Motivation
Statistical inferences are often challenged because of uncontrolled bias. There may be bias due to uncontrolled confounding variables or non-random selection into a sample. We will answer the question about what it would take to change an inference by formalizing the sources of bias and quantifying the discourse about causal inferences in terms of those sources. For example, we will transform challenges such as “But the inference of a treatment effect might not be valid because of pre-existing differences between the treatment groups” to questions such as “How much bias must there have been due to uncontrolled pre-existing differences to make the inference invalid?”

Approaches
In part I we will use Rubin’s causal model to interpret how much bias there must be to invalidate an inference in terms of replacing observed cases with counterfactual cases or cases from an unsampled population. This generates statements such as “One would have to replace qqq% of the cases with cases with no effect to invalidate the inference.” In part II, we will quantify the robustness of causal inferences in terms of correlations associated with unobserved variables or in unsampled populations. This generates statements such as “An omitted variable would have to be correlated at rrr with the treatment and outcome to invalidate the inference.” Calculations for bivariate and multivariate analysis will be presented using an app: http://konfound-it.com as well as macros in STATA and R and a spreadsheet for calculating indices [KonFound-it!].

Format
The format will be a mixture of presentation, individual exploration, and group work. Participants may include graduate students and professors, although all must be comfortable with basic regression and multiple regression. Participants should bring their own laptop, or be willing to work with another student who has a laptop. Participants may choose to bring to the course an example of an inference from a published study or their own work, as well as data analyses they are currently conducting.
How it all comes together: Integrated statement

I study the social structures of organizations and systems. How do the structures evolve, how do they affect what an organization does? I often use networks and multilevel models. Typically I study small systems, in particular schools as organizations, a topic that is part of the sociology of education.

Check out the new teachers in social media website!

Causal Inference and Robustness Indices

Summary statement

For Impact Threshold of a Confounding Variable and % Bias to Validate an Inference (can be applied to output from Stata, SAS, SPSS, R)

R Shiny app KonFound-it!

Presentations

quick examples [pdf of quick examples] 30 Minutes
powerpoint for combined frameworks 3 hours
powerpoint for replacement of cases [pdf of replacement of cases] 2 hours
  full publishable write-up for case replacement
powerpoint for correlation framework [pdf of correlation framework] 1 hour
  full publishable write-up for correlation
powerpoint for Case Replacement for Logistic Regression 10 minutes
powerpoint for ordered thresholds relative to transaction costs 10 minutes
powerpoint for alternative scenarios and related techniques 15 minutes
powerpoint for new directions

Quantifying What it Would Take to Change an Inference: Toward a Pragmatic Sociology (ASA Duncan lecture August 2018) 1 hour
powerpoint for comparison of frameworks [pdf of comparison of frameworks]
published empirical examples
The value of controlling for pretests
Motivation

• Inferences uncertain
  • it’s causal inference not determinism
• Do you have enough information to make a decision?
• Instead of “you didn’t control for xxx”
  • Personal: Granovetter to Fernandez
• What would xxx have to be to change your inference
  • Promotes scientific discourse
  • Informs pragmatic policy
overview

Replacement Cases Framework (40 minutes to reflection)
- Thresholds for inference and % bias to invalidate and inference
- The counterfactual paradigm
- Application to concerns about non-random assignment to treatments
- Application to concerns about non-random sample
  - Reflection (10 minutes)
Examples of replacement framework
- Internal validity example: Effect of kindergarten retention on achievement
  (40 minutes to break)
- External validity example: effect of Open Court curriculum on achievement
Review and Reflection

Extensions
- Extensions of the framework
- Exercise and break (20 minutes)

Correlational Framework (25 minutes to exercise)
- How regression works
- Impact of a Confounding variable
- Internal validity: Impact necessary to invalidate an inference
- Example: Effect of kindergarten retention on achievement
- Exercise (25 minutes)
- External validity (30 minutes)
  - combining estimates from different populations
  - example: effect of Open Court curriculum on achievement
Conclusion (10 minutes)
Quick Survey

Can you make a causal inference from an observational study?
Answer: Quantifying the Discourse

Can you make a causal inference from an observational study?

Of course you can. You just might be wrong. It’s causal *inference*, not determinism.

But what would it take for the inference to be wrong?
I: Replacement of Cases Framework

How much bias must there be to invalidate an inference?

Concerns about Internal Validity

• What percentage of cases would you have to replace with counterfactual cases (with zero effect) to invalidate the inference?

Concerns about External Validity

• What percentage of cases would you have to replace with cases from an unsampled population (with zero effect) to invalidate the inference?
What Would It Take to Change an Inference? Using Rubin's Causal Model to Interpret the Robustness of Causal Inferences

Abstract

We contribute to debate about causal inferences in educational research in two ways. First, we quantify how much bias there must be in an estimate to invalidate an inference. Second, we utilize Rubin's causal model (RCM) to interpret the bias necessary to invalidate an inference in terms of sample replacement. We apply our analysis to an inference of a positive effect of Open Court Curriculum on reading achievement from a randomized experiment, and an inference of a negative effect of kindergarten retention on reading achievement from an observational study. We consider details of our framework, and then discuss how our approach informs judgment of inference relative to study design. We conclude with implications for scientific discourse.

Keywords: causal inference; Rubin’s causal model; sensitivity analysis; observational studies

Figure 1
Estimated Treatment Effects in Hypothetical Studies A and B Relative to a Threshold for Inference

% bias necessary to invalidate the inference

Threshold
Quantifying the Discourse: Formalizing

Bias Necessary to Invalidate an Inference

\( \delta \) = a population effect,
\( \hat{\delta} \) = the estimated effect, and
\( \delta^\# \) = the threshold for making an inference

An inference is invalid if:

\[ \hat{\delta} > \delta^\# > \delta \]  \hspace{1cm} (1)

An inference is invalid if the estimate is greater than the threshold while the population value is less than the threshold.

Defining bias as \( \hat{\delta} - \delta \), (1) implies an inference is invalid if and only if:

\[ \text{bias}(\hat{\delta}) > \hat{\delta} - \delta^\# \]  \hspace{1cm} (2)

Expressed as a proportion of the estimate, inference invalid if:

\[ \% \text{bias}(\hat{\delta}) > \frac{(\hat{\delta} - \delta^\#)}{\hat{\delta}} = 1 - \frac{\delta^\#}{\hat{\delta}} \]
The full Algebra for Exceeding a Threshold

inference invalid if
\( \hat{\delta} > \delta^* > \delta \),
subtract \( \hat{\delta} \)
\( \hat{\delta} - \hat{\delta} > \delta^* - \hat{\delta} > \delta - \hat{\delta} \)
\( \Rightarrow 0 > \delta^* - \hat{\delta} > \delta - \hat{\delta} \)
multiply by -1
\( \Rightarrow 0 < \hat{\delta} - \delta^* < \hat{\delta} - \delta \)
subtract: bias = \( \hat{\delta} - \delta \)
\( \Rightarrow 0 < \hat{\delta} - \delta^* < \text{bias} \)
\( \Rightarrow \text{bias} > \hat{\delta} - \delta^* > 0 \)
as % of estimate, divide by \( \hat{\delta} \)
\( \frac{\text{bias}}{\hat{\delta}} > \frac{\hat{\delta} - \delta^*}{\hat{\delta}} > 0 \)
\( \Rightarrow \% \text{bias} > 1 - \frac{\delta^*}{\delta} > 0 \)
Figure 1
Estimated Treatment Effects in Hypothetical Studies A and B Relative to a Threshold for Inference $\delta^\#$

\[
\% \text{ bias}(\hat{\delta}) \text{ to invalidate} = \frac{(\hat{\delta} - \delta^\#)}{\hat{\delta}} = 1 - \frac{\delta^\#}{\hat{\delta}} = 1 - \frac{4}{6} = \frac{1}{3} = 33\%
\]

\[
1 - \frac{4}{8} = \frac{1}{2} = 50\%
\]
Interpretation of % Bias to Invalidate an Inference

% Bias is intuitive
Relates to how we think about statistical significance
   Better than “highly significant” or “barely significant”
But need a framework for interpreting
Sensitivity vs Robustness

