IT)
```

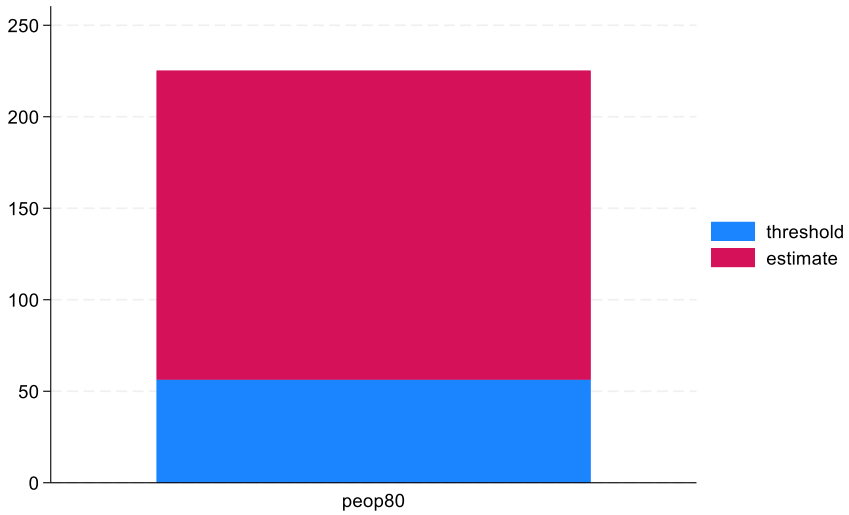
2.2.6.4 Output and Interpretation

Stata text output:

```
For variable peop80

Impact Threshold for a Confounding Variable (ITCV)

Unconditional ITCV:
The minimum impact of an omitted variable to nullify an inference
for a null hypothesis of an effect of 0 is based on a correlation of
0.319 with the outcome and 0.424 with the predictor of interest
(BEFORE conditioning on observed covariates; signs are
interchangeable if they are different). This is based on a threshold
effect of .088 for statistical significance (alpha = .05).

Correspondingly, the UNCONDITIONAL impact of an omitted variable (as
defined in Frank 2000) must be 0.319 × 0.424 = .135 to nullify an
inference for a null hypothesis of an effect of 0 (nu).

Conditional ITCV:
The minimum impact of an omitted variable to nullify an inference
for a null hypothesis of an effect of 0 is based on a correlation
of .519 with the outcome and .519 with the predictor of interest
(conditioning on all observed covariates in the model; signs are
interchangeable if they are different). This is based on a threshold
effect of .088 for statistical significance (alpha = .05).
```

```
Correspondingly, the impact of an omitted variable (as defined in
Frank 2000) must be .519 × .519 = .269 to nullify an inference for a
null hypothesis of an effect of 0 (nu).

For exact values calculated by ITCV, include 'return list' following
the konfound command.

These thresholds can be compared with the impacts of observed
covariates below.

Observed Impact Table for peop80

+-------------------------------------------------------------------+
|          Raw |         Cor(vX) |         Cor(vY) |          Impact |
|--------------+-----------------+-----------------+-----------------|
|      water80 |        0.533900 |        0.764800 |        0.408300 |
|       income |        0.284500 |        0.417800 |        0.118800 |
|       retire |       -0.358400 |       -0.273100 |        0.097900 |
|       educat |        0.057100 |        0.040400 |        0.002300 |
+-------------------------------------------------------------------+

+-------------------------------------------------------------------+
|      Partial |         Cor(vX) |         Cor(vY) |          Impact |
|--------------+-----------------+-----------------+-----------------|
|      water80 |        0.458000 |        0.726000 |        0.332500 |
|       income |        0.071400 |        0.286800 |        0.020500 |
|       educat |       -0.054500 |       -0.156700 |        0.008500 |
|       retire |       -0.225000 |       -0.006600 |        0.001500 |
+-------------------------------------------------------------------+

X represents peop80, Y represents water81, v represents each covariate.
First table is based on raw (unconditional) correlations, second table is
based on partial (conditional) correlations.

konfound command should only be run immediately after a model is estimated.
No other commands should be entered between estimating the model and running
konfound.

See Frank et al. (2013) for a description of the method.