Sensitivity of estimate assess how estimates change as a result of alternative analyses
- Different covariates
- Different sub-samples
- Different estimation techniques

Robustness of the inference quantifies what it would take to change an inference based on hypothetical data or conditions
- Unobserved covariates
- Unobserved samples
Framework for Interpreting % Bias to Invalidate an Inference: Rubin’s Causal Model and the Counterfactual

1) I have a headache
2) I take an aspirin (treatment)
3) My headache goes away (outcome)

Q) Is it because I took the aspirin?
A) We’ll never know – it is counterfactual – for the individual

This is the Fundamental Problem of Causal Inference
Approximating the Counterfactual with Observed Data

But how well does the observed data approximate the counterfactual? Difference between counterfactual values and observed values for the control implies the treatment effect of 1 is overestimated as 6 using observed control cases with mean of 4.
Using the Counterfactual to Interpret % Bias to Invalidate the Inference: Replacement with Average Values

How many cases would you have to replace with zero effect counterfactuals to change the inference? Assume threshold is 4 ($\delta^* = 4$):

$$1 - \frac{\delta^*}{\hat{\delta}} = 1 - \frac{4}{6} = \frac{1}{3}$$

The inference would be invalid if you replaced 33% (or 1 case) with counterfactuals for which there was no treatment effect. New estimate = \((1-%\text{ replaced})\hat{\delta} + %\text{ replaced(0 effect)}\) = \((1-0.33)\hat{\delta} = 0.66\hat{\delta} = 0.66(6) = 4\)
Which Cases to Replace?

• Think of it as an expectation: if you randomly replaced 1 case, and repeated 1,000 times, on average the new estimate would be 4.
• Assumes constant treatment effect.
• Conditioning on covariates and interactions already in the model.
• Assumes cases carry equal weight.
Figure 1
Estimated Treatment Effects in Hypothetical Studies A and B Relative to a Threshold for Inference

% bias(\hat{\delta}) to invalidate = \left(1 - \frac{\hat{\delta}^#}{\hat{\delta}}\right) = \left(1 - \frac{4}{6}\right) = \frac{1}{3} = 33\%

To invalidate the inference, replace 33% of cases with counterfactual data with zero effect.
Review & Reflection

Review of Framework

Pragmatism → thresholds
How much does an estimate exceed the threshold → % bias to invalidate the inference

Interpretation: Rubin’s causal model
• internal validity: % bias to invalidate → number of cases that must be replaced with counterfactual cases (for which there is no effect)

Reflect
Which part is most confusing to you?
Is there more than one interpretation?
Discuss with a partner or two
Example of Internal Validity from Observational Study:
The Effect of Kindergarten Retention on Reading and Math Achievement
(Hong and Raudenbush 2005)

1. What is the average effect of kindergarten retention policy? (Example used here)

Should we expect to see a change in children’s average learning outcomes if a school changes its retention policy?

Propensity based questions (not explored here)

2. What is the average impact of a school’s retention policy on children who would be promoted if the policy were adopted?

Use principal stratification.

Data

Early Childhood Longitudinal Study Kindergarten cohort (ECLSK)
US National Center for Education Statistics (NCES).
Nationally representative
Kindergarten and 1st grade
observed Fall 1998, Spring 1998, Spring 1999

Student
background and educational experiences
Math and reading achievement (dependent variable)
experience in class

Parenting information and style

Teacher assessment of student

School conditions

Analytic sample (1,080 schools that do retain some children)
471 kindergarten retainees
7,168 promoted students
Estimated Effect of Retention on Reading Scores (Hong and Raudenbush)
Possible Confounding Variables
(note they controlled for these)

Gender
Two Parent Household
Poverty
Mother’s level of Education (especially relevant for reading achievement)
Extensive pretests
  measured in the Spring of 1999 (at the beginning of the second year of school)
  standardized measures of reading ability, math ability, and general knowledge;
  indirect assessments of literature, math and general knowledge that include aspects of a child’s process as well as product;
teacher’s rating of the child’s skills in language, math, and science
With no statistical adjustment for selection bias, the mean differences between the 471 kindergarten retainees and the 7,168 promoted at-risk students in retention schools were −18.51 in the reading outcome and −11.06 in the math outcome. Given the pretreatment imbalance between the two groups, these mean differences were likely to be negatively biased if viewed as estimates of the causal effects of being retained.

Next, we computed the mean differences in the reading and math outcomes between the retained students and the at-risk promoted students in the retention schools within each individual-level propensity stratum. The results are displayed in Tables 9 and 10. After a year of treatment, the kindergarten retainees generally reached a lower achievement status in both reading and mathematics than did their comparable promoted peers. The within-stratum mean differences in reading ranged from −0.96 to −12.14, while those in mathematics ranged from 1.57 to −7.98.

To generate a more conclusive answer to the causal question, we specified a two-level hierarchical linear model for estimating the average retention effect, $\delta$, on the learning outcomes of at-risk children attending retention schools.

We divided the sample of at-risk students attending retention schools into 15 strata on the basis of the logit of \( \hat{q} \) as estimated from Equation 3. Table 8 compares the within-stratum distribution of the logit of this propensity score between the retained group and the promoted group. The eight retainees in the last stratum had no matches in the promoted group. Within each of the remaining 14 strata, the two treatment groups had similar distributions of the logit of \( \hat{q} \). The result of hypothesis testing showed no statistically significant between group difference over all the strata in 97% of pretreatment covariates.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>( N )</th>
<th>( M )</th>
<th>( SD )</th>
<th>( N )</th>
<th>( M )</th>
<th>( SD )</th>
<th>Mean diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L = 1 )</td>
<td>9</td>
<td>38.88</td>
<td>9.22</td>
<td>3,044</td>
<td>45.56</td>
<td>7.58</td>
<td>−6.67</td>
</tr>
<tr>
<td>( L = 2 )</td>
<td>12</td>
<td>40.87</td>
<td>11.14</td>
<td>1,654</td>
<td>43.48</td>
<td>7.70</td>
<td>−2.61</td>
</tr>
<tr>
<td>( L = 3 )</td>
<td>14</td>
<td>35.12</td>
<td>12.65</td>
<td>976</td>
<td>41.59</td>
<td>8.48</td>
<td>−6.48</td>
</tr>
<tr>
<td>( L = 4 )</td>
<td>12</td>
<td>35.62</td>
<td>12.68</td>
<td>321</td>
<td>41.77</td>
<td>8.50</td>
<td>−6.15</td>
</tr>
<tr>
<td>( L = 5 )</td>
<td>24</td>
<td>35.64</td>
<td>9.20</td>
<td>440</td>
<td>39.42</td>
<td>8.44</td>
<td>−3.78</td>
</tr>
<tr>
<td>( L = 6 )</td>
<td>23</td>
<td>31.05</td>
<td>9.70</td>
<td>153</td>
<td>39.02</td>
<td>9.41</td>
<td>−7.98</td>
</tr>
<tr>
<td>( L = 7 )</td>
<td>47</td>
<td>33.37</td>
<td>9.44</td>
<td>211</td>
<td>37.19</td>
<td>9.70</td>
<td>−3.81</td>
</tr>
<tr>
<td>( L = 8 )</td>
<td>48</td>
<td>34.49</td>
<td>10.43</td>
<td>143</td>
<td>36.97</td>
<td>10.09</td>
<td>−2.48</td>
</tr>
<tr>
<td>( L = 9 )</td>
<td>45</td>
<td>34.98</td>
<td>10.95</td>
<td>85</td>
<td>34.43</td>
<td>9.17</td>
<td>0.55</td>
</tr>
<tr>
<td>( L = 10 )</td>
<td>48</td>
<td>30.76</td>
<td>8.07</td>
<td>49</td>
<td>34.56</td>
<td>10.37</td>
<td>−3.80</td>
</tr>
<tr>
<td>( L = 11 )</td>
<td>47</td>
<td>30.97</td>
<td>9.56</td>
<td>35</td>
<td>35.63</td>
<td>8.98</td>
<td>−4.66</td>
</tr>
<tr>
<td>( L = 12 )</td>
<td>49</td>
<td>28.34</td>
<td>9.53</td>
<td>16</td>
<td>26.77</td>
<td>12.71</td>
<td>1.57</td>
</tr>
<tr>
<td>( L = 13 )</td>
<td>45</td>
<td>27.04</td>
<td>10.24</td>
<td>8</td>
<td>30.96</td>
<td>8.22</td>
<td>−3.92</td>
</tr>
<tr>
<td>( L = 14 )</td>
<td>38</td>
<td>30.10</td>
<td>8.58</td>
<td>3</td>
<td>37.45</td>
<td>5.66</td>
<td>−7.36</td>
</tr>
<tr>
<td>( L = 15 )</td>
<td>8</td>
<td>27.49</td>
<td>11.53</td>
<td>0</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
</tbody>
</table>

*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.*

Df used in model = 207 + 14 + 2 = 223

Add 2 to include the intercept and retention
Calculating % Bias to Invalidate the Inference

1) Calculate threshold $\delta^#$

   Estimated effect is statistically significant if:
   \[
   \frac{|\text{Estimated effect}|}{\text{standard error}} > |t_{\text{critical}}|
   \]
   \[\Rightarrow |\text{Estimated effect}| > |t_{\text{critical}}| \times \text{standard error} = \delta^#
   \]
   \[\Rightarrow |\text{Estimated effect}| > 1.96 \times 0.68 = 1.33 = \delta^#\]

2) Record $\hat{\delta} = |\text{Estimated effect}| = 9.01$

3) % bias to invalidate the inference is
   \[1 - \delta^#/\hat{\delta} = 1 - 1.33/9.01 = 0.85\]

85% of the estimate would have to be due to bias to invalidate the inference.

You would have to replace 85% of the cases with counterfactual cases with 0 effect of retention on achievement to invalidate the inference.
In R Shiny app KonFound-it!
(konfound-it.com/)

Estimated effect

Replacement of Cases Approach: To invalidate an inference, 85.205% of the estimate would have to be due to bias. This is based on a threshold of -1.333 for statistical significance (alpha = 0.05). To invalidate an inference, 6509 observations would have to be replaced with cases for which the effect is 0.

Number of observations
7639

Number of covariates
223

Take out your phone and try it!!!
% Bias necessary to invalidate inference = \(1 - \frac{\hat{\delta}}{\delta^*}\) 
\(= 1 - 1.33/9.01 = 85\%\)

85% of the estimate must be due to bias to invalidate the inference.

85% of the cases must be replaced with null hypothesis cases to invalidate the inference.
Using the Counterfactual to Interpret % Bias to Invalidate the Inference: Replacement with Average Values

How many cases would you have to replace with zero effect counterfactuals to change the inference? Assume threshold is $\delta^\# = 4$:

$$1 - \frac{\delta^\#}{\hat{\delta}} = 1 - \frac{4}{6} = 0.33 = (1/3)$$

The inference would be invalid if you replaced 33% (or 1 case) with counterfactuals for which there was no treatment effect. New estimate = $(1-% replaced) \hat{\delta} + % replaced (no effect)$ = $(1-.33)6 = .66(6) = 4$
Example Replacement of Cases with Counterfactual Data to Invalidate Inference of an Effect of Kindergarten Retention

- Retained
- Counterfactual: No effect
- Promoted

Legend:
- Original cases that were not replaced
- Replacement counterfactual cases with zero effect
- Original distribution
Interpretation

1) Consider test scores of a set of children who were retained that are considerably lower (9 points) than others who were candidates for retention but who were in fact promoted. No doubt some of the difference is due to advantages the comparable others had before being promoted. But now to believe that retention did not have an effect one must believe that 85% of those comparable others would have enjoyed most (7.2) of their advantages whether or not they had been retained.

This is even after controlling for differences on pretests, mother’s education, etc.

2a) The inference is invalid if we replace 85% cases with cases in which an omitted variable is perfectly correlated with kindergarten retention.

2b) The inference is invalid if we replace 85% of the cases with cases in which an omitted variable is perfectly correlated with the achievement.

3) Could 85% of the children have manifest an effect because of unadjusted differences (even after controlling for prior achievement, motivation and background) rather than retention itself?
Which Cases to Replace?

• Graphics are rigged to make $p = .06$
• Generally: thought experiment of repeating replacement 1,000 times. Average of new estimates will be at the threshold for inference
• If data are weighted (sample weights, IPTW, HLM, Logistic), then if you remove a unit with case weight of 3, then you replace with a unit with a case weight of 3
Evaluation of % Bias Necessary to Invalidate Inference

• 50% cut off– for every case you remove, I get to keep one
• Compare bias necessary to invalidate inference with bias accounted for by background characteristics
  1% of estimated effect accounted for by background characteristics (including mother’s education), once controlling for pretests
  e.g. estimate of retention before controlling for mother’s education is -9.1, after controlling for mother’s education it is -9.01, a change of .1 (or about 1% of the final estimate. The estimate would have to change another 85% to invalidate the inference.
  More than 85 times more unmeasured bias necessary to invalidate the inference
• Compare with % bias necessary to invalidate inference in other studies:
  • Use correlation metric
  • Adjusts for differences in scale
% Bias Necessary to Invalidate Inference based on Correlation to Compare across Studies

\[ r = \frac{t}{\sqrt{(n-q-1) + t^2}} = \frac{-13.25}{\sqrt{(7639 - 223 - 1) + (13.25)^2}} = -0.152 \]

t taken from HLM: \(-9.01/0.68 = -13.25\)
n is the sample size
q is the number of parameters estimated

\[ \text{threshold} = r^* = \frac{t_{\text{critical}}}{\sqrt{(n-q-1) + t_{\text{critical}}^2}} = \frac{-1.96}{\sqrt{(7639 - 223 - 1) + 1.96^2}} = -0.023 \]

Where t is critical value for df>200

% bias to invalidate inference = 1 - (-0.023)/(-0.152) = 85%

Accounts for changes in regression coefficient and standard error
Because \( t(r) = t(\beta) \)
Figure 1
Estimated Treatment Effects in Hypothetical Studies A and B Relative to a Threshold for Inference

% bias(r) to invalidate = \( \frac{(r - r^#)}{r} \) = \( 1 - \frac{r^#}{r} = 1 - \frac{4}{6} = \frac{1}{3} = 33\% \)

scale \( r - r^# \) by \( r \)

\( \Delta \)
### Compare with Bias other Observational Studies

#### Table 2

Quantifying the Robustness of Inferences from Observational Studies

<table>
<thead>
<tr>
<th>Study (author, year)</th>
<th>Predictor of interest</th>
<th>Condition on pretest</th>
<th>Population</th>
<th>Outcome</th>
<th>Estimated effect, standard error, source</th>
<th>Effect size (correlation)</th>
<th>% bias to make inference invalid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effects of kindergarten retention policy on children’s cognitive growth in reading and mathematics (Hong &amp; Raudenbush, 2005)</td>
<td>Kindergarten retention versus promotion</td>
<td>Multiple</td>
<td>7639 kindergarteners in 1080 retention schools in ECLS-K</td>
<td>ECLS-K Reading IRT scale score</td>
<td>9 (.68) Table 11, model based estimate</td>
<td>.67 (.14)</td>
<td>85%</td>
</tr>
<tr>
<td>Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning, (Morgan, 2001)</td>
<td>Catholic versus public school</td>
<td>Single</td>
<td>10835 high school students nested within 973 schools in NELS</td>
<td>NELS Math IRT scale score</td>
<td>.99 (.33) Table 1, (model with pretest + family background)</td>
<td>.23 (.10)</td>
<td>34%</td>
</tr>
<tr>
<td>Effects of teachers’ mathematical knowledge for teaching on student achievement (Hill et al., 2005)</td>
<td>Content knowledge for teacher mathematics</td>
<td>Gain score</td>
<td>1773 third graders nested within 365 teachers</td>
<td>Terra Nova math scale score</td>
<td>2.28 (.75) Table 7, model 1, (third graders)</td>
<td>NA (.16)</td>
<td>36%</td>
</tr>
</tbody>
</table>
% Bias to Invalidate versus p-value: a better language?