Citation: Frank, K.A., Maroulis, S., Duong, M., and Kelcey, B. (2013).
What would it take to change an inference? Using Rubin's causal model to
interpret the robustness of causal inferences. Educational Evaluation and
Policy Analysis, 35, 437-460.
```

Stata graphic output



2.3 Benchmarks for the ITCV (Correlations associated with Observed Variables) of ITCV

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; Rosenbuam, 1986). To begin, for models such as (2) that already include an observed covariate, z, the expression in (7) generates the ITCV|z. This can be converted to an expression that is a function of zero-order (unadjusted) correlations (under the assumption that $r_{z\cdot cv}$=0):

$$ITCV \mid z = \frac{r_{x\cdot y|z} - r^{\#}}{1 - |\, r^{\#} \,|} = r_{y\cdot cv|z} r_{x\cdot cv|z} = \frac{r_{y\cdot cv} r_{x\cdot cv}}{\sqrt{\left(1 - r_{y\cdot z}^2\right)\left(1 - r_{x\cdot z}^2\right)}} = \frac{ITCV}{\sqrt{\left(1 - r_{y\cdot z}^2\right)\left(1 - r_{x\cdot z}^2\right)}} \text{ . (9)}$$

$$\Rightarrow ITCV = ITCV \mid z \sqrt{\left(1 - r_{y\cdot z}^2\right)\left(1 - r_{x\cdot z}^2\right)}$$

Then the ITCV can be expressed relative to the impact for an observed benchmark covariate z[4]:

$$ITCV(benchmark) = \frac{ITCV}{r_{y\cdot z} r_{x\cdot z}} = \frac{r_{y\cdot cv} r_{x\cdot cv}}{r_{y\cdot z} r_{x\cdot z}}. \text{ (10)}$$

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 nullify 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).

For example, Hong and Raudenbush (2005) analyzed a sample size of 7,639 students with 221 covariates to estimate the kindergarten retention effects (for 7,168 promoted students, 471 retained students). The results showed that the expected reading achievement for a retained student would be 9.01 points lower at the end of the treatment year, with a standard error of 0.68 and an observed t-ratio of -13.27 (Hong & Raudenbush, 2005, p. 217). An omitted confounding variable would have to have an impact of $r_{x\cdot cv}$ x $r_{y\cdot cv}$=-.132, with component correlations of $.132^{1/2}$=.36 (taking opposite signs) to result in a partial correlation of −.023 (associated with a p-value of .05). Correspondingly, if an omitted variable had an impact greater in magnitude than .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 (3) through (8).

[4] 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 $r_{x\cdot cv}$ / $r_{y\cdot z}$, with their specification of the maximum R^2 from (2) implying a value of $r_{y\cdot cv}$ (Frank et al.,2022).

TABLE 11
Model-Based Estimation of Kindergarten Retention Effect on Reading in Retention Schools

Fixed effect		Coefficient	*SE*	*t*
Retention school promoted at-risk kid intercept, γ		53.99	0.28	192.15
Retention effect in retention schools, δ_Z		−9.01	0.68	−13.27
Random effect	**Variance**	***df***	**χ^2**	***p* value**
School mean intercept, u_j	55.02	230	1,863.46	.000
School retention effect, $\Delta_{Z.uj}$	18.83	230	280.30	.013
Correlation between u_j and $\Delta_{Z.uj}$		−0.27		
Level-1 effect, e_{ij}	88.22			

Note. For brevity, we have omitted in this table the coefficient estimates for the covariates including the propensity strata, the logit of propensity score, and the duration of treatment up to the time of assessment.

In the study of Hong and Raudenbush (2006), the strongest covariate identified is "student approaches to learning (SAL)" and its impact is $r_{SAL,retention}r_{SAL,achievement}=(-.1849)(.4442)=-.08$. Correspondingly, the impact of an omitted variable would have to be −.132/(−.08)=1.65 or 65% stronger than the impact of the strongest covariate to change the inference.

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

$$\text{outcome}=\beta 0+\beta 1\text{treatment}+B'Z. \quad (11)$$

Frank (2000) shows that the expression in (10) can be generalized for the model in (11):

$$ITCV(benchmark)=\frac{r_{y\cdot cv}r_{x\cdot cv}}{R_{y\cdot Z}R_{x\cdot Z}}, \quad (12)$$

where $R_{Y\cdot \mathbf{z}}$ is the multiple correlation between *Y* and the vector of covariates **z** ($0<R_{Y\cdot Z}<1$), $R_{X\cdot Z}$ is the multiple correlations between X and the vector of covariates z ($0<R_{X\cdot Z}<1$). Knaeble and Dutter (2017) and Knaeble et al. (2020) show that the OLS estimate for β_1 for the model in (11) is

$$\hat{\beta}_1=\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\frac{r_{X\cdot Y}-R_{X\cdot \mathbf{Z}}R_{Y\cdot \mathbf{Z}}\rho_{\hat{X}\hat{Y}}}{1-R^2_{X\cdot \mathbf{Z}}}, \quad (13)$$

where $\rho_{\hat{X}\hat{Y}}$ is the correlation between $\hat{X}$ (the predicted value from regressing *X* on the elements in **Z**) and $\hat{Y}$ (the predicted value from regressing *Y* on the elements in **Z**), with $-1<\rho_{\hat{X}\hat{Y}}<1$. The corresponding partial correlation is

$$r_{X\cdot Y|\mathbf{Z}}=\frac{r_{X\cdot Y}-R_{X\cdot \mathbf{Z}}R_{Y\cdot \mathbf{Z}}\rho_{\hat{X}\hat{Y}}}{\sqrt{1-R^2_{Y\cdot \mathbf{Z}}}\sqrt{1-R^2_{X\cdot \mathbf{Z}}}}. \quad (14)$$

Therefore, under the assumption that $\rho_{\hat{X}\hat{Y}}$ =1 (the strength of a covariate's prediction of X corresponds to its strength of prediction of *Y*), the expression in (10) can be generalized for the model in (11) using the expression in (12). Note that the terms $R_{y\cdot Z}$ and $R_{x\cdot Z}$ can be obtained directly from the overall R^2 from (11) and $\hat{\beta}_1$, $se(\hat{\beta}_1)$, $\hat{\sigma}_X$ and $\hat{\sigma}_Y$. Thus, one can benchmark using one, some, or all covariates in Z (Lonati & Wulff, 2025). 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, Clark, and Steiner, 2008; Steiner et al., 2010, 2011; see review in Wong, Valentine, and Miller-Bains 2017).

Assuming that an omitted variable is independent of observed covariates, the unconditional $r_{cv\cdot y}$ and $r_{cv\cdot y}$ can be expressed as:

$$r_{y\cdot cv} = r_{y\cdot cv|z}\sqrt{\left(1-R^2_{y\cdot z}\right)\left(1-R^2_{cv\cdot z}\right)} + \sqrt{R^2_{y\cdot z}R^2_{cv\cdot z}} = r_{y\cdot cv|z}\sqrt{\left(1-R^2_{y\cdot z}\right)} = \sqrt{ITCV_Z\left(1-R^2_{y\cdot z}\right)} \quad (15)$$

$$r_{x\cdot cv} = r_{x\cdot cv|z}\sqrt{\left(1-R^2_{x\cdot z}\right)\left(1-R^2_{cv\cdot z}\right)} + \sqrt{R^2_{x\cdot z}R^2_{cv\cdot z}} = r_{x\cdot cv|z}\sqrt{\left(1-R^2_{x\cdot z}\right)} = \sqrt{ITCV_Z\left(1-R^2_{x\cdot z}\right)} \quad (16)$$

3. ROBUSTNESS OF INFERENCE TO REPLACEMENT (RIR)[5]

3.1 Overview of RIR

An alternative to expressing confounding in terms of the dual components rx·cv ry·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 of Desmond and Kimbro (2015), it is impossible to observe a mother who is simultaneously evicted and not evicted.

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., economic hardship) that would be observed if unit *i* were exposed to the treatment (e.g., being evicted); 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., not being evicted). If the Stable Unit Treatment Value Assumption (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

$$\delta_i = Y_i^t - Y_i^c . \quad (17)$$

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 non-random sampling. 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 & Rubin, 1983) to graphical representations (e.g., Cinelli & Hazlett, 2020; Imbens, 2003), to computation of the properties of covariates (e.g., Jesson et al., 2021; Kallus et al., 2019) 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

[5] This chapter is modified from Frank et al.'s (2023) paper published in *Social Science Research* and Frank et al.'s (2021) paper published in *Journal of Clinical Epidemiology*.

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 & Rubin, 1983).

Here we leverage the Robustness of Inference to Replacement (RIR; Frank et al., 2013; Frank et al., 2021) 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 nullify, or undo, the inference. The more the estimate exceeds the threshold, the more robust the inference with respect to that threshold

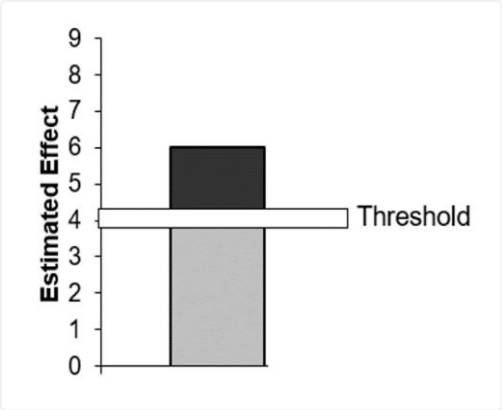


Figure 2: Estimated effect relative to a threshold for inference.

Consider the idealized example in Figure 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 x 2 = 3.92 (where 1.96 represents the critical value of a t-distribution for probability level of .05 for a two-tailed test and sample size of 260). Thus, we draw the threshold for inference at 3.92, or about a value of 4 shown on the graph in Figure 2.

Frank et al.'s, (2013) interpretation of Figure 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 nullify 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 nullify an inference in terms of replacing observed cases with counterfactual cases where a null hypothesis of zero treatment effect held. Specifically, one would expect to 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 ($\bar{\hat{\delta}}$) as a function of the observed estimated effect ($\hat{\delta}_o$) and the hypothesized effect in the unobserved (e.g., counterfactual) replacement data (δ_u). See Cronbach (1982) or Frank & Min (2007) for details. Assuming the proportion of units receiving the treatment is the same in the observed and unobserved data, an expression for $\bar{\hat{\delta}}$ is:

$$\bar{\hat{\delta}} = (1\text{-}\pi)\hat{\delta}_O + \pi\delta_U \text{ , (18)}$$

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, $\bar{\hat{\delta}}$ is a mixture, according to π, of $\hat{\delta}_o$ 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 & Hazlett, 2020; Frank et al., 2013; VanderWeele & 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., 2013). Or, $\delta_u = 0$ holds if there are no variables confounded with treatment and outcome. Then set $\bar{\hat{\delta}} = \delta^{\#}$ where $\delta^{\#}$ defines the threshold for making an inference (such as an estimate associated with an effect size of specific clinical significance; Angst, Aeschlimann, & Angst, 2017) or with a specific p-value (e.g., .05) and finally solving for π yields:

$$\pi = 1 - \frac{\delta^{\#}}{\bar{\hat{\delta}}_O} = \text{Robustness of Inference to Replacement (RIR) . (19)}$$

The closed form expression in (19) 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 ($\bar{\hat{\delta}}$) below the threshold ($\delta^{\#}$) for making an inference (Frank et al., 2013). For instance, in the simple example in Figure 2 where $\hat{\delta}_o = 6$ 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.

There are two important points to make about the RIR. First, by conceptualizing robustness in terms of how much of the data would have to be replaced to change the inference, the RIR quantifies the robustness of the inference in terms of experiences of people expressed as potential outcomes. Second, the RIR is essentially non-parametric in that it applies regardless of the functional form relating the covariate to the outcome or to the treatment. As such, the RIR has been extended to a broad class of models across the social sciences including propensity models (Frank et al., 2008), logistic regression for dichotomous outcomes (Frank et al., 2021), omitted variables in multilevel models (Dietz et al., 2015), omitted levels of analysis in multilevel models (Chen, 2020), Bayesian analysis (Li & Frank, 2020), social network models (Xu & Frank, 2021),

and multisite randomized control trials and spillover violations of SUTVA in value added models (Lin, 2019).