Df=973 based on Morgan's analysis of Catholic school effects, Functional form not sensitive to df
Not very Sensitive to Sample Size
Beyond *, **, and ***

• P values
  – sampling distribution framework
  – Must interpret relative to standard errors
  – Information lost for modest and high levels of robustness

• % bias to invalidate
  – counterfactual framework
  – Interpret in terms of case replacement
  – Information along a continuous scale
<table>
<thead>
<tr>
<th>Study (1st author)</th>
<th>Random sample</th>
<th>Predictor of interest</th>
<th>Condition on pretest</th>
<th>Population (df)*</th>
<th>Outcome</th>
<th>source</th>
<th>r</th>
<th>% bias to make inference invalid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Effects of Summer Instruction and Retention in New York City (Mariano)</td>
<td></td>
<td>Summer instruction and retention</td>
<td>RD</td>
<td>5th graders in New York (29,987)</td>
<td>English &amp; Language Arts achievement 6th grade</td>
<td></td>
<td>.029</td>
<td>60%</td>
</tr>
<tr>
<td>Unintended Consequences of an Algebra-for-All Policy on High-Skill Students (Nomi)</td>
<td></td>
<td>schools that increased algebra enrollment post-policy for low-ability students by more than 5%</td>
<td>Interrupted Time Series</td>
<td>9th graders in Chicago (17,987)</td>
<td>Peer ability levels for high skill students (1999 difference for affected schools)</td>
<td></td>
<td>.030</td>
<td>50%</td>
</tr>
<tr>
<td>Impact of Dual Enrollment on College Degree Attainment (An)</td>
<td>yes</td>
<td>Dual enrollment</td>
<td>NELS (8,754)</td>
<td>Any degree obtained</td>
<td></td>
<td></td>
<td>.043</td>
<td>50%</td>
</tr>
<tr>
<td>Life After Vouchers (Carlison)</td>
<td></td>
<td>Student transferred to public school</td>
<td>No</td>
<td>3rd, 10th grade in Milwaukee public</td>
<td>Math</td>
<td></td>
<td>.092</td>
<td>36%</td>
</tr>
<tr>
<td>Teaching Students What They Already Know? (Engel)</td>
<td>yes</td>
<td>Teaching practices</td>
<td>yes</td>
<td>Kindergarten teachers in ECLS-K (2,174)</td>
<td>Math (spring of kindergarten)</td>
<td></td>
<td>.059</td>
<td>28%</td>
</tr>
<tr>
<td>Unintended Consequences of an Algebra-for-All Policy on High-Skill Students (Nomi)</td>
<td></td>
<td>schools that increased algebra enrollment post-policy for low-ability students by more than 5%</td>
<td>Interrupted Time Series</td>
<td>9th graders in Chicago (17,987)</td>
<td>9th grade math scores (1999 difference for affected schools)</td>
<td></td>
<td>.020</td>
<td>26%</td>
</tr>
<tr>
<td>Can High Schools Reduce College Enrollment Gaps With a New Counseling Model?</td>
<td></td>
<td>Coach in school for college going</td>
<td>9th, 12th grade schools in Chicago (54)</td>
<td>Applied to 3+ colleges</td>
<td></td>
<td></td>
<td>.35</td>
<td>23%</td>
</tr>
<tr>
<td>Can Research Design Explain Variation in Head Start Research Results? (Shager)</td>
<td></td>
<td>Elements of study design</td>
<td>Studies of Head Start on Cognition (20)</td>
<td>Reported effect size</td>
<td></td>
<td></td>
<td>.524</td>
<td>16%</td>
</tr>
<tr>
<td>Can High Schools Reduce College Enrollment Gaps With a New Counseling Model?</td>
<td></td>
<td>Coach in school for college going</td>
<td>9th, 12th grade schools</td>
<td>Enrollment in less selective 4-yr college</td>
<td></td>
<td></td>
<td>.27</td>
<td>2%</td>
</tr>
</tbody>
</table>

| Model? (Stephan)                                                                 |               |                                                                                       |                      |                  |                                                                         |       |       |                               |
| Balancing Career and Technical Education With Academic Coursework (Bozick)        |               |                                                                                       |                      |                  |                                                                         |       |       |                               |

* Degrees of freedom for calculating threshold value of r are defined at the level of the predictor of interest, supplementing the df.
% Bias to Invalidate Inference for Observational Studies
On-line EEPA July 24-Nov 15 2012

Kindergarten retention effect
Quantify the Robustness of Causal Inferences

KonFound-It! takes four values - the estimated effect (such as an unstandardized regression coefficient or the group mean difference), its standard error, the number of observations, and the number of publishable statements as well as figures to support the interpretation of the output.

Change or set any of the values below and then click run to see output from KonFound-It!

<table>
<thead>
<tr>
<th>Estimated Effect</th>
<th>Standard Error</th>
<th>Number of Observations</th>
<th>Number of Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.4</td>
<td>100</td>
<td>3</td>
</tr>
</tbody>
</table>

R

More information on the R package can be found here

For R (presently in-development), issue the following commands:

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

Then you use the following functions for already-published studies, models (including mixed effects models) fit in R, and meta-analyses, respectively:

```
pkonfound()
konfound()
mkonfound()
```

STATA

More information on the STATA module can be found here

For STATA, issue the following commands:
In R: Pkonfound
(published example)

install.packages("konfound")
library(konfound)
pkonfound(est_eff = -9.01, std_err = .68, n_obs = 7639, n_covariates = 223)

> pkonfound(est_eff = -9.01,
+     std_err = .68,
+     n_obs = 7639,
+     n_covariates = 223)
Percent Bias Necessary to Invalidate the Inference:
To invalidate an inference, 85.205% of the estimate would have to be due to bias. This is based on a threshold of -1.333 for statistical significance (alpha = 0.05).
To invalidate an inference, 6509 observations would have to be replaced with cases for which the effect is 0.

Impact Threshold for a Confounding Variable:
An omitted variable would have to be correlated at 0.364 with the outcome and at 0.364 with the predictor of interest (conditioning on observed covariates) to invalidate an inference based on a threshold of -0.021 for statistical significance (alpha = 0.05). Correspondingly, the impact of an omitted variable (as defined in Frank (2012)) must be 0.364 X 0.364 = 0.132 to invalidate an inference.
For other forms of output, change `to_return` to table, raw_output, thres_plot, or corr_plot.
For models fit in R, consider use of konfound().
Sensitivity on Regression Run in R: Konfound (data in R)

data <- read.table(url("https://msu.edu/~kenfrank/p025b.txt"), header = T)
model <- lm(Y1 ~ X1 + X4, data = data)
model
konfound(model, X1)

--

Call: 
  lm(formula = Y1 ~ X1 + X4, data = data)

Coefficients: 
  (Intercept)       X1        X4
  4.33291    0.45073   -0.09873

konfound(model, X1)

Percent Bias Necessary to Invalidate the Inference:
To invalidate an inference, 29.778% of the estimate would have to be due to bias. This is based on a threshold of 0.317 for statistical significance (alpha = 0.05).
To invalidate an inference, 3 observations would have to be replaced with cases for which the effect is 0.

Impact Threshold for a Confounding Variable:
An omitted variable would have to be correlated at 0.592 with the outcome and at 0.592 with the predictor of interest (conditioning on observed covariates) to invalidate an inference based on a threshold of 0.6 for statistical significance (alpha = 0.05).
Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be 0.592 x 0.592 = 0.35 to invalidate an inference.

For more detailed output, consider setting `to_return` to table
To consider other predictors of interest, consider setting `test_all` to TRUE.
In STATA

```
.ssc install konfound
.ssc install moss
.ssc install matsort
.ssc install indeplist
.pkonfound -9.01 .68 7639 223
/* pkonfound estimate standard_error n number_of_covariates */
```
Step 1: search konfound in the Viewer window

Step 2: click on the link, you get this window

Step 3: click “click here to install” to install it!
Sensitivity on Regression Run in STATA

```
use https://msu.edu/~kenfrank/p025b.dta, clear
regress y1 x1 x4
konfound x1
```

```
use http://www.ats.ucla.edu/stat/stata/examples/chp/p025b, clear

 regress y1 x1 x4

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>28.3141343</td>
<td>2</td>
<td>14.1570671</td>
<td>F( 2, 8) = 8.74</td>
</tr>
<tr>
<td>Residual</td>
<td>12.9585573</td>
<td>9</td>
<td>1.61981966</td>
<td>Prob &gt; F = 0.0097</td>
</tr>
<tr>
<td>Total</td>
<td>41.2726916</td>
<td>10</td>
<td>4.12726916</td>
<td>R-squared = 0.6860</td>
</tr>
</tbody>
</table>

| Coef.    | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|----------|-----------|-------|------|---------------------|
| x1       | 0.4507273 | 0.140122 | 3.22 | 0.012 | 0.1276054 0.7738492 |
| x4       | -0.0987272 | 0.140122 | -0.70 | 0.501 | -0.4218492 0.2243947 |
| _cons    | 4.332909  | 2.217738 | 1.95 | 0.086 | -0.7812041 9.447022 |

konfound x1

The Threshold for % Bias to Invalidate/Sustain the Inference

For x1:
To invalidate the inference 29.67% of the estimate would have to be due to bias; to invalidate the inference 29.67% (3) cases would have to be replaced with cases for which there is an effect of 0.
Exercise 1: % Bias necessary to Invalidate an Inference for Internal Validity

Take an example from an observational study (your own data, the toy example, or a published example)
Calculate the % bias necessary to invalidate the inference (using konfound or pkonfound)
[ignore output for correlation based approach with impact]

Interpret the % bias in terms of sample replacement
What are the possible sources of bias?
Would they all work in the same direction?
What happens if you change the
sample size
# of covariates
standard error
Debate your inference with a partner
Approximating the Unsampling Population with Observed Data from an RCT (External Validity)

<table>
<thead>
<tr>
<th>Unit</th>
<th>s</th>
<th>Y</th>
<th>combined</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>t</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>t</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>t</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>c</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>c</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>c</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>t</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>t</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>t</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>c</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>c</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>c</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>c</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>c</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>c</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Texas Notre Dame

How many cases would you have to replace with cases with zero effect to change the inference? Assume threshold is: $\delta^* = 4$:

$$1 - \frac{\delta^*}{\hat{\delta}} = 1 - \frac{4}{6} = .33 = (1/3)$$
Figure 1
Estimated Treatment Effects in Hypothetical Studies A and B Relative to a Threshold for Inference

% bias(\(\hat{\delta}\)) to invalidate = \(\frac{\hat{\delta} - \delta^#}{\hat{\delta}}\) = 1 - \(\frac{\delta^#}{\hat{\delta}}\) = 1 - \(\frac{4}{6}\) = \(\frac{1}{3}\) = 33%

To invalidate the inference, replace 33% of cases with cases from unsampled population with zero effect.
Application to Randomized Experiment: Effect of
Open Court Curriculum on Reading Achievement

Open Court “scripted” curriculum versus business as usual
917 elementary students in 49 classrooms
Comparisons within grade and school
Outcome Measure: Terra Nova comprehensive reading score

<table>
<thead>
<tr>
<th>Randomized Block</th>
<th>Open Court (n = 27)</th>
<th>Control (n = 22)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>State</td>
<td>Grade Level</td>
</tr>
<tr>
<td>1</td>
<td>ID</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>ID</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>ID</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>ID</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>FL</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>FL</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>FL</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>NC</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>NC</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>TX</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>TX</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>TX</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>TX</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>TX</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>IN</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>IN</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>IN</td>
<td>1</td>
</tr>
<tr>
<td>Means</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Condition</td>
<td>N</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>--------------</td>
<td>-----</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>Open Court</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>22</td>
</tr>
<tr>
<td>Reading Vocabulary</td>
<td>Open Court</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>22</td>
</tr>
<tr>
<td>Reading Composite</td>
<td>Open Court</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>22</td>
</tr>
<tr>
<td>% Minority</td>
<td>Open Court</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>22</td>
</tr>
<tr>
<td>% English as a Second Language</td>
<td>Open Court</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>22</td>
</tr>
<tr>
<td>% Special Education</td>
<td>Open Court</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>22</td>
</tr>
<tr>
<td>% Free Lunch</td>
<td>Open Court</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>22</td>
</tr>
</tbody>
</table>
Value of Randomization

Few differences between groups
But done at classroom level
  Teachers might talk to each other
School level is expensive (Slavin, 2008)

### TABLE 3
**Posttest Outcomes for the Open Court and Control Samples by Grade Level**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Posttest</th>
<th>Open Court</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Classes</td>
<td>Students</td>
</tr>
<tr>
<td>1</td>
<td>Reading Composite</td>
<td>9</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>9</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>Reading Comprehension</td>
<td>9</td>
<td>165</td>
</tr>
<tr>
<td>2</td>
<td>Reading Composite</td>
<td>6</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>6</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>Reading Comprehension</td>
<td>6</td>
<td>121</td>
</tr>
<tr>
<td>3</td>
<td>Reading Composite</td>
<td>5</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>5</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Reading Comprehension</td>
<td>5</td>
<td>93</td>
</tr>
<tr>
<td>4</td>
<td>Reading Composite</td>
<td>4</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>4</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Reading Comprehension</td>
<td>4</td>
<td>81</td>
</tr>
<tr>
<td>5</td>
<td>Reading Composite</td>
<td>3</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>3</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Reading Comprehension</td>
<td>3</td>
<td>51</td>
</tr>
<tr>
<td>All Grades</td>
<td>Reading Composite</td>
<td>27</td>
<td>507</td>
</tr>
<tr>
<td></td>
<td>Vocabulary</td>
<td>27</td>
<td>507</td>
</tr>
<tr>
<td></td>
<td>Reading Comprehension</td>
<td>27</td>
<td>511</td>
</tr>
</tbody>
</table>

\[n = 27 + 22 = 49\]
Differences between Open Court and Business as Usual

Difference across grades: about 10 units

7.95 using statistical model

“statistically significant” unlikely (probability < 5%) to have occurred by chance alone if there were really no differences in the population

But is the Inference about Open Court valid in other contexts?
Obtaining # parameters estimated, t critical, estimated effect and standard error

3 parameters estimated, 
Df=n of classrooms-
# of parameters estimated= 49-3=46. 
t critical = $t_{.05, df=46} = 2.014$

Estimated effect ($\hat{\delta}$) = 7.95  
Standard error=1.83
Quantifying the Discourse for Borman et al: What would it take to change the inference?

$\delta =$ a population effect,
$\hat{\delta} =$ the estimated effect $= 7.95$, and
$\delta =$ the threshold for making an inference =
$se \times t_{\text{critical, df}=46} = 1.83 \times 2.014 = 3.69$

% Bias necessary to invalidate inference =
$1 - \frac{\delta}{\hat{\delta}} = 1 - \frac{3.69}{7.95} = 54\%$

54% of the estimate must be due to bias to invalidate the inference
Exercise 2: % Bias Necessary to Invalidate an Inference for External Validity

Take an example of a randomized experiment in your own data or an article or Borman’s example of Open Court

Calculate the % bias necessary to invalidate the inference

Interpret the % bias in terms of sample replacement

What are the possible sources of bias?

Would they all work in the same direction?

What happens if you change the sample size
standard error
degrees of freedom

Debate your inference with a new partner
Calculating the % Bias to Invalidate the Inference:
Inside the Calculations

\( \hat{\delta} \) = the estimated effect = 7.95, standard error =1.83, sample size=49, covariates=3

. pkonfound 7.95 1.83 49 3

------------------
The Threshold for % Bias to Invalidate/Sustain the Inference

To invalidate the inference 53.64% of the estimate would have to be due to bias; to invalidate the inference 53.64% (26) cases would have to be replaced with cases for which there is an effect of 0.
% Exceeding Threshold for Open Court Estimated Effect

\[ \hat{\delta} = 7.95 \]
\[ \delta^# = 3.68 \]

54% above threshold = 1 - 3.68/7.95 = 0.54

54% of the estimate must be due to bias to invalidate the inference.

Diagram showing the estimated effect with 54% above the threshold.
Interpretation of Amount of Bias Necessary to Invalidate the Inference: Sample Representativeness

To invalidate the inference:

54% of the estimate must be due to sampling bias to invalidate Borman et al.’s inference

You would have to replace 54% of Borman’s cases (about 30 classes) with cases in which Open Court had no effect to invalidate the inference

Are 54% of Borman et al.’s cases irrelevant for non-volunteer schools?

We have quantified the discourse about the concern of external validity