Importantly, the RIR has also been extended to quantify concerns about external validity (Frank et al., 2013; Frank & Min, 2007) across disciplines, including education (Broda et al., 2018; Saw, Kunisaki, et al., 2025; Saw, Lin et al., 2025b), psychology (Ansari & Gottfried, 2021; Golec de Zavala et al., 2021), environmental studies (Chung et al., 2018; Mayer et al., 2021), sociology (Paxton et al., 2020; Pyne, 2019), political science (Chua, 2024; Ciobanu, 2024), economics (Lapatinas & Litina, 2019; Shen & Zhang, 2024), and business and management (Busenbark et al., 2021; Martino et al., 2024; Shin & You, 2023). The idea here is that one would like to make a general statement about causality that applies in a population that includes both a sampled and unsampled population. In the study example of Desmond and Kimbro (2015), one might seek to generalize the inference of an effect of eviction on economic hardship among women to a population in another subsequent time point. Of course, conditions of economy and housing could change generally between any two time points, eliminating the ability to claim absolutely that those in the study represent any other time. One question then is how much must the estimated effect of eviction based on data from a given time point be biased by the time frame to nullify an inference beyond that time point? This can be answered by applying the RIR. Interpreted in terms of sampling, to nullify an inference that applies beyond the study year of Desmond and Kimbro (2015), 43.4% of the mothers in the data would have to be replaced with cases from a subsequent year and for which eviction had no effect on economic hardship. This complements other techniques that characterize the surface similarity (Cook et al., 2002) of the sampled and unsampled populations (e.g., Tipton, 2014). Note the parallels of external validity robustness to the original presentation of the RIR. In both cases we acknowledge that the desired data are not observed, and we use the sensitivity analysis to quantify how much the observed data would have to be perturbed by unobserved data (from the counterfactual or sampling) to change the inference.

In the application to dichotomous patient outcomes, such as mortality, the RIR maps to the existing concept of “fragility” which has been gaining increasing attention in clinical epidemiology (Atal et al., 2019; Garcia-Retamero & Hoffrage, 2013; Walsh et al., 2014) with applications in oncology (Forrester et al., 2020) and pediatrics (Rickard et al., 2020). The Fragility Index indicates how many patients from the treatment group would have to have different outcomes, or experience event switches, to change an inference (Walsh et al., 2014). The RIR directly extends the Fragility Index in two fundamental ways. First, using RIR as in Figure 2, any threshold can be used as a basis for inference. Second, the RIR accounts for the likelihood that an outcome for a case will be switched. Consider the example in Table 1 below, drawn from Walter et al. (2020). These results are from a hypothetical experiment where 90/95 patients given Treatment A survived (versus died), 96/96 patients given Treatment B survived, with a p-value of .029 (based on Fisher’s exact test) leading to the inference that Treatment B is more effective than Treatment A. Walter et al. (2020) note that the Fragility Index of this inference is 1 – if one alive case in Treatment B were switched to died, the success rate would change to 95/96 in the treatment, and the *p*-value would change to 0.118. Correspondingly, if one uses a threshold of p=.05, the one switch would lead to an inference that there is no difference between Treatments A and B. Walter et al. (2020) note that the “fallacy” in this [the Fragility Index] argument is that the change from 0 to 1 death in treatment group B may actually be unlikely to occur because of the rarity of death.

Table 1. Robustness of inference for hypothetical treatment and mortality. Example taken from Walter et al [10]. Cells represent number of cases. Fragility Index= number of cases to switch to change the inference; RIR represents the robustness of the inference to replacement.

	Died		Alive	Total
Treatment A	5		90	95
Treatment B	0	Fragility Index = 1 ←	96 [RIR = 19]	96
Total	5		186	191

Walter et al's (2020) concern can be expressed by considering how switches are generated from case replacement. In particular, we ask how many of the 96 Treatment B Survived cases would have to be replaced with Treatment A cases to change the inference that Treatment B was more efficacious than Treatment A. We begin by drawing the replacement cases from a population represented by Treatment A with an estimated mortality rate of 5/95 or 5.3%. Using the 5.3% mortality rate, for every 19 Treatment B Survival cases replaced, we would expect 18 to remain classified as alive, and 1 to be reclassified as died. Therefore, we expect to have to replace 19 Treatment B alive cases to generate the one Treatment died case necessary to change the inference (p = 0.118). RIR = 19 out of 96 while the Fragility Index = 1.

Formally, the Fragility Index can be expressed as the expected number of replaced treatment cases with positive outcomes multiplied by the observed probability of negative outcomes in the control group: Fragility Index = RIR x $\hat{p}$, where $\hat{p}$ is the observed probability of a negative outcome in the control group. This implies that RIR = Fragility Index / $\hat{p}$. In the example, 19=1/.053. Thus, RIR is a funciton of $\hat{p}$, addressing Walter et al.'s (2020) critique of the Fragility Index by incorporating the prevalence of positive and negative outcomes in the data.

*** A further reading list on RIR can be found on KonFound-It! Website resources page.

3.2 An Application of RIR with Continuous Outcomes: A Step-by-Step Guide

The calculations of RIR with continuous outcomes can be performed with (1) a Shiny app KonFound-it! At https://konfound-project.shinyapps.io/konfound-it/, (2) Konfound commands in R software, (3) Konfound commands in Stata software, or (4) a Konfound-it! Spreadsheet (in Microsoft Excel). To compute the RIR of an estimated effect in a linear model (i.e., regression), a researcher will need the four following values from the data or estimated model: (1) estimated coefficient for the predictor of interest, (2) standard error, (3) sample size, and (4) number of covariates.

As an example of how to apply the RIR, consider Yeager et al.'s (2015) estimated effect of a growth mindset intervention on core course GPAs among lower-achieving adolescents, which is 0.10 grade points (standard error = 0.03; sample size = 6,320; number of covariates = 5; see the abstract and results of the paper below).

A national experiment reveals where a growth mindset improves achievement

David S. Yeager[1]*, Paul Hanselman[2]*, Gregory M. Walton[3], Jared S. Murray[1], Robert Crosnoe[1], Chandra Muller[1], Elizabeth Tipton[4], Barbara Schneider[5], Chris S. Hulleman[6], Cintia P. Hinojosa[7], David Paunesku[8], Carissa Romero[9], Kate Flint[10], Alice Roberts[10], Jill Trott[10], Ronaldo Iachan[10], Jenny Buontempo[1], Sophia Man Yang[1], Carlos M. Carvalho[1], P. Richard Hahn[11], Maithreyi Gopalan[12], Pratik Mhatre[1], Ronald Ferguson[13], Angela L. Duckworth[14] & Carol S. Dweck[3]

A global priority for the behavioural sciences is to develop cost-effective, scalable interventions that could improve the academic outcomes of adolescents at a population level, but no such interventions have so far been evaluated in a population-generalizable sample. Here we show that a short (less than one hour), online growth mindset intervention—which teaches that intellectual abilities can be developed—improved grades among lower-achieving students and increased overall enrolment to advanced mathematics courses in a nationally representative sample of students in secondary education in the United States. Notably, the study identified school contexts that sustained the effects of the growth mindset intervention: the intervention changed grades when peer norms aligned with the messages of the intervention. Confidence in the conclusions of this study comes from independent data collection and processing, pre-registration of analyses, and corroboration of results by a blinded Bayesian analysis.

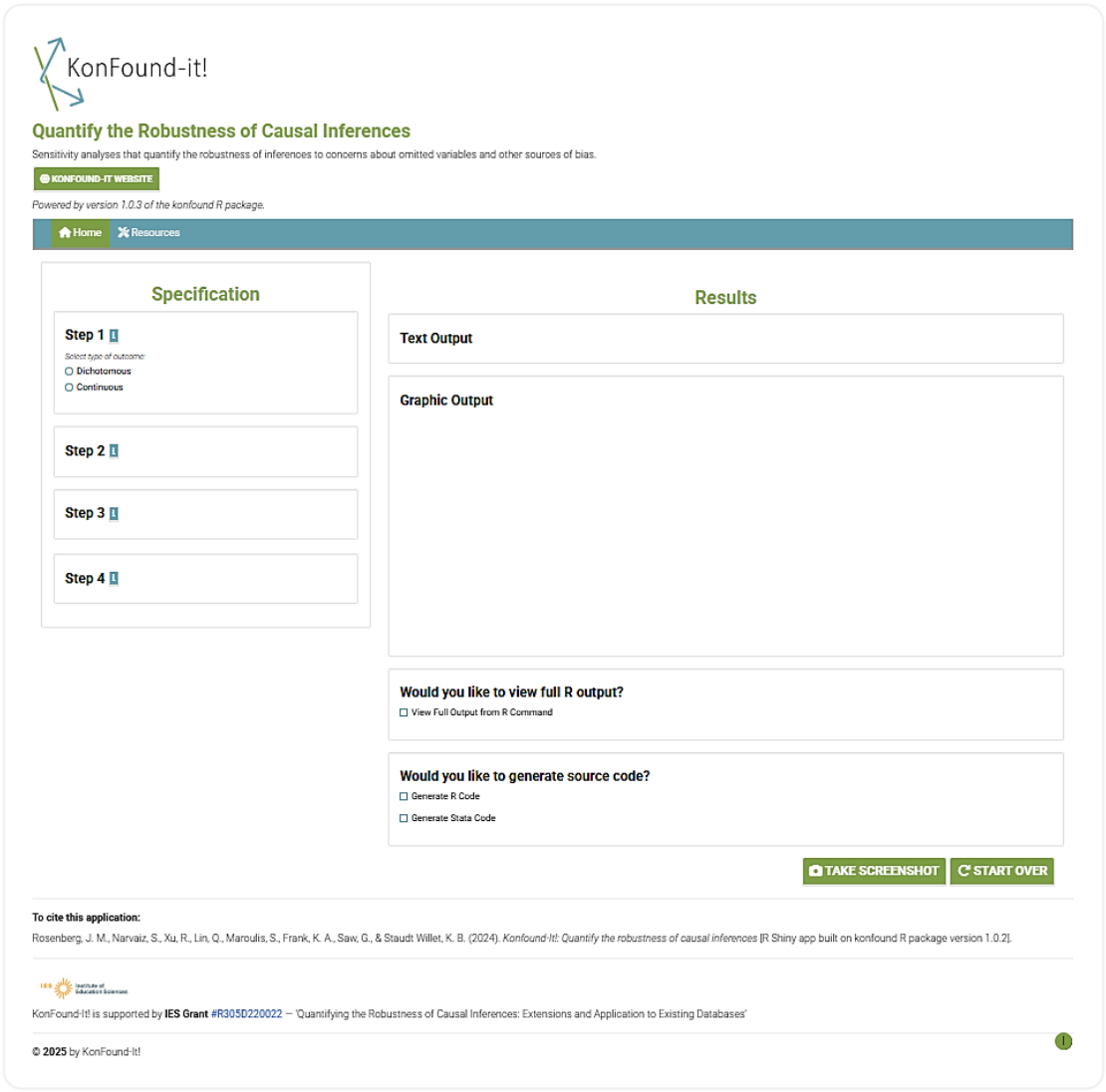
Average effects on core course GPAs

In line with our first major prediction, lower-achieving adolescents earned higher GPAs in core classes at the end of the ninth grade when assigned to the growth mindset intervention, $B = 0.10$ grade points (95% confidence interval = 0.04, 0.16), s.e. = 0.03, $n = 6{,}320$, $k = 65$, $t = 3.51$, $P = 0.001$, standardized mean difference effect size of 0.11, relative to comparable students in the control condition. This conclusion is robust to alternative model specifications that deviate from the pre-registered model (Extended Data Fig. 1).

3.2.1 Computing RIR with KonFound-it! Shiny App

3.2.1.1 Access

To use the KonFound-it! Shiny App, go to https://konfound-project.shinyapps.io/konfound-it/. As of the release of this practical guide, the KonFound-it! Shiny App is built with version 0.4.0 of the konfound R package, which would be updated over time.



3.2.1.2 Calculating RIR

To calculate the RIR, follow the steps illustrated below:

Step 1

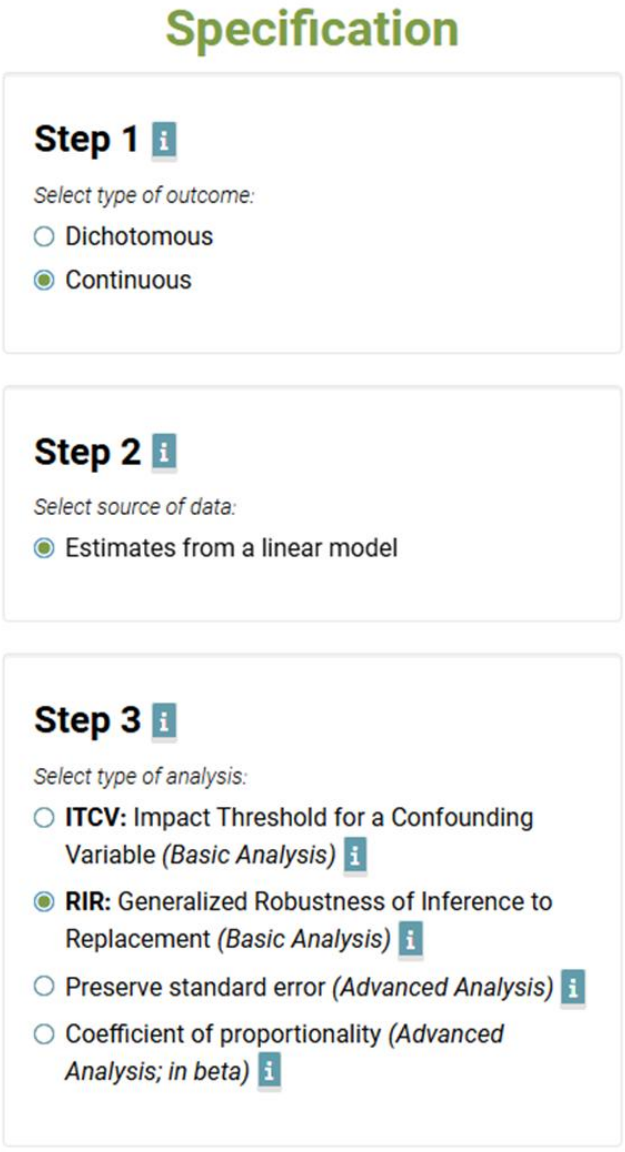
Choose the option of "Continuous" as the type of outcome

Step 2

The option of "Estimates from a linear model" will be automatically chosen

Step 3

Choose the option of "**RIR:** Generalized Robustness of Inference to Replacement (*Basic Analysis*)"



Step 4.1

Enter the coefficient of the predictor of interest

Step 4.2

Enter the standard error of the estimated effect

Step 4.3

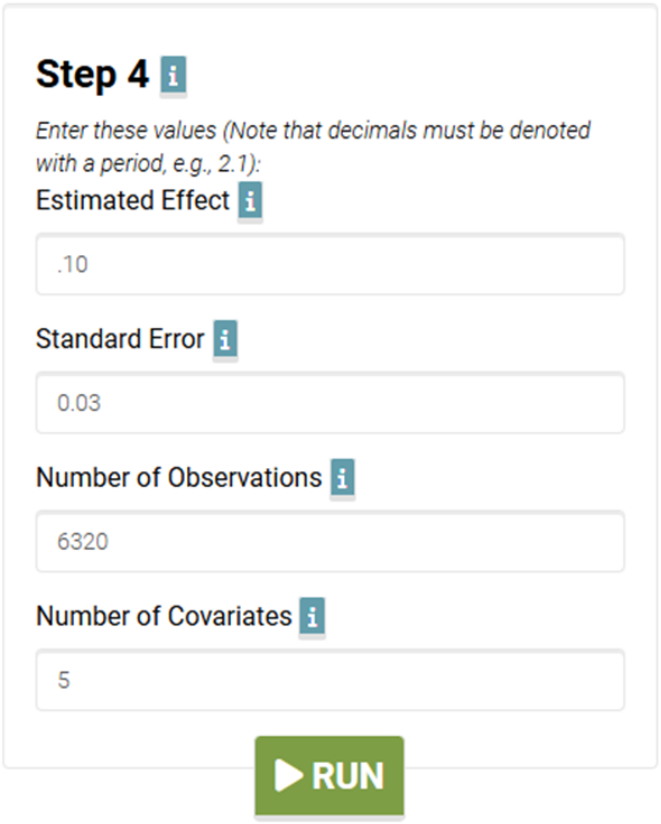
Enter the number of observations (or sample size) of the estimated model

Step 4.4

Enter the number of covariates included in the model other than the predictor of interest

Step 4.5

Click “RUN”



3.2.1.3 Output and Interpretation

Results

The **Text Output** presents a statement interpreting the calculated RIR.

The **Graphic Output** presents an estimated effect (0.100) relative to a threshold for inference (0.059).

Text Output

Robustness of Inference to Replacement (RIR):

RIR = 2603
To nullify the inference of an effect using the threshold of 0.059 for statistical significance (with null hypothesis = 0 and alpha = 0.05), 41.19% of the estimate of 0.1 would have to be due to bias. This implies that to nullify the inference one would expect to have to replace 2603 (41.19%) observations with data points for which the effect is 0 (RIR = 2603).

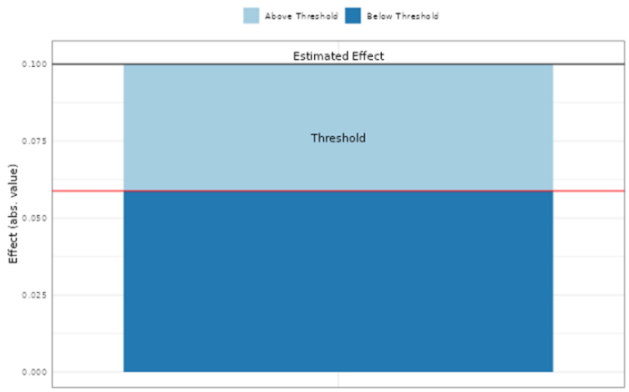
See Frank et al. (2013) for a description of the method.

Citation:
Frank, K.A., Maroulis, S., Duong, M., and Kelcey, B. (2013).
What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences.
Education, Evaluation and Policy Analysis, 35, 437-460.

Accuracy of results increases with the number of decimals reported. This analysis assumes the use of default parameters. For greater flexibility, use the R or Stata versions of the konfound package, beginning with the advanced code provided below on this page.

Calculated with konfound R package version 1.0.3

Graphic Output



Suggested Interpretation:
A suggested statement for interpreting the calculated RIR of Desmond and Kimbro's (2015) estimated effect of growth mindset intervention in a RCT on core course GPA reads: "to nullify the inference of the estimated effect of growth mindset intervention on core course GPAs among lower-achieving adolescents (coefficient = 0.10; standard error = 0.03; sample size = 6,320; number of covariates = 5), 41.19% of the cases (or 2,603 cases) would have to be replaced with counterfactual cases for which the treatment has zero effect" (Frank et al., 2013).

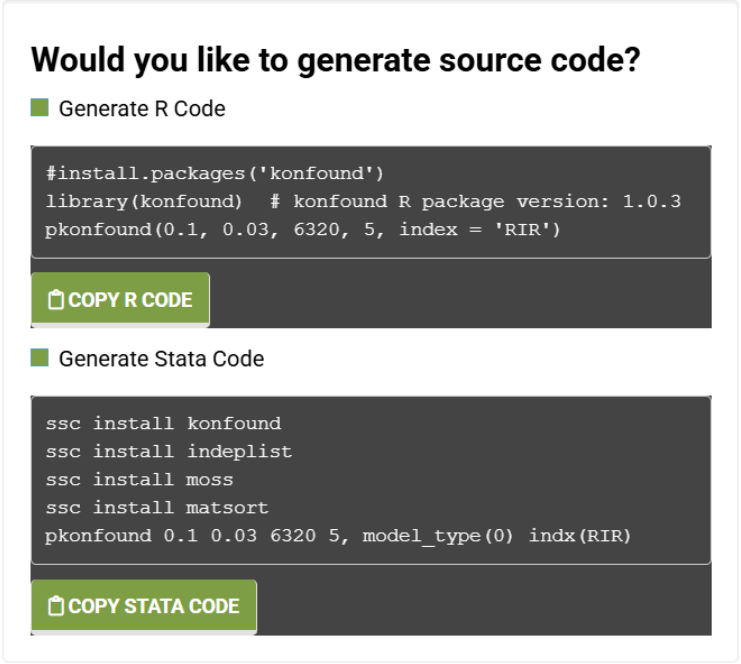
*** Other published empirical examples with ITCV interpretation can be found on KonFound-It! Website resources page.

3.2.1.4 Generating R and Stata codes

To generate R and Stata codes:

To generate R codes, check the box on the left of “**Generate R Code**”

To generate Stata codes, check the box on the left of “**Generate Stata Code**”



3.2.2 Computing RIR with R Software

3.2.2.1 Installation

To install the CRAN version of konfound:

```
install.packages("konfound")
```

To install the development version from GitHub:

```
install.packages("devtools")

devtools::install_github("jrosen48/konfound")
```

3.2.2.2 Calculating RIR

To calculate the RIR by manually entering results using long-form code:

```
library(konfound)
pkonfound(est_eff = 0.10,
          std_err = 0.03,
          n_obs = 6320,
          n_covariates = 5,
          index = 'RIR')
```

To calculate the RIR by manually entering results using short-form code:

```
pkonfound(0.10, 0.03, 6320, 5, index = 'RIR')
```

3.2.2.3 Output and Interpretation

R output (the same for both long- and short-form approach):

```
Robustness of Inference to Replacement (RIR):
RIR = 2603

To nullify the inference of an effect using the threshold of 0.059
for statistical significance (with null hypothesis = 0 and alpha =
0.05), 41.19% of the (0.1) estimate would have to be due to bias.
This implies that to nullify the inference one would expect to have
to replace 2603 (41.19%) observations with data points for which the
effect is 0 (RIR = 2603).

See Frank et al. (2013) for a description of the method.

Citation: Frank, K.A., Maroulis, S., Duong, M., and Kelcey, B. (2013).
```

What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Education Evaluation and Policy Analysis, 35*, 437-460.

Accuracy of results increases with the number of decimals reported.

3.2.3 Computing RIR with Stata Software

3.2.3.1 Installation

To install the Stata konfound command:

```
ssc install konfound
ssc install indeplist
ssc install moss
ssc install matsort
```

3.2.3.2 Calculating RIR

To calculate RIR by manually entering results:

```
pkonfound 0.10 0.03 6320 5, indx("RIR")
```

3.2.3.3 Output and Interpretation

Stata output:

```
Robustness of Inference to Replacement (RIR):

RIR = 2603

To nullify the inference of an effect using the threshold of 0.059 for
statistical significance (with null hypothesis = 0 and alpha = .05), 41.190%
of the (0.10) estimate would have to be due to bias. This implies that to
nullify the inference one would expect to have to replace 2603 (41.190%)
observations with data points for which the effect is 0 (RIR = 2603).

See Frank et al. (2013) for a description of the method.

Citation: Frank, K.A., Maroulis, S., Duong, M., and Kelcey, B. (2013). What
 would it take to change an inference? Using Rubin's causal model to
 interpret the robustness of causal inferences. Educational Evaluation and
 Policy Analysis, 35, 437-460.