Example Replacement of Cases from Non-Volunteer Schools to Invalidate Inference of an Effect of the Open Court Curriculum

Business as Usual

Unsampled Population: No effect

Open Court

Original volunteer cases that were not replaced
Replacement cases from non-volunteer schools with no treatment effect
Original distribution for all volunteer cases
Pragmatism and the Fundamental Problem of External Validity

Pre-experiment population

\[ X \rightarrow Y \]

\[ \neq \]

Post-experiment population

Fundamental problem of external validity:
The more influential a study the more different the pre and post populations, the less the results apply to the post experimental population (Ben-David; Kuhn)

All the more so if it is due to the design (Burtless, 1995)
### % Bias to Invalidate Inferences across Randomized Experiments

(Quantifying the Robustness of Inferences from Randomized Experiments)

<table>
<thead>
<tr>
<th>Study (author, year)</th>
<th>Treatment vs Control</th>
<th>Blocking</th>
<th>Population</th>
<th>Outcome</th>
<th>Estimated effect, standard error, source</th>
<th>Effect size (correlation)</th>
<th>% bias to make the inference invalid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A multi-site cluster randomized field trial of Open Court Reading (Borman et al., 2008)</td>
<td>Open Court curriculum versus business as usual</td>
<td>Within grade and school</td>
<td>917 students in 49 classrooms</td>
<td>Terra Nova comprehensive reading score</td>
<td>7.95 (1.83) Table 4, results for reading composite score</td>
<td>.16 (.54)</td>
<td>47%</td>
</tr>
<tr>
<td>Answers and questions about class size. A statewide experiment (Finn and Achilles, 1990)</td>
<td>Small classes versus all others</td>
<td>By school</td>
<td>6500 students in 328 classrooms</td>
<td>Stanford Achievement Test, reading</td>
<td>13.14 (2.34) Table 5 mean for other classes is based on the regular and aide classes combined proportional to their sample sizes</td>
<td>.23 (.30)</td>
<td>64%</td>
</tr>
<tr>
<td>Experimental Evaluation of the Effects of a Research-based Preschool Mathematics Curriculum (Clements and Sarama, 2008)]</td>
<td>Building Blocks research to practice curriculum versus alternate math intensive curriculum</td>
<td>By program type</td>
<td>276 children within 35 classrooms randomly sampled from volunteers within income strata</td>
<td>Change in Early Mathematics Childhood Assessment IRT scale score (mean=50, std=10)</td>
<td>3.55 (1.16) Building Blocks vs comparison group Table 6 (df of 19 used based on footnote b)</td>
<td>.5 (.60)</td>
<td>31%</td>
</tr>
</tbody>
</table>
Table B1
Quantifying the Robustness of Inferences from Randomized Experiments in EEPA posted on-line July 24-Nov 15 2012

<table>
<thead>
<tr>
<th>Study (1st author)</th>
<th>Predictor of interest</th>
<th>Population (df)</th>
<th>Outcome source</th>
<th>r</th>
<th>% bias to make inference invalid</th>
<th>Multiplier to make estimate significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices (Yuan)</td>
<td>Pay-for performance</td>
<td>5th-8th grade teachers in Nashville (143)</td>
<td>Test Preparation</td>
<td>.250</td>
<td>33%</td>
<td>1.0</td>
</tr>
<tr>
<td>Effects of Two Scientific Inquiry Professional Development Interventions on Teaching Practice (Grigg)</td>
<td>Professional development for inquiry in science</td>
<td>4th-5th grade Classrooms in Los Angeles (70 schools)</td>
<td>Inquiry used</td>
<td>.277</td>
<td>15%</td>
<td>1.0</td>
</tr>
<tr>
<td>A Randomized Controlled Trial Evaluation of <em>Time to Read</em>, a Volunteer Tutoring Program for 8- to 9-Year-Olds (Miller)</td>
<td>Time to Read (tutoring)</td>
<td>8-9 year old children in Northern Ireland (734)</td>
<td>Future aspiration</td>
<td>.078</td>
<td>7%</td>
<td>1.0</td>
</tr>
<tr>
<td>Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices (Yuan)</td>
<td>Pay-for performance</td>
<td>5th-8th grade teachers in Nashville (143)</td>
<td>Extra hours worked/week</td>
<td>.077</td>
<td>-154%</td>
<td>2.53</td>
</tr>
</tbody>
</table>

* Degrees of freedom for calculating threshold value of r are defined at the level of the predictor of interest, subtracting the number of parameters estimated at that level. Calculations can be conducted from spreadsheet at [url to be provided, would reveal author identity]. Studies listed in order of % robustness.
Distribution of % Bias to Invalidate Inference for Randomized Studies EEPA: On-line Jul 24-Nov 5 2012

Mean = .43
Std. Dev. = .21
N = 6

Open Court
Review & Reflection

Review of applications

Concern about internal validity: Kindergarten retention (Hong and Raudenbush)
- 85% of cases must be replaced counterfactual data (with no effect) to invalidate the inference of a negative effect of retention on reading achievement
  - Comparison with other observational studies

Concern about external validity: Open Court Curriculum
- 54% of cases must be replaced with data from unobserved population to invalidate the inference of a positive effect of Open Court on reading achievement in non-volunteer schools
  - Comparison with other randomized experiments

Reflect
Which part is most confusing to you?
Is there more than one interpretation?
Discuss with a partner or two
Case Replacement for Logistic Regression

• Replace cases with null hypothesis cases.
• What is a null hypothesis case? One in which probability of success is independent of predictor
• → You switch some treatment success cases to treatment failure case
State of the Art: Rosenbaum Bounds

Odds ratio relating treatment to outcome
Odds ratio relating omitted variable to outcome
Odds ratio relating omitted variable to treatment

<table>
<thead>
<tr>
<th>$\Gamma$</th>
<th>$\Delta = 1.9$</th>
<th>$\Delta = 2.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>(1.17–4.44)</td>
<td>(1.17–4.44)</td>
</tr>
<tr>
<td></td>
<td>2.28</td>
<td>2.28</td>
</tr>
<tr>
<td></td>
<td>(1.17–4.45)</td>
<td>(1.17–4.45)</td>
</tr>
<tr>
<td></td>
<td>2.19</td>
<td>2.16</td>
</tr>
<tr>
<td></td>
<td>(1.13–4.32)</td>
<td>(1.11–4.24)</td>
</tr>
<tr>
<td></td>
<td>2.11</td>
<td>2.10</td>
</tr>
<tr>
<td>1.0</td>
<td>(1.16–4.46)</td>
<td>(1.16–4.46)</td>
</tr>
<tr>
<td></td>
<td>2.28</td>
<td>2.28</td>
</tr>
<tr>
<td></td>
<td>(1.11–4.29)</td>
<td>(1.05–4.16)</td>
</tr>
<tr>
<td></td>
<td>2.06</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>(1.05–4.10)</td>
<td>(1.02–3.97)</td>
</tr>
<tr>
<td>2.2</td>
<td>(1.16–4.49)</td>
<td>(1.10–4.27)</td>
</tr>
<tr>
<td></td>
<td>2.01</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>(1.05–4.10)</td>
<td>(0.99–3.87)</td>
</tr>
</tbody>
</table>

NOTE.—$\Gamma =$ effect of hypothetical unobserved covariate on treatment (odds ratio). $\Delta =$ effect of hypothetical unobserved covariate on outcome (odds ratio).

* The standard bias of the propensity score is .0002.


Increments of $\Gamma$ matter more as $\Delta$ increases. Correlation between $\Gamma$ and odd ratio is -.96
Replacement of Cases for Logistic: Toy Example Oberved Data

<table>
<thead>
<tr>
<th></th>
<th>Failure</th>
<th>Success</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>15</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>Treatment</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>30</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Failure</th>
<th>Success</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
</tr>
<tr>
<td>Treatment</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
</tr>
<tr>
<td>Total</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
<td>☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐</td>
</tr>
</tbody>
</table>

Odds ratio=(15x20)/(10x5)=6; Ln(Odds ratio)=ln(6)=1.79
Standard error=(1/15+1/10+1/5+1/20)\(^{1/2}\)=.64
Threshold=1.96x.64=1.27
% bias to invalidate=1-1.27/1.79=29%
Replace 29% of treatment successes (n=20)=6 cases
Replace 29% of treatment successes (n=20)=6 cases
Replace with null hypothesis cases (p of success=30/50=.6).
So some of the treatment successes are replaced with other treatment successes
→ Ask instead, how many treatment successes must be switched to treatment failure to invalidate the inference

### Replacement Cases, P(success)=.6

<table>
<thead>
<tr>
<th></th>
<th>Failure</th>
<th>Success</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control</strong></td>
<td>◊◊◊◊◊◊◊◊◊</td>
<td>◊◊◊◊◊◊◊◊</td>
<td>◊◊◊◊◊◊◊◊◊</td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td>◊◊◊◊◊◊◊◊◊</td>
<td>◊◊◊◊◊◊◊◊</td>
<td>◊◊◊◊◊◊◊◊◊</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>◊◊◊◊◊◊◊◊◊</td>
<td>◊◊◊◊◊◊◊◊</td>
<td>◊◊◊◊◊◊◊◊◊</td>
</tr>
</tbody>
</table>
### Switching Treatment success to Treatment Failure

#### Before Switches, from data

<table>
<thead>
<tr>
<th></th>
<th>Failure</th>
<th>Success</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>15</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>Treatment</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>30</td>
<td>50</td>
</tr>
</tbody>
</table>

Odds ratio = \((15 \times 10)/(10 \times 8)\) = 6; \(\text{Ln(Odds ratio)} = \ln(6) = 1.79\)

Standard error = \(\sqrt{1/15 + 1/10 + 1/5 + 1/20} = .64\)

T ratio = 2.27

#### After Switching 3 treatment success to treatment failure

<table>
<thead>
<tr>
<th></th>
<th>Failure</th>
<th>Success</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>15</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>Treatment</td>
<td>8</td>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>30</td>
<td>50</td>
</tr>
</tbody>
</table>

Odds ratio = \((15 \times 17)/(10 \times 8)\) = 3.19; \(\text{Ln(Odds ratio)} = 1.15\)

Standard error = \(\sqrt{1/15 + 1/10 + 1/8 + 1/17} = .59\)

T ratio = 1.958
setwd("C:/Users/userDropbox (Personal)/sensitivity for logistics")
rm(list = ls())
A <- 29
B <- 26
C <- 15
D <- 40
x <- matrix(c(A,B,C,D), byrow = TRUE, 2, 2) # this is the 2 by 2 table we start with
p.CD <- p.value <- chisq.test(x,correct = FALSE)$p.value
N.CD <- 0
while ( p.value <0.05 ) {
  C <- C + 1
  D <- D - 1
  N.CD <- N.CD + 1
  x <- matrix(c(A,B,C,D), byrow = TRUE, 2, 2)
  print(x)
  print(chisq.test(x,correct = FALSE))
  p.value <- chisq.test(x,correct = FALSE)$p.value
  p.CD <- c(p.CD, chisq.test(x,correct = FALSE)$p.value)
}
A <- 29
B <- 26
C <- 15
D <- 40
x <- matrix(c(A,B,C,D), byrow = TRUE, 2, 2)
p.AB <- p.value <- chisq.test(x,correct = FALSE)$p.value
N.AB <- 0
while ( p.value <0.05 ) {
  A <- A - 1
  B <- B + 1
  N.AB <- N.AB + 1
  x <- matrix(c(A,B,C,D), byrow = TRUE, 2, 2)
  print(x)
  print(chisq.test(x,correct = FALSE))
  p.value <- chisq.test(x,correct = FALSE)$p.value
  p.AB <- c(p.AB, chisq.test(x,correct = FALSE)$p.value)
}

# so p.AB and p.CD record the p values each time we switch one case
# N.AB = 5 (5 cases needed to be changed from A to B)
# N.CD = 4 (4 cases needed to be changed from D to C)
KonFound-it for Logistic Regression

https://jmichaelrosenberg.shinyapps.io/shinykonfound/
In R: Pkonfound for nonlinear (published example)

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

pkonfound(2.5, 0.5, 20888, 3, n_trm = 17888, switch_trm=TRUE, non_linear = TRUE)
```

> pkonfound(2.5, 0.5, 20888, 3, n_trm = 17888, switch_trm=TRUE, non_linear = TRUE)

```
$Implied_Table
  Fail Success
Control 12   29888
Treatment 6    17882

$Transfer_Table
  Fail Success
Control 12   29888
Treatment 38   17850
```

```
[[4]]
[1] "To invalidate the inference, 32 cases need to be transferred from treatment success to treatment failure, as shown from the Implied Table to the Transfer Table."

[[5]]
[1] "For the Implied Table, we have estimate of 2.482, with standard error of 0.500 and t-ratio of 4.961."

[[5]]
[1] "For the Transfer Table, we have estimate of 0.635, with standard error of 0.332 and t-ratio of 1.913."
```
Extensions

Powerpoint for ordered thresholds relative to transaction costs
Specify threshold based on context of value and cost of treatment

Powerpoint for alternative scenarios and related techniques
Includes Type II errors, alternative replacement data

Powerpoint for new directions
Includes value added, Bayesian, integrated framework

Powerpoint for comparison of frameworks [pdf of comparison of frameworks]
Shows algebraic and graphical relationships between comparison of cases and correlation/regression frameworks
Quantifying the Robustness of Inferences: Correlation/regression Framework

Work in terms of partial correlations
How strongly must an omitted variable be correlated with the predictor of interest (e.g., retention) and the outcome (e.g., reading achievement) to invalidate an inference of an effect of retention on achievement?
Estimated Effect of Retention on Reading Scores (Hong and Raudenbush)
In R Shiny app KonFound-it!
(konfound-it.com/)

Quantify the Robustness of Causal Inferences

KonFound-It! takes four values - the estimated effect (such as an unstandardized regression coefficient or the group mean difference), its standard error, the number of observations, and the number of covariates. KonFound-It! returns output in the forms of publishable statements as well as figures to support the interpretation of the output.

Estimated effect

Correlation-based Approach: An omitted variable would have to be correlated at 0.364 with the outcome and at 0.364 with the predictor of interest (conditioning on observed covariates) to invalidate an inference based on a threshold of -0.023 for statistical significance (alpha = 0.05). Correspondingly the impact of an omitted variable (as defined by Frank 2000) must be 0.364 * 0.364 = 0.132 to invalidate an inference.

Number of observations: 7639

Number of covariates: 223

Take out your phone and try it!!!
The Impact of a Confounding Variable: Accounting for the Dual Relationships Associated with an Alternative Factor

Each correlation is weighted in proportion to the size of the other.
How Regression Works: Partial Correlation

Partial Correlation: correlation between x and y, where x and y have been controlled for the confounding variable

\[
r_{xy|cv} = \frac{r_{x\cdot y} - r_{x\cdot cv} \times r_{y\cdot cv}}{\sqrt{1-r_{x\cdot cv}^2} \sqrt{1-r_{y\cdot cv}^2}} = \frac{r_{\text{ret\cdot ach}} - r_{\text{ret\cdot momed}} \times r_{\text{ach\cdot momed}}}{\sqrt{1-r_{\text{ach\cdot momed}}^2} \sqrt{1-r_{\text{ret\cdot momed}}^2}}
\]

Implied by multivariate: \( B = [X'X]^{-1}X'Y \)

\[
-0.152 - 0.075 \times 0.203 \\
\sqrt{1-0.075^2} \sqrt{1-0.203^2} = -0.140
\]

<table>
<thead>
<tr>
<th></th>
<th>momed</th>
<th>retention</th>
<th>achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td>momed</td>
<td>1</td>
<td>-0.075</td>
<td>-0.152</td>
</tr>
<tr>
<td>retention</td>
<td>-0.075</td>
<td>1</td>
<td>0.203</td>
</tr>
<tr>
<td>achievement</td>
<td>0.203</td>
<td>-0.152</td>
<td>1</td>
</tr>
</tbody>
</table>
The Impact of a Confounding Variable: Accounting for the dual relationships associated with an alternative factor, linear relationships

Each correlation is weighted in proportion to the size of the other.
Use the Impact of a Confounding Variable on a Regression Coefficient to Quantify Robustness of Inference

• Statements such as:
  – An omitted variable must have an impact of \( q \) to invalidate the inference of an effect of \( X \) on \( Y \).
  – An omitted variable must be correlated at \( r \) with \( X \) and \( Y \) to invalidate an inference of an effect of \( X \) on \( Y \).