Accuracy of results increases with the number of decimals reported.
```

3.2.4 Computing RIR with Konfound-it! Spreadsheet

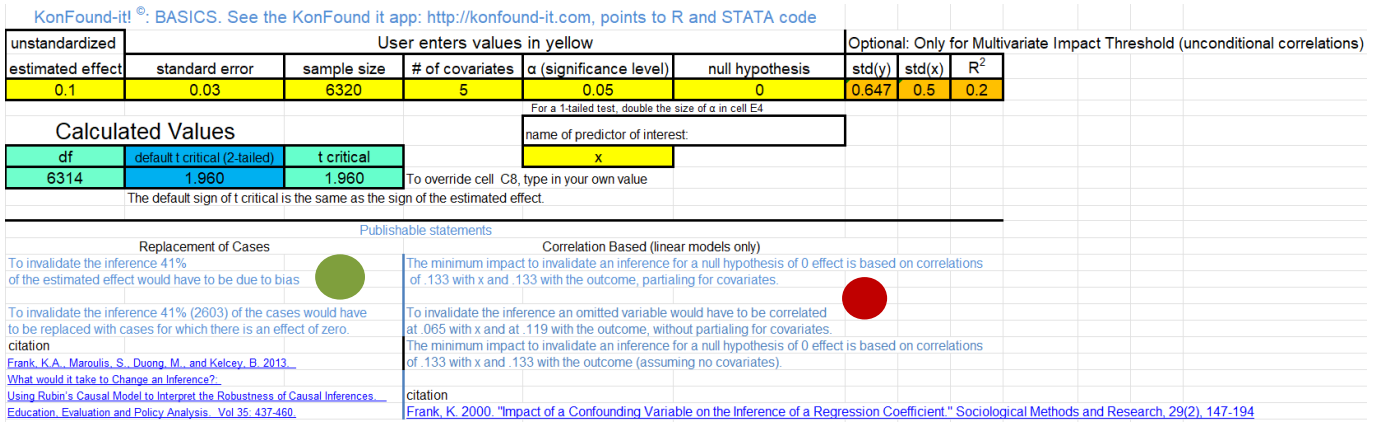
3.2.4.1 Download

Go to KonFound-it! Website Resources page and download the KonFound-it! spreadsheet for calculating indices.

3.2.4.2 Calculating ITCV, Output, and Interpretation

Follow the steps illustrated below to calculate the ITCV and obtain the output and interpretation:

1. Enter the estimated effect
2. Enter the standard error
3. Enter the sample size
4. Enter the number of covariates



KonFound-it! ©: BASICS. See the KonFound it app: http://konfound-it.com, points to R and STATA code

unstandardized estimated effect	standard error	sample size	# of covariates	α (significance level)	null hypothesis	Optional: Only for Multivariate Impact Threshold (unconditional correlations) std(y)	std(x)	R^2
	User enters values in yellow							
0.1	0.03	6320	5	0.05	0	0.647	0.5	0.2

For a 1-tailed test, double the size of α in cell E4

Calculated Values

df	default t critical (2-tailed)	t critical
6314	1.960	1.960

name of predictor of interest: x

To override cell C8, type in your own value

The default sign of t critical is the same as the sign of the estimated effect.

Publishable statements

Replacement of Cases

To invalidate the inference 41% of the estimated effect would have to be due to bias

To invalidate the inference 41% (2603) of the cases would have to be replaced with cases for which there is an effect of zero.

citation

Frank, K.A., Maroulis, S., Duong, M., and Kelcey, B. 2013. What would it take to Change an Inference?. Using Rubin's Causal Model to Interpret the Robustness of Causal Inferences. Education, Evaluation and Policy Analysis. Vol 35: 437-460.

Correlation Based (linear models only)

The minimum impact to invalidate an inference for a null hypothesis of 0 effect is based on correlations of .133 with x and .133 with the outcome, partialing for covariates.

To invalidate the inference an omitted variable would have to be correlated at .065 with x and at .119 with the outcome, without partialing for covariates.

The minimum impact to invalidate an inference for a null hypothesis of 0 effect is based on correlations of .133 with x and .133 with the outcome (assuming no covariates).

citation

Frank, K. 2000. "Impact of a Confounding Variable on the Inference of a Regression Coefficient." Sociological Methods and Research, 29(2), 147-194

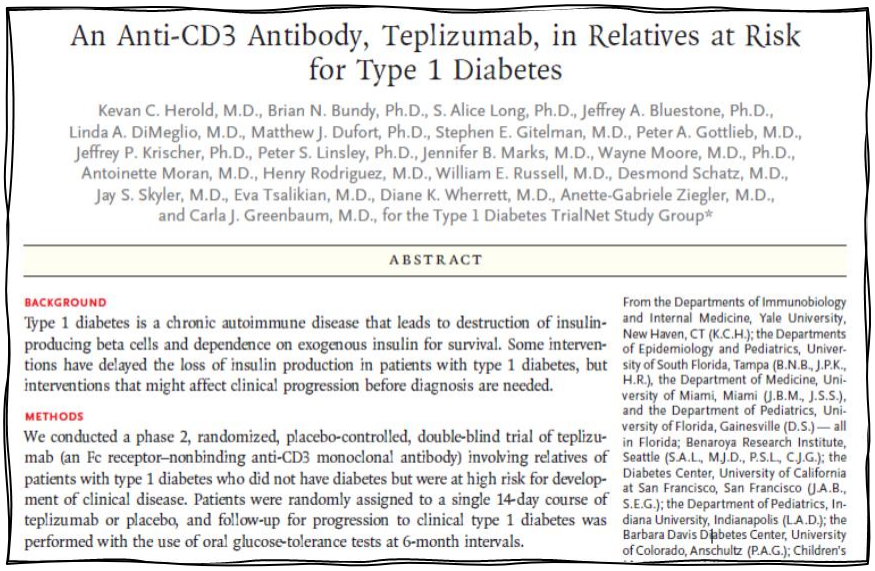
USE this set of output for the RIR

IGNORE this set of output for the RIR analysis

3.3 An Application of RIR with Dichotomous Outcomes for a 2x2 Table: A Step-by-Step Guide

The calculations of RIR with dichotomous outcomes in a logistic model can be performed with (1) a Shiny app KonFound-it! At https://konfound-project.shinyapps.io/konfound-it/, (2) pkonfound commands in R software, or (3) pkonfound commands in Stata software. To compute the RIR of an estimated effect in a logistic regression model, a researcher will need the five following values from the data or estimated model: (1) estimated coefficient for the predictor of interest (log odds), (2) standard error (of the log odds), (3) sample size, (4) number of covariates, and (5) number of cases in treatment condition.

As an example of how to apply the RIR with dichotomous outcomes for a 2x2 table, consider Herold et al.'s (2019) estimated RCT effect of Teplizumab (an antibody) on the development of type 1 diabetes in high-risk participants, where the disease was diagnosed in 19 (out of 44) of the participants who received teplizumab and in 23 (out of 32) of those who received placebo (see the abstract and results of the paper below).



An Anti-CD3 Antibody, Teplizumab, in Relatives at Risk for Type 1 Diabetes

Kevan C. Herold, M.D., Brian N. Bundy, Ph.D., S. Alice Long, Ph.D., Jeffrey A. Bluestone, Ph.D., Linda A. DiMeglio, M.D., Matthew J. Dufort, Ph.D., Stephen E. Gitelman, M.D., Peter A. Gottlieb, M.D., Jeffrey P. Krischer, Ph.D., Peter S. Linsley, Ph.D., Jennifer B. Marks, M.D., Wayne Moore, M.D., Ph.D., Antoinette Moran, M.D., Henry Rodriguez, M.D., William E. Russell, M.D., Desmond Schatz, M.D., Jay S. Skyler, M.D., Eva Tsalikian, M.D., Diane K. Wherrett, M.D., Anette-Gabriele Ziegler, M.D., and Carla J. Greenbaum, M.D., for the Type 1 Diabetes TrialNet Study Group*

ABSTRACT

BACKGROUND

Type 1 diabetes is a chronic autoimmune disease that leads to destruction of insulin-producing beta cells and dependence on exogenous insulin for survival. Some interventions have delayed the loss of insulin production in patients with type 1 diabetes, but interventions that might affect clinical progression before diagnosis are needed.

METHODS

We conducted a phase 2, randomized, placebo-controlled, double-blind trial of teplizumab (an Fc receptor–nonbinding anti-CD3 monoclonal antibody) involving relatives of patients with type 1 diabetes who did not have diabetes but were at high risk for development of clinical disease. Patients were randomly assigned to a single 14-day course of teplizumab or placebo, and follow-up for progression to clinical type 1 diabetes was performed with the use of oral glucose-tolerance tests at 6-month intervals.

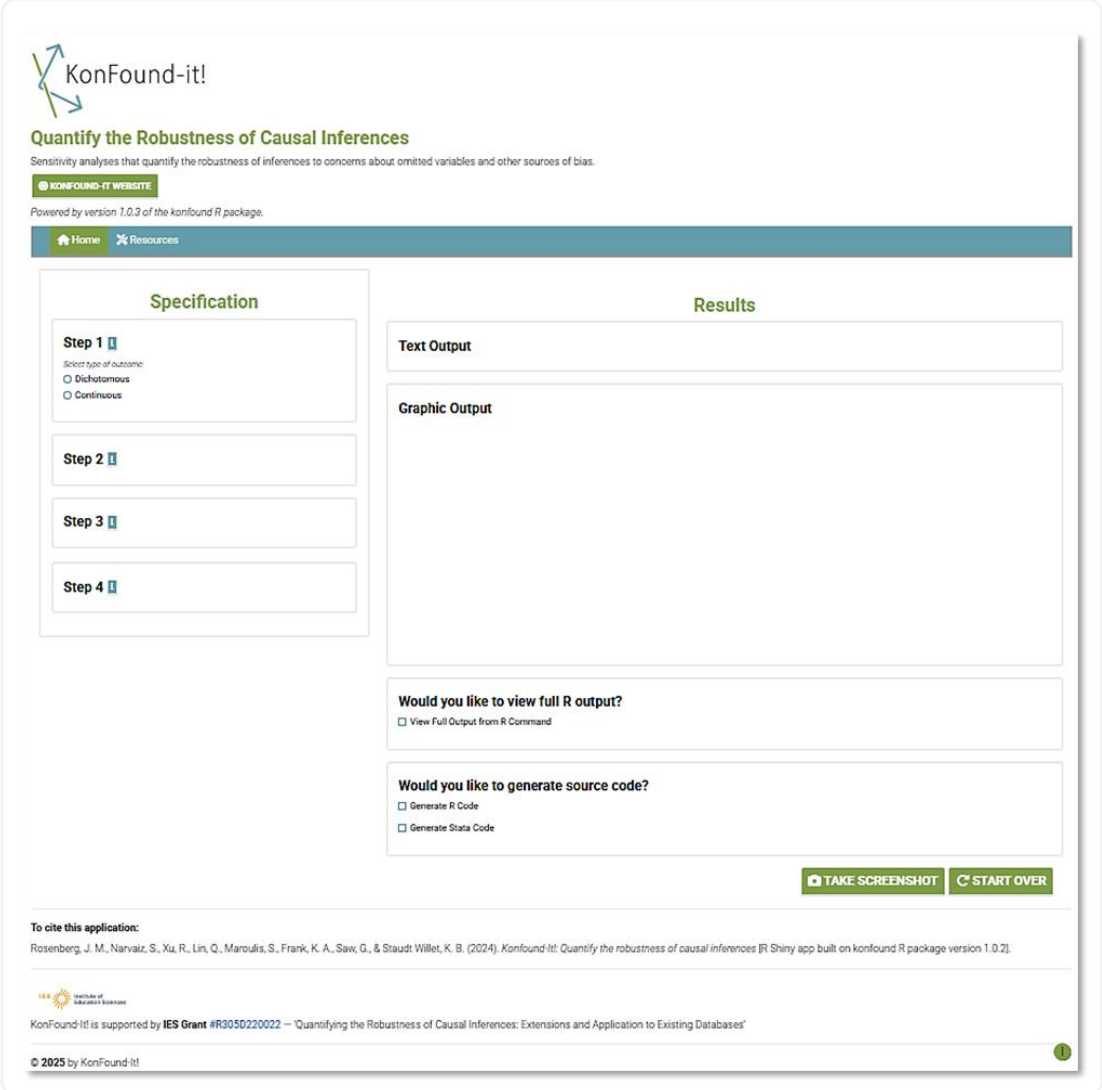
From the Departments of Immunobiology and Internal Medicine, Yale University, New Haven, CT (K.C.H.); the Departments of Epidemiology and Pediatrics, University of South Florida, Tampa (B.N.B., J.P.K., H.R.), the Department of Medicine, University of Miami, Miami (J.B.M., J.S.S.), and the Department of Pediatrics, University of Florida, Gainesville (D.S.) — all in Florida; Benaroya Research Institute, Seattle (S.A.L., M.J.D., P.S.L., C.J.G.); the Diabetes Center, University of California at San Francisco, San Francisco (J.A.B., S.E.G.); the Department of Pediatrics, Indiana University, Indianapolis (L.A.D.); the Barbara Davis Diabetes Center, University of Colorado, Anschultz (P.A.G.); Children's

	No. without Diagnosis	No. with Diagnosis
Teplizumab	25	19
Placebo	9	23

3.3.1 Computing RIR with KonFound-it! Shiny App

3.3.1.1 Access

To use the KonFound-it! Shiny App, go to https://konfound-project.shinyapps.io/konfound-it/. As of the release of this practical guide, the KonFound-it! Shiny App is built with version 1.0.3 of the konfound R package, which would be updated over time.



3.3.1.2 Calculating RIR

To calculate the RIR, follow the steps illustrated below:

Step 1

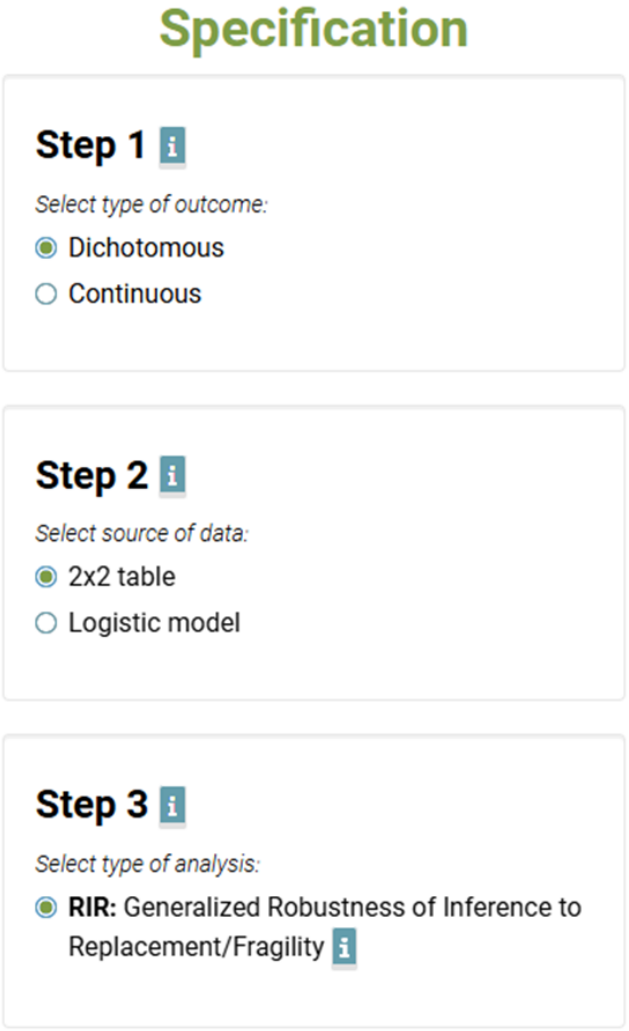
Choose the option of "Dichotomous" as the type of outcome

Step 2

Choose the option of "Logistic model"

Step 3

The option of "Generalized Robustness of Inference to Replacement (RIR)" will be automatically chosen



Step 4.1

Enter the number of control failure cases

Step 4.2

Enter the number of control success cases

Step 4.3

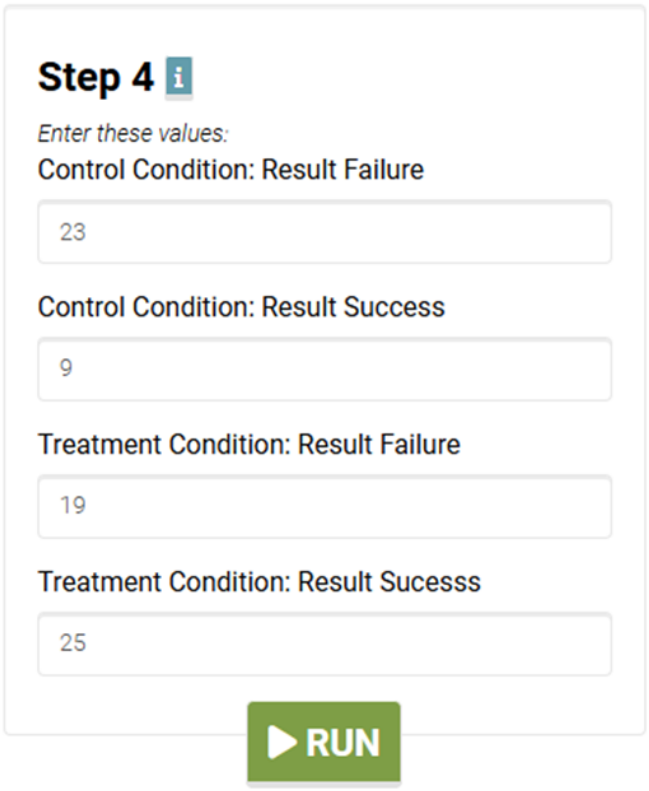
Enter the number of treatment failure cases

Step 4.4

Enter the number of treatment success cases

Step 4.5

Click “RUN”



3.3.1.3 Output and Interpretation

The **Text Output** presents a statement interpreting the calculated RIR. It also presents an User-Enetred Table and a Transfer Table showing the results of transferring cases from treatment success to treatment failure.

Results

Text Output

Robustness of Inference to Replacement (RIR):
RIR = 3
Fragility = 2

This function calculates the number of data points that would have to be replaced with zero-effect data points (RIR) to nullify or sustain the inference made about the association between the rows and columns in a 2x2 table. One can also interpret this as switches (Fragility) from one cell to another, such as from the treatment success cell to the treatment failure cell.

To nullify the inference that the effect is different from 0 (alpha = 0.05), one would need to transfer 2 data points from treatment success to treatment failure as shown, from the User-entered Table to the Transfer Table (Fragility = 2). This is equivalent to replacing 3 (12%) treatment success data points with data points for which the probability of failure in the control group (71.875%) applies (RIR = 3).

Note that RIR = Fragility/P(destination) = 2/0.719 ~ 3.

For the User-entered Table, the estimated odds ratio is 3.307, with p-value of 0.019:

User-Entered Table:

Group	Failures	Successes	Success Rate
Control	23	9	28.12%
Treatment	19	25	56.82%
Total	42	34	44.74%

For the Transfer Table, the estimated odds ratio is 2.76, with p-value of 0.059:

Transfer Table:

Group	Failures	Successes	Success Rate
Control	23	9	28.12%
Treatment	21	23	52.27%
Total	44	32	42.11%

See Frank et al. (2021) for a description of the method.

Citation:
*Frank, K. A., *Lin, Q., *Maroulis, S., *Mueller, A. S., Xu, R., Rosenberg, J. M., ... & Zhang, L. (2021).
Hypothetical case replacement can be used to quantify the robustness of trial results.
Journal of Clinical Epidemiology, 134, 150-159.
*Authors are listed alphabetically.

Accuracy of results increases with the number of decimals entered.
This analysis assumes the use of default parameters. For greater flexibility, use the R or Stata versions of the konfound package, beginning with the advanced code provided below on this page.

Calculated with konfound R package version 1.0.3

Suggested Interpretation:
A suggested statement for interpreting the calculated RIR of Herold et al.'s (2019) estimated effect of teplizumab on development of type 1 diabetes reads: "to nullify the inference of the estimated effect of teplizumab on the development of type 1 diabetes (19 of the 44 participants who received teplizumab and 23 of the 32 participants who received placebo had type 1 diabetes diagnosed), one would need to transfer two data points from treatment success to treatment failure (Fragility = 2), which is equivalent to replacing three treatment success data points with data points for which the probability of failure in the control group applies (RIR = 3)."

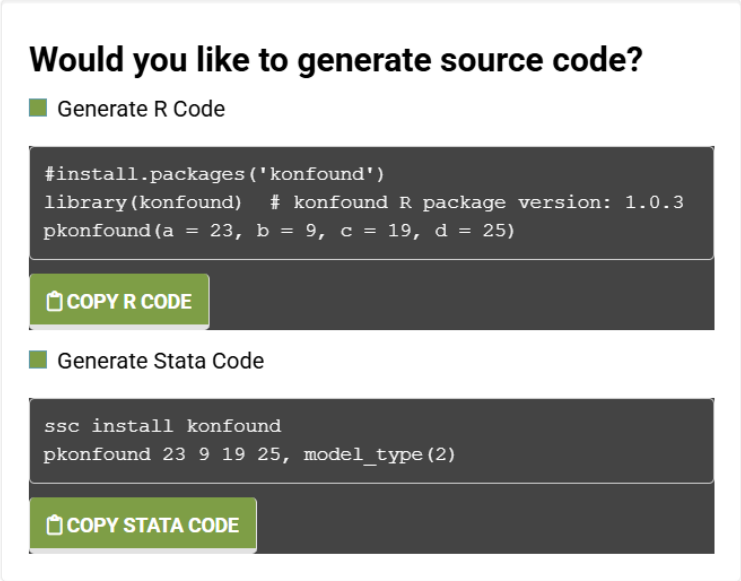
*** Other published empirical examples with RIR interpretation can be found on KonFound-It! Website resources page.

3.3.1.4 Generating R and Stata Codes

To generate R and Stata codes:

To generate R codes, check the box on the left of "**Generate R Code**"

To generate Stata codes, check the box on the left of "**Generate Stata Code**"



3.3.2 Computing RIR with R Software

3.3.2.1 Installation

To install the CRAN version of konfound:

```
install.packages("konfound")
```

To install the development version from GitHub:

```
install.packages("devtools")
devtools::install_github("jrosen48/konfound")
```

3.3.2.2 Calculating RIR

To calculate the RIR:

```
library(konfound)
pkonfound(a = 23, b = 9, c = 19, d = 25)
```

3.3.2.3 Output and Interpretation

R output:

```
Robustness of Inference to Replacement (RIR):
RIR = 3
Fragility = 2

This function calculates the number of data points that would have
to be replaced with zero effect data points (RIR) to nullify the
inference made about the association between the rows and columns
in a 2x2 table. One can also interpret this as switches (Fragility)
from one cell to another, such as from the treatment success cell
to the treatment failure cell.

To nullify the inference that the effect is different from 0 (alpha
= 0.05), one would need to transfer 2 data points from treatment
success to treatment failure as shown, from the User-entered Table
to the Transfer Table (Fragility = 2). This is equivalent to
replacing 3 (12.000%) treatment success data points with data points
for which the probability of failure in the control group (71.875%)
applies (RIR = 3).

RIR = Fragility/P(destination)

For the User-entered Table, the estimated odds ratio is 3.307, with
p-value of 0.019:
User-entered Table:
          Fail Success Success_Rate
Control     23       9       28.12%
Treatment   19      25       56.82%
Total       42      34       44.74%

For the Transfer Table, the estimated odds ratio is 2.760, with p-
value of 0.059:
Transfer Table:
          Fail Success Success_Rate
Control     23       9       28.12%
Treatment   21      23       52.27%
Total       44      32       42.11%

See Frank et al. (2021) for a description of the methods.

*Frank, K. A., *Lin, Q., *Maroulis, S., *Mueller, A. S., Xu, R.,
Rosenberg, J. M., ... & Zhang, L. (2021). Hypothetical case
replacement can be used to quantify the robustness of trial results.
Journal of Clinical Epidemiology, 134, 150-159. *authors are listed
alphabetically.
```

3.3.3 Computing RIR with Stata Software

3.3.3.1 Installation

To install the Stata konfound command:

```
ssc install konfound
ssc install indeplist
ssc install moss
ssc install matsort
```

3.3.3.2 Calculating RIR

To calculate RIR:

```
pkonfound 23 9 19 25, replace(1) model_type(2)
```

Note: replace(#) – indicates whether to use the entire sample or the control group to calculate the base rate; the default value is control replace(0), to change to entire use replace(1)

3.3.3.3 Output and Interpretation

Stata output:

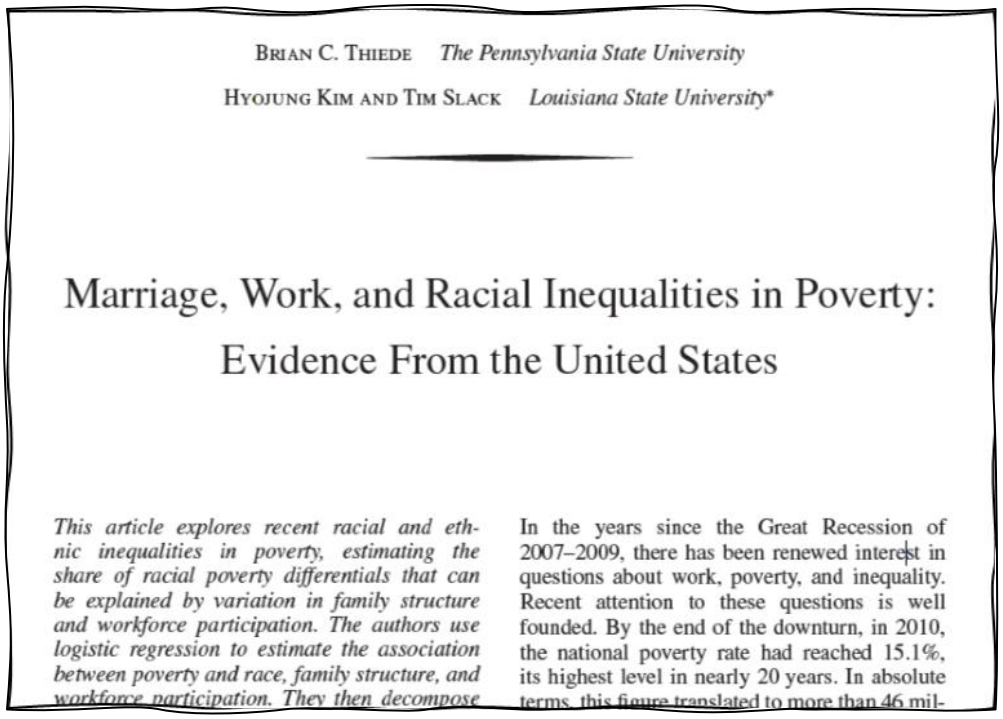
```
Robustness of Inference to Replacement (RIR):

RIR = 3
Fragility = 2

This function calculates the number of data points that would have to be
replaced with zero effect data points (RIR) to nullify the inference made
about the association between the rows and columns in a 2x2 table.
One can also interpret this as switches (Fragility) from one cell to
another, such as from the treatment success cell to the treatment failure
cell.

To nullify the inference that the effect is different from 0 (alpha = .05),
one would need to transfer 2 data points from treatment success to
treatment failure as shown, from the User-entered Table to the Transfer
Table (Fragility = 2). This is equivalent to replacing 3 (12.000%)
treatment success data points with data points for which the probability
of failure in the control group (71.875%) applies (RIR = 3).

RIR = Fragility/P(destination)
```

```
For the User-entered Table, the estimated odds ratio is 3.363, with p-
value of 0.019.
User-entered Table:

             |      Fail    Success  Success_%
-------------+-------------------------------
     Control |        23          9      28.13
   Treatment |        19         25      56.82
       Total |        42         34      44.74

For the Transfer Table, the estimated odds ratio of 2.799, with p-value
of 0.059.
Transfer table:

             |      Fail    Success  Success_%
-------------+-------------------------------
     Control |        23          9      28.13
   Treatment |        21         23      52.27
       Total |        44         32      42.11

See Frank et al. (2021) for a description of the methods.

*Frank, K. A., *Lin, Q., *Maroulis, S., *Mueller, A. S., Xu, R., Rosenberg,
J. M., ... & Zhang, L. (2021). Hypothetical case replacement can be used
to quantify the robustness of trial results. Journal of Clinical
Epidemiology, 134, 150-159. *authors are listed alphabetically.
```

3.4 An Application of RIR with Dichotomous Outcomes in Logistic Model: A Step-by-Step Guide

The The calculations of RIR with dichotomous outcomes in a logistic model can be performed with (1) a Shiny app KonFound-it! at https://konfound-project.shinyapps.io/konfound-it/(2) pkonfound commands in R software, or (3) pkonfound commands in Stata software. To compute the RIR of an estimated effect in a logistic regression model, regression, a researcher will need the five following values from the data or estimated model: (1) estimated coefficient for the predictor of interest (log odds), (2) standard error (of the log odds), (3) sample size, (4) number of covariates, and (5) number of cases in treatment condition.

As an example of how to apply the RIR with dichotomous outcomes in a logistic regression model, consider Thiede et al.'s (2017) estimated effect (log odds) of cohabitation, compared with being married, on poverty among Hispanic households in the U.S., which is 0.388 (standard error = 0.139; sample size = 14,082; number of covariates = 23; number of cases in treatment condition = 1,267; see the abstract and results of the paper below).



Brian C. Thiede *The Pennsylvania State University*

Hyojung Kim and Tim Slack *Louisiana State University**

Marriage, Work, and Racial Inequalities in Poverty: Evidence From the United States

This article explores recent racial and ethnic inequalities in poverty, estimating the share of racial poverty differentials that can be explained by variation in family structure and workforce participation. The authors use logistic regression to estimate the association between poverty and race, family structure, and workforce participation. They then decompose

In the years since the Great Recession of 2007–2009, there has been renewed interest in questions about work, poverty, and inequality. Recent attention to these questions is well founded. By the end of the downturn, in 2010, the national poverty rate had reached 15.1%, its highest level in nearly 20 years. In absolute terms, this figure translated to more than 46 mil-

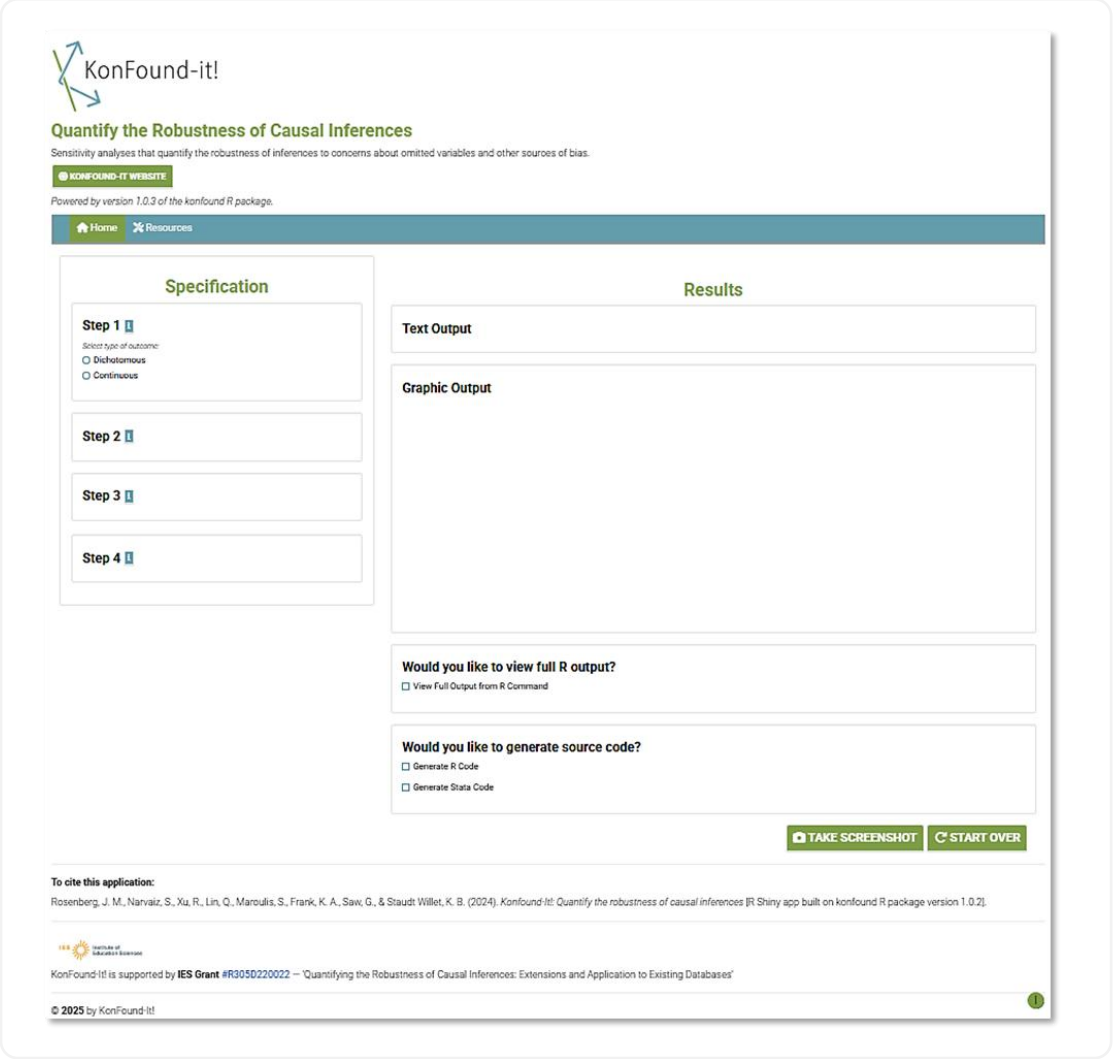
Table 3. *Logistic Regression Models Predicting Poverty, by Race*

	Non-Hispanic White		Non-Hispanic Black		Hispanic	
Variable	*B*	*SE*	*B*	*SE*	*B*	*SE*
Number of adult workers (ref. = zero)						
One worker	−2.560***	0.059	−2.596***	0.094	−2.699***	0.116
Two + workers	−4.393***	0.094	−4.507***	0.157	−4.496***	0.142
Family structure (ref. = married)						
Cohabitating	0.522***	0.099	0.508*	0.210	0.388**	0.139
Single male-headed	0.530***	0.103	0.723***	0.134	0.194	0.124
Single female-headed	0.977***	0.096	0.803***	0.115	0.853***	0.108

3.4.1 Computing RIR with KonFound-it! Shiny App

3.4.1.1 Access

To use the KonFound-it! Shiny App, go to https://konfound-project.shinyapps.io/konfound-it/. As of the release of this practical guide, the KonFound-it! Shiny App is built with version 0.4.0 of the konfound R package, which would be updated over time.



3.4.1.2 Calculating RIR

To calculate the RIR, follow the steps illustrated below:

Step 1

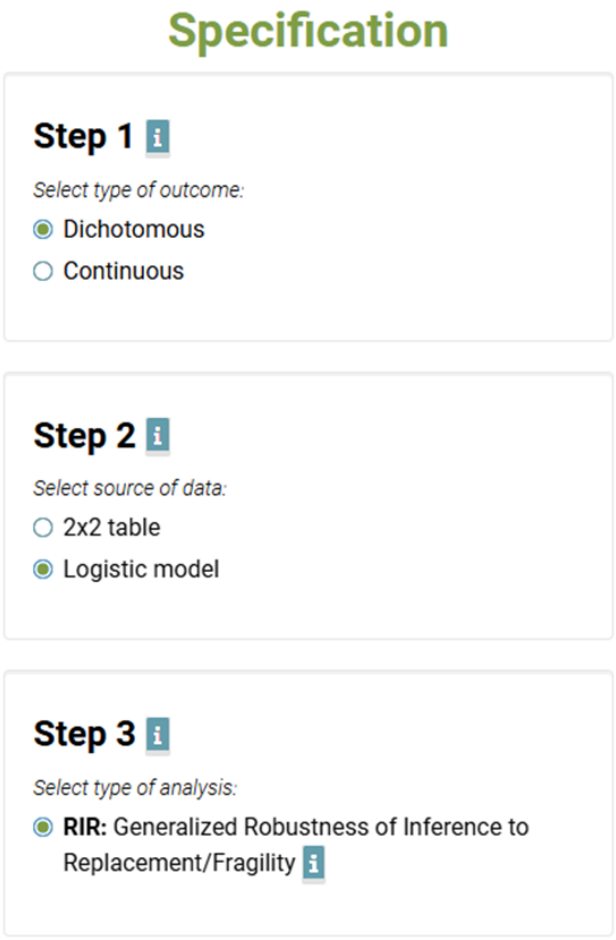
Choose the option of "Dichotomous" as the type of outcome

Step 2

Choose the option of "Logistic model

Step 3

Choose the option of "RIR: Generalized Robustness of Inference to Replacement/Fragility"



Step 4.1

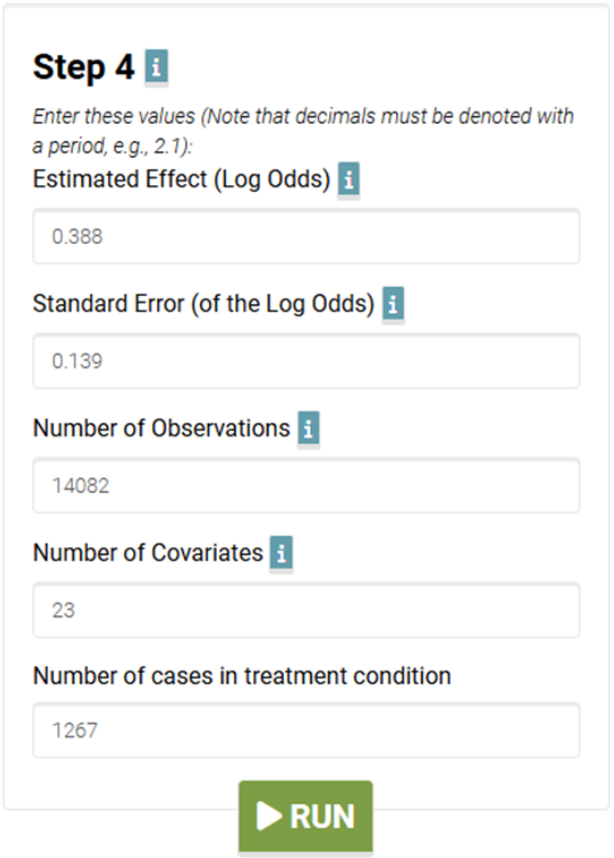
Enter the coefficient of the predictor of interest (log odds)

Step 4.2

Enter the standard error of the estimated effect (log odds)

Step 4.3

Enter the number of observations (or sample size) of the estimated model



Step 4.4

Enter the number of covariates included in the model other than the predictor of interest

Step 4.5

Enter the number of cases in treatment condition

Step 4.6

Click "RUN"

***The software takes these values to generate an implied 2x2 table. If there is a cell with a count smaller than five, an error will be generated. In some situations, this could be corrected by increasing the value of standard error, which will be reported in the output.

3.4.1.3 Output and Interpretation

The **Text Output** presents a statement interpreting the calculated RIR.

Results

Text Output

Robustness of Inference to Replacement (RIR):
RIR = 7
Fragility = 6

You entered: log odds = 0.388, SE = 0.139, with p-value = 0.006.
The table implied by the parameter estimates and sample sizes you entered:
User-Entered Table:

	Failures	Successes	Success Rate
Control	12382	433	3.38%
Treatment	1205	62	4.89%
Total	13587	495	3.52%

Values in the table have been rounded to the nearest integer. This may cause a small change to the estimated effect for the table.

To nullify the inference that the effect is different from 0 (alpha = 0.050), one would need to transfer 6 data points from treatment success to treatment failure (Fragility = 6).
This is equivalent to replacing 7 (11.29%) treatment success data points with data points for which the probability of failure in the control group (96.621%) applies (RIR = 7).

Note that RIR = Fragility/P(destination) = 6/0.966 ~ 7.

The transfer of data points yields the following table:
Transfer Table:

	Failures	Successes	Success Rate
Control	12382	433	3.38%
Treatment	1211	56	4.42%
Total	13593	489	3.47%

The log odds (estimated effect) = 0.279, SE = 0.145, p-value = 0.054.
This p-value is based on t = estimated effect / standard error.

Benchmarking RIR for Logistic Regression
The benchmark value helps interpret the RIR necessary to nullify an inference by comparing the change needed to nullify the inference with the changes in the estimated effect due to observed covariates.Currently this feature is available only when the reported results are statistically significant.

The benchmark is used to compare the bias needed to nullify the inference / bias reduction due to observed covariates. Specifically, change in data from implied to transfer table / change in data from unconditional table to implied table.

To calculate this benchmark value, a range of treatment success values is automatically generated based on the assumption that the marginals are constant between the implied table and the raw unadjusted table.
The benchmark value is visualized as a graph, allowing the user to interpret how the benchmark changes with hypothesized treatment success values.

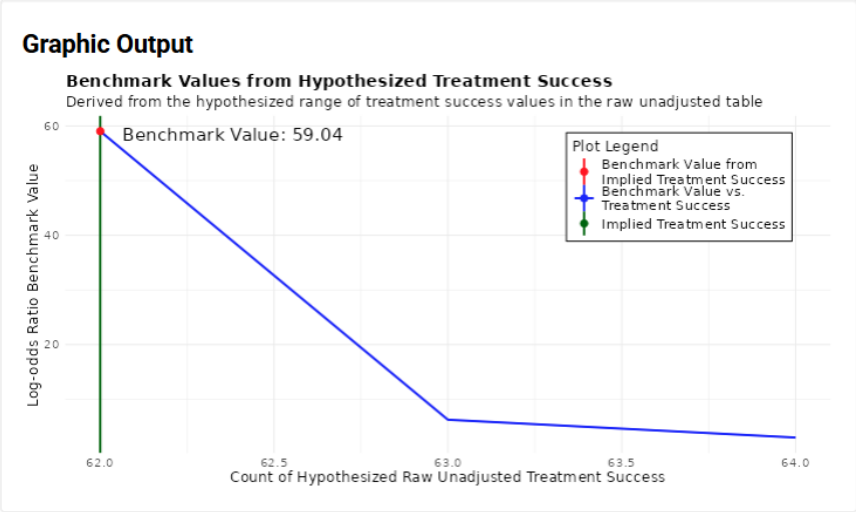
See Frank et al. (2021) for a description of the methods.

Citation:
*Frank, K. A., *Lin, Q., *Maroulis, S., *Mueller, A. S., Xu, R., Rosenberg, J. M., ... & Zhang, L. (2021). Hypothetical case replacement can be used to quantify the robustness of trial results. *Journal of Clinical Epidemiology, 134*, 150-159.
*Authors are listed alphabetically.

Accuracy of results increases with the number of decimals entered.
This analysis assumes the use of default parameters. For greater flexibility, use the R or Stata versions of the konfound package, beginning with the advanced code provided below on this page.

Calculated with konfound R package version 1.0.3

The **Graphic Output** presents benchmark values from hypothesized treatment success.



Suggested Interpretation:
A suggested statement for interpreting the calculated RIR of Thiede et al.'s (2017) estimated effect of cohabitation on poverty reads: "to nullify the inference of the estimated effect of cohabitation on poverty among Hispanic households in the U.S. (log odds = 0.388; standard error = 0.139; sample size = 14,082; number of covariates = 23; number of cases in treatment condition = 1,267), one would need to replace 7 (11.3%) treatment success data points with data points for which the probability of failure in the control group (96.6%) applies (RIR = 7)." (Frank et al., 2021).

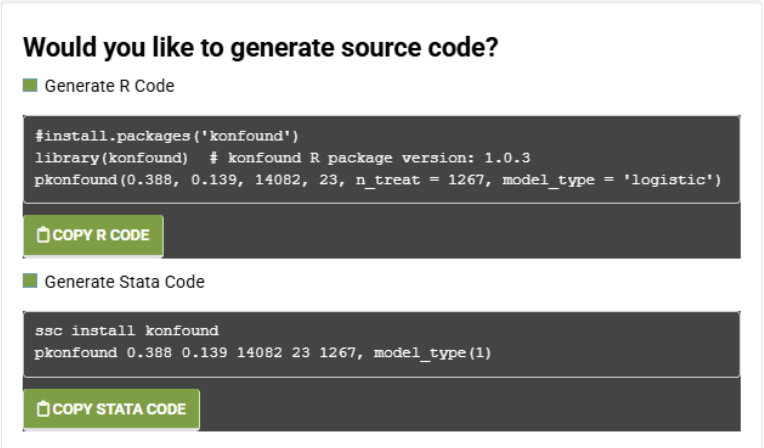
*** Other published empirical examples with RIR interpretation can be found on KonFound-It! Website resources page.

3.4.1.4 Generating R Code (Stata code coming soon)

To generate R code:

To generate R codes, check the box on the left of "**Generate R Code**"

To generate Stata codes, check the box on the left of "**Generate Stata Code**"



3.4.2 Computing RIR with R Software

3.4.2.1 Installation

To install the CRAN version of konfound:

```
install.packages("konfound")
```

To install the development version from GitHub:

```
install.packages("devtools")

devtools::install_github("jrosen48/konfound")
```

3.4.2.2 Calculating RIR

To calculate the RIR by manually entering results with long-form code:

```
library(konfound)
pkonfound(est_eff = 0.388,
          std_err = 0.139,
          n_obs = 14082,
          n_covariates = 23,
          n_treat = 1267,
          model_type = 'logistic')
```

To calculate the RIR by manually entering results with short-form code:

```
pkonfound(0.388, 0.139, 14082, 23, n_treat = 1267, model_type = 'lo
  gistic')
```

3.4.2.3 Output and Interpretation

R output (the same for both long- and short-form approach):

```
Robustness of Inference to Replacement (RIR):
RIR = 7
Fragility = 6

You entered: log odds = 0.388, SE = 0.139, with p-value = 0.006.
The table implied by the parameter estimates and sample sizes you
entered:
User-entered Table:
           Fail Success Success_Rate
Control   12382     433        3.38%
Treatment  1205      62        4.89%
Total     13587     495        3.52%

Values in the table have been rounded to the nearest integer. This
may cause a small change to the estimated effect for the table.
```

To nullify the inference that the effect is different from 0 (alpha = 0.050), one would need to transfer 6 data points from treatment success to treatment failure (Fragility = 6). This is equivalent to replacing 7 (11.290%) treatment success data points with data points for which the probability of failure in the control group (96.621%) applies (RIR = 7).

Note that RIR = Fragility/P(destination) = 6/0.966 ~ 7.

The transfer of 6 data points yields the following table:

Transfer Table:

	Fail	Success	Success_Rate
Control	12382	433	3.38%
Treatment	1211	56	4.42%
Total	13593	489	3.47%

The log odds (estimated effect) = 0.279, SE = 0.145, p-value = 0.054. This is based on t = estimated effect/standard error

Benchmarking RIR for Logistic Regression

The benchmark value helps interpret the RIR necessary to nullify an inference by comparing the change needed to nullify the inference with the changes in the estimated effect due to observed covariates. Currently this feature is available only when the reported results are statistically significant.

The benchmark is used to compare the bias needed to nullify the inference / bias reduction due to observed covariates. Specifically, change in data from implied to transfer table / change in data from unconditional table to implied table

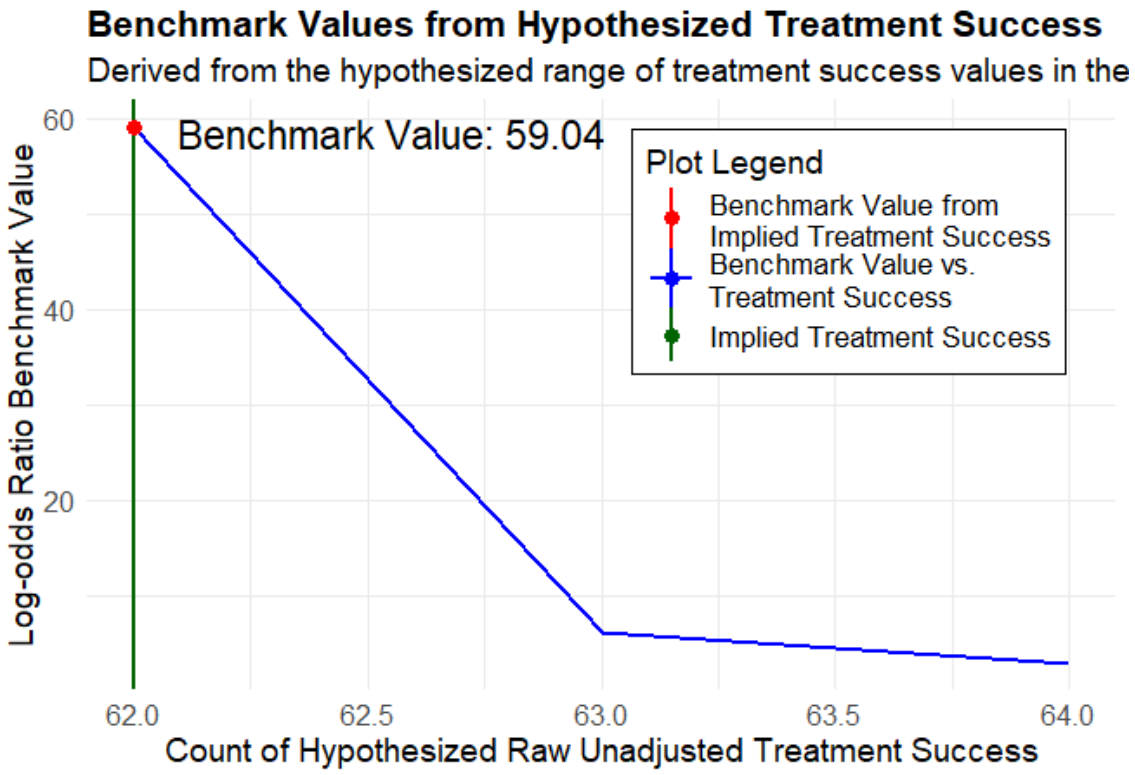
To calculate this benchmark value, a range of treatment success values is automatically generated based on the assumption that the marginals are constant between the implied table and the raw unadjusted table. The benchmark value is visualized as a graph, allowing the user to interpret how the benchmark changes with hypothesized treatment success values.

To calculate a specific benchmark value, locate the number of treatment successes in the raw data on the graph below.

See Frank et al. (2021) for a description of the methods.

*Frank, K. A., *Lin, Q., *Maroulis, S., *Mueller, A. S., Xu, R., Rosenberg, J. M., ... & Zhang, L. (2021). Hypothetical case replacement can be used to quantify the robustness of trial results. *Journal of Clinical Epidemiology, 134*, 150-159. *authors are listed alphabetically.

Accuracy of results increases with the number of decimals entered.


Benchmark Values from Hypothesized Treatment Success
Derived from the hypothesized range of treatment success values in the
Benchmark Value: 59.04
Plot Legend
Benchmark Value from Implied Treatment Success
Benchmark Value vs. Treatment Success
Implied Treatment Success
Log-odds Ratio Benchmark Value
60
40
20
62.0
62.5
63.0
63.5
64.0
Count of Hypothesized Raw Unadjusted Treatment Success

3.4.3 Computing RIR with Stata Software

3.4.3.1 Installation

To install the Stata konfound command:

```
ssc install konfound
ssc install indeplist
ssc install moss
ssc install matsort
```

3.4.3.2 Calculating RIR

To calculate RIR by manually entering results:

```
pkonfound .388 .139 14082 23 1267, model_type(1)
```

3.3.3.3 Output and Interpretation

Stata output:

```
Robustness of Inference to Replacement (RIR):

RIR = 7
Fragility = 6

The table implied by the parameter estimates and sample sizes you entered:
User-entered Table:

             |      Fail    Success  Success_%
-------------+----------------------------------
     Control |     12382        433   3.378853
   Treatment |      1205         62   4.893449
       Total |     13587        495   3.515126

The reported effect size = .388, SE = .139, p-value = 0.006. Values in
the table have been rounded to the nearest integer. This may cause a small
change to the estimated effect for the table.

To nullify the inference that the effect is different from 0 (alpha =
0.050), one would need to replace 6 data points from treatment success to
treatment failure (Fragility = 6). This is equivalent to replacing 7
(11.290%) treatment success data points with data points for which the
probability of failure in the control group (96.621%) applies (RIR = 7).

RIR = Fragility/P(destination)
```

```
The transfer of 6 data points yields the following table:
Transfer Table

             |      Fail    Success  Success_%
-------------+------------------------------
     Control |     12382        433   3.378853
   Treatment |      1211         56    4.41989
       Total |     13593        489   3.472518

The log odds (estimated effect) = 0.279, SE = 0.145, p-value = 0.054. This
is based on t = estimated effect/standard error

See Frank et al. (2021) for a description of the methods.

*Frank, K. A., *Lin, Q., *Maroulis, S., *Mueller, A. S., Xu, R., Rosenberg,
J. M., ... & Zhang, L. (2021). Hypothetical case replacement can be used
to quantify the robustness of trial results. Journal of Clinical
Epidemiology, 134, 150-159. *authors are listed alphabetically.

Accuracy of results increases with the number of decimals entered.
```

3.5 Benchmarks for the RIR using data from What Works Clearinghouse (WWC)

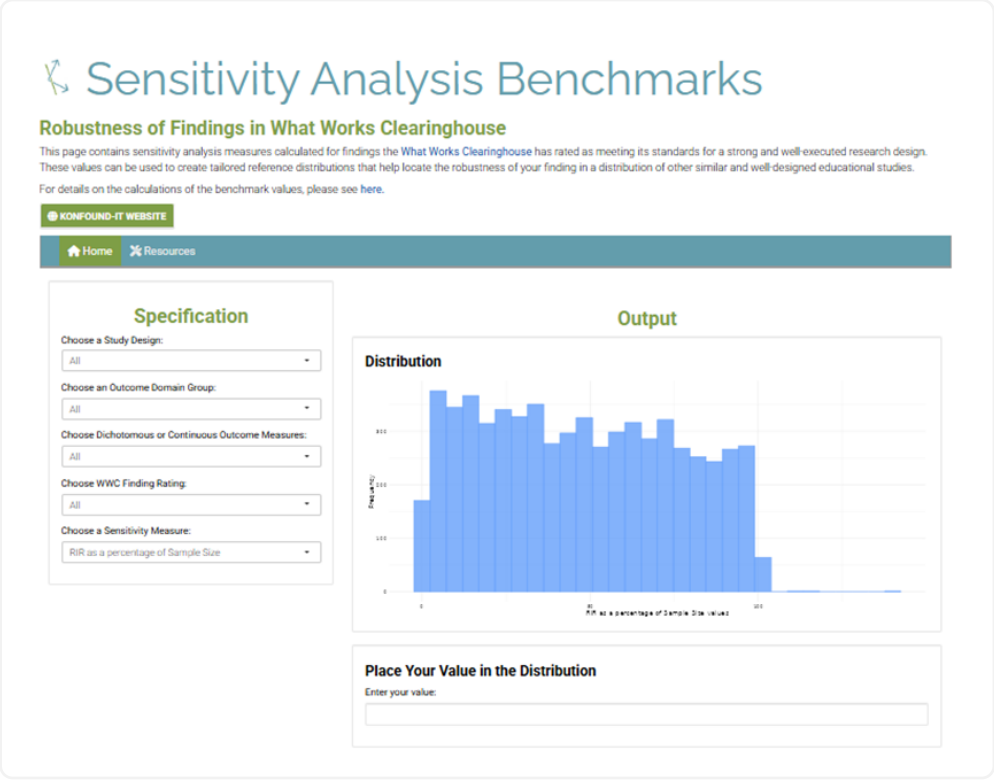
To aid educational researchers in characterizing the robustness of a given finding, we have calculated the RIR values for over 6,300 findings the What Works Clearinghouse (WWC) has rated as meeting its standards for a strong and well-executed research design. All findings meeting the WWC criteria, with or without reservations, were included. Researchers can access these values on the Sensitivity Analysis Benchmarks page of the KonFound-It! website.

After calculating RIR for a finding from their educational study, a researcher can use the Sensitivity Analysis Benchmarks webpage to generate a desired reference distribution, enabling them to locate their own finding in a distribution of other similar and well-designed educational studies. Reference distributions can be tailored to match the finding on the following dimensions: study design (RCT or Quasi-Experiment), outcome domain, continuous or dichotomous outcome, WWC rating, etc., and sensitivity measure. Thus, a researcher could state for example that their RIR, as a percentage, was greater than ___ percent of studies in the WWC that match the same criteria

3.5.1 Sensitivity Analysis Benchmarks: A Step-by-Step Guide

3.5.1.1 Access

To access the Sensitivity Analysis Benchmarks, go https://konfound-project.shinyapps.io/wwc-sensitivity-benchmark/. As of the release of this practical guide, the benchmarks were calculated with version 1.0.2 of the konfound R package using findings available on the WWC website as of September 13, 2023.



3.5.1.2 Generating a reference distribution of RIR

To generate a corresponding distribution of RIR, follow the steps illustrated below:

Step 1

Choose a study design: “All,” “Randomized Controlled Trial,” or “Quasi-Experiment.”

Step 2

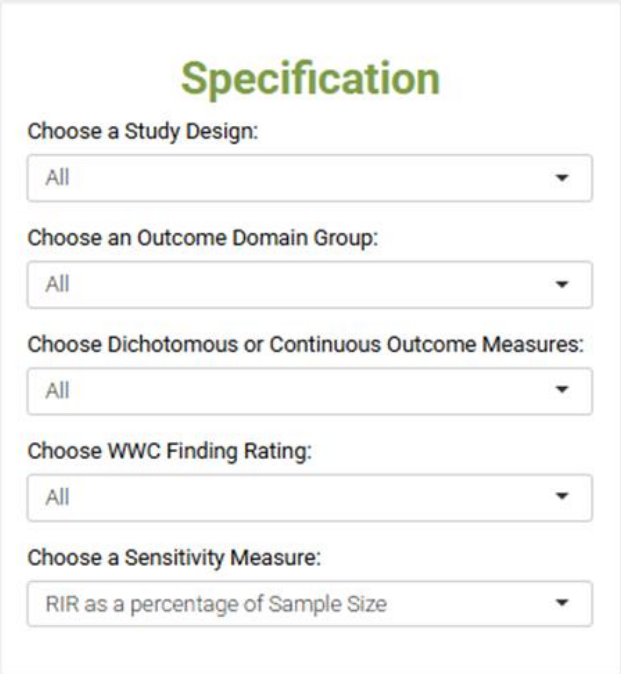
Choose an outcome domain group: “Academic Readiness, Knowledge, or Skills,” “College Readiness, Progress, and Completion,” “Pre-K-12 Progress and Completion,” “School Leader Outcomes,” “School Outcomes and Educational Opportunity,” “Social, Emotional, Behavioral, and Mental Health Outcomes,” “Teacher Outcomes,” or “Workforce Outcomes.”

Step 3

Choose a outcome measure type: “All,” “continuous,” or “dichotomous.”

Step 4:

Choose WWC Finding Rating: “All,” “Meets WWC standards without reservations,” or “Meets WWC standards with reservations.”



Step 5:

Choose a sensitivity measure: “Robustness of Inferences to Replacement (RIR),” “RIR as a percentage of Sample Size,” “Fragility (dichotomous only),” or “Unselected.”

3.5.1.3 An Example: Yeager et al. (2015)

As an example, consider Yeager et al.'s (2015) study in Section 3.2 above. The calculated RIR for the estimated effect of a growth mindset intervention in a RCT on core course GPAs among lower-achieving adolescents is 2,603 (41.19% of total sample 6,320). This following example of sensitivity analysis benchmarking focuses on using RIR as a percentage of sample size.

Step 1

choose "Randomized Controlled Trial" as study design

Step 2

choose an outcome domain group: "Academic Readiness, Knowledge, or Skills," as outcome domain group

Step 3

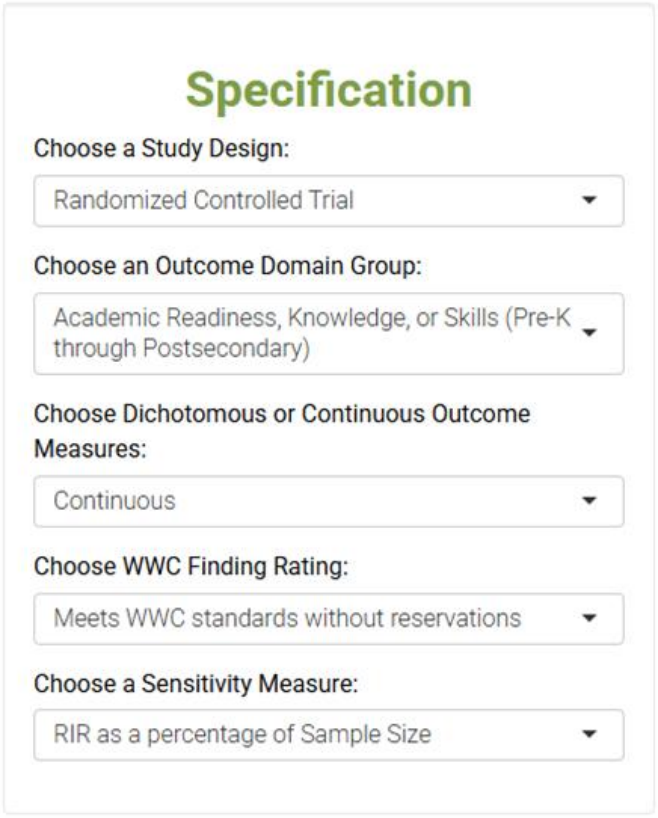
choose "continuous" as outcome measure

Step 4

enter the number of covariates included in the model other than the predictor of interest

Step 5

choose "RIR as a percentage of Sample Size" as sensitivity measure



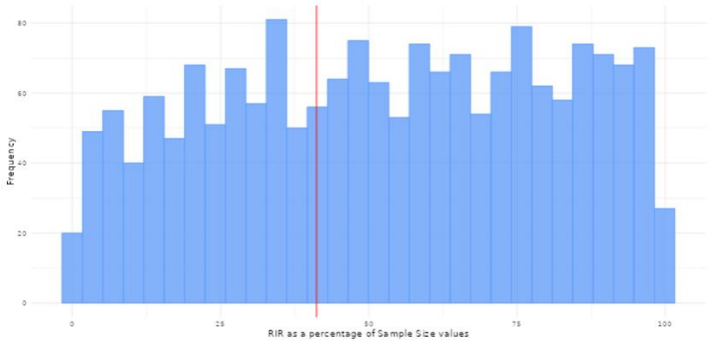
3.5.1.4 Output and Interpretation

The **Output** presents a graph showing a reference distribution of RIR as a percentage of sample size.

Enter Yeager's et al.'s (2015) RIR as a percentage of sample size here.

Output

Distribution



Place Your Value in the Distribution

Enter your value:

41.19

Your value of 41.19 is equal to or greater than 37% of the values in the selected reference distribution.

Descriptive Statistics

Statistic	Value
Count	1798.00
Minimum	0.00
1st Quartile	29.00
Median	53.00
Mean	52.62
3rd Quartile	76.00
Maximum	100.00
Standard Deviation	28.20

Suggested Interpretation:

A suggested statement for interpreting the sensitivity analysis benchmark for the calculated RIR of Yeager et al.'s (2015) estimated effect of growth mindset intervention reads: "The calculated RIR value of 2,603 for the estimated effect of growth mindset intervention in an RCT on core course GPA, when expressed as a percentage of the sample size (41.19%), is equal or greater than 37% of the values in an RIR reference distribution using findings from the What Works Clearinghouse (WWC). This reference distribution is based on 1,798 RCT findings related to academic readiness, knowledge, or skills outcomes and rated as meeting WWC standards without reservations."

Appendix A

Overview of ITCV and RIR Techniques[6]

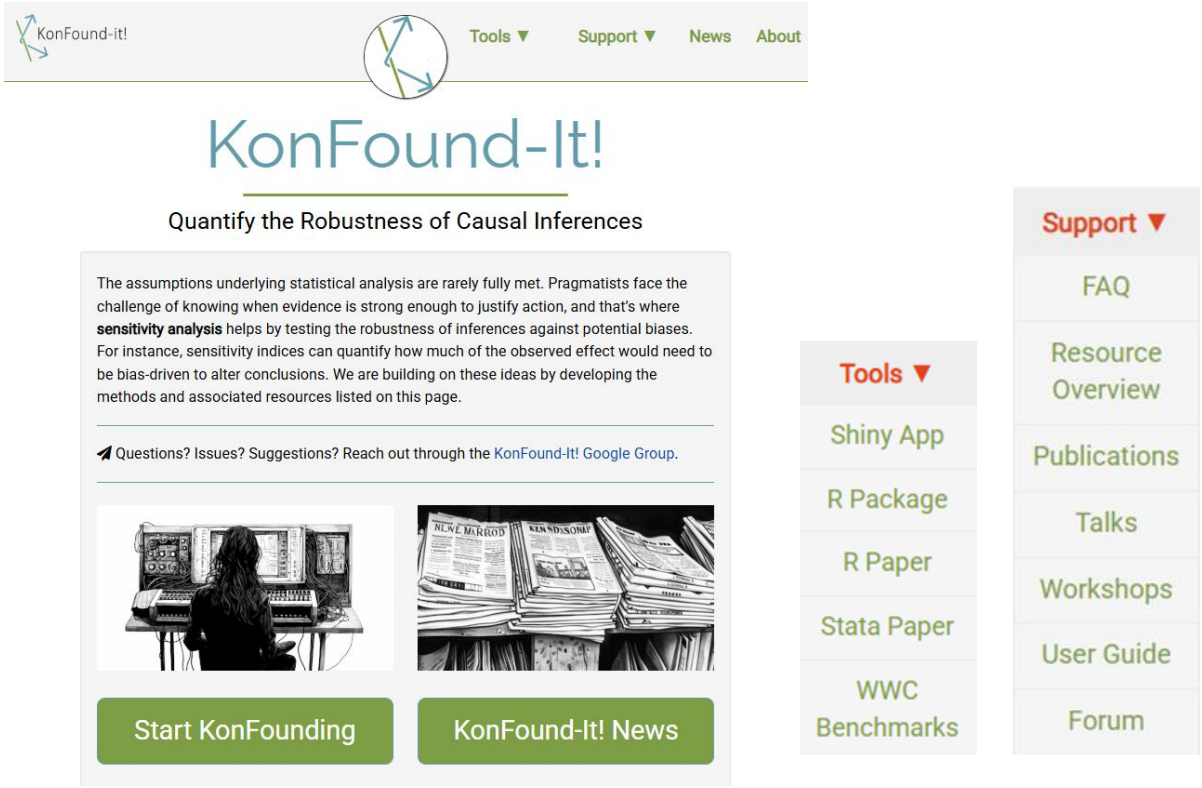
Function	Assumptions	Threshold	Output	Software	Step-by-Step Guide
Impact Threshold of a Confounding Variable (ITCV) Frank (2000), Xu et al. (2019)	Estimate and standard error change when confounding variable is added; Linear model	Statistical significance or any partial correlation	ITCV; component correlations; Unconditional correlations (forthcoming)	Shiny app, R, Stata, spreadsheet	See Section 2.2
Robustness of Inference to Replacement Frank & Min (2007), Frank et al. (2013)	Standard error does not change when cases are replaced	Statistical significance or effect size	% bias to nullify and inference; % of cases to replace with cases with 0 effect	Shiny app, R, Stata, spreadsheet	See Section 3.2
RIR for 2x2 table Frank et al. (2021)		Statistical significance or effect size	% of cases to replace with cases with 0 effect (RIR); Number of cases to switch from treatment success to treatment failure (Fragility). Other switches possible	Shiny app, R, Stata, spreadsheet	See Section 3.4
Robustness of Inference for replacement for logistic regression Based on Frank et al. (2021)	Initial inputs can be converted to 2x2 table (adjustment to standard error may be required)	Statistical significance or effect size	% of cases to replace with cases with 0 effect (RIR); Number of cases to switch from treatment success to treatment failure (Fragility). Other switches possible	Shiny app, R, Stata, spreadsheet	See Section 3.3

[6] Updated as of April 14, 2025

Appendix B

KonFound-It Website

Latest news, resources, and training of ITCV and RIR can be found on KonFound-It! website (https://konfound-it.org/).



REFERENCES

Acharya, A., Blackwell, M., & Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review, 110*(3), 512–529.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of political economy, 113*(1), 151–184.

An, W., & N. Glynn, A. (2021). Treatment effect deviation as an alternative to blinder–oaxaca decomposition for studying social inequality. *Sociological Methods & Research, 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. *Journal of Clinical Epidemiology, 82*, 128–136.

Ansari, A., & Gottfried, M. A. (2021). The grade-level and cumulative outcomes of absenteeism. *Child Development, 92*(4), e548–e564.

Atal, I., Porcher, R., Boutron, I., & Ravaud, P. (2019). The statistical significance of meta-analyses is frequently fragile: Definition of a fragility index for meta-analyses. *Journal of Clinical Epidemiology, 111*, 32–40.

Baer, B. R., Gaudino, M., Charlson, M., Fremes, S. E., & Wells, M. T. (2021). Fragility indices for only sufficiently likely modifications. *Proceedings of the National Academy of Sciences, 118*(49), e2105254118. https://doi.org/10.1073/pnas.2105254118

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

Bulmer, M. (2015). *The uses of social research (Routledge Revivals): Social investigation in public policy making*. Routledge.

Broda, M., Yun, J., Schneider, B., Yeager, D. S., Walton, G. M., & Diemer, M. (2018). Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *Journal of Research on Educational Effectiveness, 11*(3), 317–338.

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. *Statistics in Medicine, 23*(5), 749–767.

Busenbark, J. R., Frank, K. A., Maroulis, S. J., Xu, R., & Lin, Q. (2021). Quantifying the robustness of empirical inferences in strategic management: The impact threshold of a confounding variable

and robustness of inference to replacement. In *Research Methodology in Strategy and Management* (pp. 123–150). (Research Methodology in Strategy and Management; Vol. 13). Emerald Group Holdings Ltd.

Burawoy, M. (2005). For public sociology. *American sociological review, 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.

Chen, Z. (2020). *Quantifying the bias of standard error estimates due to omitted cluster levels in complex multilevel data: A sensitivity analysis for empirical researchers* [Doctoral dissertation, Michigan State University]. ProQuest Dissertations & Theses Global.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., & Syrgkanis, V. (2021). *Long story short: Omitted variable bias in causal machine learning*. arXiv. https://arxiv.org/html/2112.13398v5

Chua, Y. T. (2024). "We want you!" Applying social network analysis to online extremist communities. *Terrorism and Political Violence*. Advance online publication. https://doi.org/10.1080/09546553.2024.2304800

Chung, M. G., Dietz, T., & Liu, J. (2018). Global relationships between biodiversity and nature-based tourism in protected areas. *Ecosystem Services, 34*, 11–23.

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82*(1), 39–67.

Ciobanu, C. (2024). The electoral risks of austerity. *European Journal of Political Research, 63*, 348–69

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society, 5*(2), 171–176.

Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences*. Psychology press.

Cook, T. D., & Campbell, D. T. & Shadish, W. R. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (pp. xxi, 623). Houghton, Mifflin and Company.

Copas, J. B., & Li, H. G. (1997). Inference for Non-Random Samples. Journal of the Royal Statistical Society, Series B (Methodological), 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. *Journal of the National Cancer institute, 22*(1), 173–203.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

Desmond, M., & Kimbro, R. T. (2015). Eviction's fallout: housing, hardship, and health. *Social forces, 94*(1), 295–324.

Dietz, T., Frank, K. A., Whitley, C. T., Kelly, J., & Kelly, R. (2015). Political influences on greenhouse gas emissions from US states. *Proceedings of the National Academy of Sciences, 112*(27), 8254–8259.

DiPrete, T. A., & Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology, 34*, 271–310.

Dorie, V., Harada, M., Carnegie, N. B., & Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine, 35*(20), 3453–3470.

Fisher, R. A. (1936). *Statistical methods for research workers* (6th ed.). Oliver and Boyd.

Forrester, L. A., Jang, E., Lawson. M. M., Capi, A., & Tyler, W. K. (2020). Statistical fragility of surgical and procedural clinical trials in orthopaedic oncology. *Journal of the American Academy of Orthopaedic Surgeons Global Research and Reviews, 4*(6), e19.00152. doi: 10.5435/JAAOSGlobal-D-19-00152

Frank, K. A. (2000). Impact of a confounding variable on the inference of a regression coefficient. *Sociological Methods and Research, 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, September 21-24). *Exact calculation of coefficient of proportionality including evaluation of Oster's δ*, corresponding bounds, and alternatives* [Conference presentation]. Society for Research on Educational Effectiveness Conference, Washington, DC, United States.

*Frank, K. A., *Lin, Q., *Maroulis, S., *Mueller, A. S., Xu, R., Rosenberg, J. M., Hayter, C. S., Mahmoud, R. A., Kolak, M., Dietz, T., & Zhang, L. (2021). Hypothetical case replacement can be used to quantify the robustness of trial results. *Journal of Clinical Epidemiology, 134*, 150–159. *authors listed alphabetically.

*Frank, K. A., *Lin, Q., *Maroulis, S., *Mueller, A. S., Xu, R., Rosenberg, J. M., Hayter, C. S., Mahmoud, R. A., Kolak, M., Dietz, T., & Zhang, L. (2022). Response to "three comments on the RIR method". *Journal of Clinical Epidemiology, 146*, 124–127.

Frank, K. A., Lin, Q., Xu, R., Maroulis, S., & Mueller, A. (2023). Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science. *Social Science Research, 110*, 1–18.

Frank, K. A., Maroulis, S., Duong, M., & Kelcey, B. (2013). What would it take to change an inference?: Using Rubin's causal model to interpret the robustness of causal inferences. *Education Evaluation and Policy Analysis, 35*(4), 437–460.

*Frank, K. A., & *Min, K. (2007). Indices of robustness for sample representation. *Sociological Methodology, 37*(1), 349–392. * co first authors.

Frank K. A., Muller, C., & Mueller, A. S. (2013). The embeddedness of adolescent friendship nominations: The formation of social capital in emergent network structures. *American Journal of Sociology, 119*, 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? *Educational Evaluation and Policy Analysis, 30*(1), 3–30.

Franks, A., D'Amour, A., & Feller, A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association, 115*, 1730–1746.

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 Behavioral Research, 51*(5), 681–697.

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine, 83*, 27–33

Gastwirth, J. L., Krieger, A. M., & Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika, 85*(4), 907-920.

Golec de Zavala, A., Bierwiaczonek, K., Baran, T., Keenan, O., & Hase, A. (2021). The COVID-19 pandemic, authoritarianism, and rejection of sexual dissenters in Poland. *Psychology of Sexual Orientation and Gender Diversity, 8*(2), 250–260.

Harding, D. J. (2003). Counterfactual models of neighborhood effects: The effect of neighborhood poverty on dropping out and teenage pregnancy. *American Journal of Sociology, 109*(3), 676–719.

Herold, K. C., Bundy, B. N., Long, S. A., Bluestone, J. A., DiMeglio, L. A., Dufort, M. J., Gitelman, S. E., Gottlieb, P. A., Krischer, J. P., Linsley, P. S., Marks, J. B., Moore, W., Moran, A., Rodriguez, H., Russell, W. E., Schatz, D., Skyler, J. S., Tsalikian, E., Wherrett, D. K., … Greenbaum, C. J. (2019). An anti-cd3 antibody, teplizumab, in relatives at risk for type 1 diabetes. *New England Journal of Medicine, 381*(7), 603–613.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology, 2*(3), 259–278.

Hoke, M. K., & Boen, C. E. (2021). The health impacts of eviction: Evidence from the national longitudinal study of adolescent to adult health. *Social Science and Medicine*. Advance online publication. https://doi.org/10.1016/j.socscimed.2021.113742

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–70.

Hong, G., Qin, X., & Yang, F. (2018). Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics, 43*(1), 32–56.

Hong, G., Yang, F., & Qin, X. (2021). Did you conduct a sensitivity analysis? A new weighting-based approach for evaluations of the average treatment effect for the treated. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184*(1), 227–254.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 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. *The Annals of Applied Statistics, 4*(2), 849–870.

Imbens, G. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review, 93*(2), 126–132.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309–334.

Jesson, A., Mindermann, S., Gal, Y., & Shalit, U. (2021). Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *Proceedings of the 38th International Conference on Machine Learning, PMLR 139*, 4829–4838.

Kallus, N., Mao, X., & Zhou, A. (2019). Interval estimation of individual-level causal effects under unobserved confounding. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR 89*, 2281–2290.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational Researcher, 49(4), 241–253.

Lapatinas, A., & Litina, A. (2019). Intelligence and economic sophistication. *Empirical Economics, 57*(5), 1731–1750.

Lash, T. L., Fox, M. P., & Fink, A. K. (2009). *Applying quantitative bias analysis to epidemiologic data*. Springer.

Li, T., & Frank, K. (2022). The probability of a robust inference for internal validity. *Sociological Methods & Research, 51*(4), 1947–1968.

Lin, Q. (2019). *Quantifying strength of evidence in education research: Accounting for spillover, heterogeneity, and mediation* [Doctoral dissertation, Michigan State University]. ProQuest Dissertations & Theses Global.

Lin, Q., Nuttall, A. K., Zhang, Q., & Frank, K. A. (2023). How do unobserved confounding mediators and measurement error impact estimated mediation effects and corresponding statistical

inferences? Introducing the R package ConMed for sensitivity analysis. *Psychological Methods, 28*(2), 339–358.

Liu, X., & Wang, L. (2021). The impact of measurement error and omitting confounders on statistical inference of mediation effects and tools for sensitivity analysis. *Psychological Methods, 26*(3), 327–342.

Mayer, A., Malin, S., McKenzie, L., Peel, J., & Adgate, J. (2021). Understanding self-rated health and unconventional oil and gas development in three Colorado communities. *Society & Natural Resources, 34*(1), 60–81.

Martino, P., Vanacker, T., Filatotchev, I., & Bellavitis, C. (2024). (De)centralized governance and the value of platform-based new ventures: The moderating role of teams and transparency. *Small Business Economics.* Advance online publication. https://doi.org/10.1007/s11187-024-00964-6

Mauro, R. (1990). Understanding LOVE (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin, 108*(2), 314.

Middleton, J. A., Scott, M. A., Diakow, R., & Hill, J. L. (2016). Bias amplification and bias unmasking. *Political Analysis, 24*(3), 307–323.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.

National Research Council. (2002). *Scientific research in education*. The National Academies 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. *British Medical Journal, 317*(7176), 1239–1242.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics, 37*(2), 187–204.

Paxton, P., Velasco, K., & Ressler, R. W. (2020). Does use of emotion increase donations and volunteers for nonprofits? *American Sociological Review, 85*(6), 1051–1083.

Plümper, T., & Traunmüller, R. (2020). The sensitivity of sensitivity analysis. *Political Science Research and Methods, 8*(1), 149–159.

Pyne J. (2019). Suspended attitudes: Exclusion and emotional disengagement from school. *Sociology of Education, 92*(1), 59–82.

Rickard, M., Keefe, D. T., Drysdale, E., Erdman, L., Hannick, J. H., Milford, K., Santos, J. D., Mistry, N., Koyle, M. A., & Lorenzo, A. J. (2020). Trends and relevance in the bladder and bowel dysfunction literature: PlumX metrics contrasted with fragility indicators. *Journal of Pediatric Urology, 16*(4), 477.e1–477.e7. https://doi.org/10.1016/j.jpurol.2020.06.015

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 M. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 1–94). Springer.

Rosenbaum, P. R. (2002). *Observational studies*. Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological), 45*(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. *Journal of Educational Statistics, 11*(3), 207–224.

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

Rubin, D. B. (1986). Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference." *Journal of American Statistical Association, 83*, 396.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference, 25*, 279–292.

Saw, G. K., Kunisaki, L., Lin, S., Culbertson, R., & Megyesi-Brem, K. (2025). Opportunities for deeper learning and adolescents' creative thinking competency in STEM: The mediating role of situated expectancy-value beliefs. *International Journal of Science and Mathematics Education*. Advance online publication. https://doi.org/10.1007/s10763-025-10567-6

Saw, G. K., Lin, S., Kunisaki, L., Culbertson, R., & Megyesi-Brem, K. (2025). Adolescents' perceived opportunities for creative thinking, creative thinking competency belief and career interest in STEM: Joint consideration of situated expectancy-value beliefs and gender. *Journal of Research in Science Teaching*. Advance online publication. http://doi.org/10.1002/tea.22032

Scharfstein, D. O., Nabi, R., Kennedy, E. H., Huang, M. Y., Bonvini, M., & Smid, M. (2021). Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *Biometrics, 80*(4), ujae106. https://doi.org/10.1093/biomtc/ujae106

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects. Using experimental and observational design.* American Educational Research Association.

Schwartz, G. L., Feldman, J. M., Wang, S. S., & Glied, S. A. (2022). Eviction, healthcare utilization, and disenrollment among New York City Medicaid patients. *American Journal of Preventive Medicine, 62*(2), 157–164.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association, 103*(484), 1334–1344.

Shen, X., & Zhang, J. (2024). Are female leaders hiring more female workers? Evidence from developing countries. *Applied Economics*. Advance online publication.
https://doi.org/10.1080/00036846.2024.2387872

Shin, T., & You, J. (2023). Faults and faultlines: The effects of board faultlines on CEO dismissal. *Journal of Management, 49*(4), 1344–1393.

Slee, G., & Desmond, M. (2023). Eviction and voter turnout: The political consequences of housing instability. *Politics & Society, 51*(1), 3–29.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250–267.

Thiede, B. C., Kim, H., & Slack, T. (2017). Marriage, work, and racial inequalities in poverty: Evidence from the United States. *Journal of Marriage and Family, 79*(5), 1241–1257.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. II. The estimation of magnitudes. *Psychological Review, 8*(4), 384–395.

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics, 39*(6), 478–501.

Treglia, D., Byrne, T., & Tamla Rai, V. (2023). Quantifying the impact of evictions and eviction filings on homelessness rates in the United States. *Housing Policy Debate*. Advance online publication. https://doi.org/10.1080/10511482.2023.2186749

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. *Annals of Internal Medicine, 167*(4), 268–274.

Walsh, M., Srinathan, S. K., McAuley, D. F., Mrkobrada, M., Levine, O., Ribic, C., Molnar, A. O., Dattani, N. D., Burke, A., Guyatt, G., Thabane, L., Walter, S. D., Pogue, J., & Devereaux, P. J. (2014). The statistical significance of randomized controlled trial results is frequently fragile: A case for a Fragility Index. *Journal of Clinical Epidemiology, 67*(6), 622–628.

Walter, S. D., Thabane, L., & Briel, M. (2020). The fragility of trial results involves more than statistical significance alone. *Journal of Clinical Epidemiology, 124*(6), 622–628.

Weiss, C. H., & Bucuvalas, M. J. (1980). *Social science research and decision-making*. Columbia University Press

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* (2nd ed.). MIT press.

Xu, R., & Frank, K. A. (2021). Sensitivity analysis for network observations with applications to inferences of social influence effects. *Network Science, 9*(1), 73–98.

Xu, R., Frank, K. A., Maroulis, S. J., & Rosenberg, J. M. (2019). konfound: Command to quantify robustness of causal inferences. *The Stata Journal, 19*(3), 523–550.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., … Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*(7774), 364–369.