• I call this the Impact Threshold of a Confounding Variable (ITCV).
Quantifying the Robustness of the Inference: Calculating the Impact Threshold of a Confounding Variable on a Regression Coefficient

Step 1: Establish Correlation Between predictor of interest and outcome
Step 2: Define a Threshold for Inference
Step 3: Calculate the Threshold for the Impact Necessary to Invalidate the Inference
Obtain t critical, estimated effect and standard error

\[ y = \beta_0 + \beta_1 x + \beta_2 \text{unobserved confound} \]

\[ \text{reading achievement} = \beta_0 + \beta_1 \text{retention} + \beta_2 \text{unobserved confound} \]

With no statistical adjustment for selection bias, the mean differences between the 471 kindergarten retainees and the 7,168 promoted at-risk students in retention schools were −18.51 in the reading outcome and −11.06 in the math outcome. Given the pretreatment imbalance between the two groups, these mean differences were likely to be negatively biased if viewed as estimates of the causal effects of being retained.

Next, we computed the mean differences in the reading and math outcomes between the retained students and the at-risk promoted students in the retention schools within each individual-level propensity stratum. The results are displayed in Tables 9 and 10. After a year of treatment, the kindergarten retainees generally reached a lower achievement status in both reading and mathematics than did their comparable promoted peers. The within-stratum mean differences in reading ranged from −0.96 to −12.14, while those in mathematics ranged from 1.57 to −7.98.

To generate a more conclusive answer to the causal question, we specified a two-level hierarchical linear model for estimating the average retention effect, \( \delta \), on the learning outcomes of at-risk children attending retention schools.

\[ \hat{\delta} = -9.01 \]

\[ \text{Standard error} = .68 \]

Establish Correlation and Threshold for Kindergarten Retention on Achievement

**Step 1: calculate treatment effect as correlation**

\[
r = \frac{t}{\sqrt{(n-q-1) + t^2}} = \frac{-13.25}{\sqrt{(7639 - 2 - 1) + (13.25)^2}} = -.150
\]

- \( t \) taken from HLM: \(-9.01/0.68\) = 13.25
- \( n \) is the sample size = 7639
- \( q \) is the number of parameters estimated (predictor + omitted variable = 2)

**Step 2: calculate threshold for inference (e.g. statistical significance)**

\[
\text{threshold} = r# = \frac{t_{\text{critical}}}{\sqrt{(n-q-1) + t^2_{\text{critical}}}} = \frac{-1.96}{\sqrt{(7639 - 2 - 1) + 1.96^2}} = -.022
\]

Where \( t \) is critical value for \( df > 200 \)

\( r# \) can also be defined in terms of effect sizes
Step 3a: Calculate the Impact Necessary to Invalidate the Inference

Assume $r_{x\cdot cv} = r_{y\cdot cv}$ (maximizes the difference between $r_{xy}$ and $r_{xy|cv}$). Then $r_{x\cdot cv} \times r_{y\cdot cv} = r_{x\cdot cv} \times r_{x\cdot cv} = r_{y\cdot cv} \times r_{y\cdot cv} = \text{impact}$ and

$$r_{x\cdot y|cv} = \frac{r_{x\cdot y} - r_{x\cdot cv} \times r_{y\cdot cv}}{\sqrt{1 - r^2_{x\cdot cv}} \sqrt{1 - r^2_{y\cdot cv}}} = \frac{r_{x\cdot y} - \text{impact}}{1 - \text{impact}}$$

Set $r_{x\cdot y|cv} = r^#$ and solve for impact to find the impact threshold of a confounding variable ($ITCV$).

$$ITCV = \frac{r_{x\cdot y} - r^#}{1-|r^#|} \quad ITCV = \frac{-.150 - .022}{1-.|-.022|} = -.130$$

ITCV < 0 because $r_{xy} < 0$

The inference would be invalid if $|\text{impact of a confounding variable}| > .130$. 
Step 3b: Component correlations and Interpretation

ITCV = .130

If \( r_{x\cdot cv} = r_{y\cdot cv} = r \), then impact = \( r_{x\cdot cv} \times r_{y\cdot cv} = r^2 \)

\[ r^2 = .130 \Rightarrow r = \sqrt{.130} = .361 \]

An omitted variable must be correlated at .361 with retention and with achievement (with opposite signs) to invalidate the inference of an effect of retention on achievement.

Conceptualize multiple omitted vars
- Latent variable
- Declining conditional impacts
Impact Threshold for a Confounding variable: Path Diagram

$ r_{x,y} = -0.150 $  

$ r_{x,y|CV} = -0.022 $  

$r_{CV,x} = -0.361$  

$r_{CV,x} \times r_{CV,y} = -0.361 \times 0.361 = -0.13 = \text{impact}$  

$ r_{CV,y} = 0.361 $  

$ y \ (\text{outcome}) $
Look at both Correlations

Smallest impact to invalidate inference: $r_{xcv} = r_{ycv} = .361$


Other Features of Impact Threshold for a Confounding Variable (Frank, 2000)

- Covariates already in the model
- can report ITCV as partial or zero-order correlations
- Suppression (or type II error for estimate not significant)
- Derived initially in terms of estimate and standard error
- Can apply to unreliably measured covariate
Absorption: Impacts Before and After Pretests

If \( k > .109 \) (or .130 without covariates) then the inference is invalid.

After controlling for pre-tests, impact of strongest measured covariate (student approaches to learning) is -.00126 (sign indicates controlling for it reduces estimated effect of retention).

→ Impact of unmeasured confound would have to be about 100 times greater than the impact of the strongest observed covariate to invalidate the inference. Hmmmm....

The value of controlling for pretests.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Control for School Membership</th>
<th>Control for School Membership + Pretests</th>
<th>Reduction (absorption)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Impact</td>
<td>Ix retention</td>
<td>Iy achievement</td>
</tr>
<tr>
<td>Mother's education</td>
<td>-0.01522</td>
<td>-0.0749</td>
<td>0.2032</td>
</tr>
<tr>
<td>Two parent home</td>
<td>-0.00304</td>
<td>-0.0317</td>
<td>0.0958</td>
</tr>
<tr>
<td>Poverty</td>
<td>-0.00765</td>
<td>0.0559</td>
<td>-0.1369</td>
</tr>
<tr>
<td>Student Approaches to Learning (SAL)</td>
<td>-0.08213</td>
<td>-0.1849</td>
<td>0.4442</td>
</tr>
<tr>
<td>Controls</td>
<td>Est</td>
<td>Se</td>
<td>t</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>School</td>
<td>-21.24</td>
<td>.63</td>
<td>-33.49</td>
</tr>
<tr>
<td>School+Pre2 (spring Kindergarten)</td>
<td>-12.01</td>
<td>.45</td>
<td>-26.48</td>
</tr>
<tr>
<td>School+Pre2+(Pre2-Pre1)+</td>
<td>-12.10</td>
<td>.47</td>
<td>-26.28</td>
</tr>
<tr>
<td>School+Pre2+(Pre2-Pre1)+ Momed</td>
<td>-12.00</td>
<td>.47</td>
<td>-26.26</td>
</tr>
<tr>
<td>School+Pre2+(Pre2-Pre1)+ Female</td>
<td>-12.07</td>
<td>.46</td>
<td>-25.18</td>
</tr>
<tr>
<td>School+Pre2+(Pre2-Pre1)+ 2parent</td>
<td>-12.01</td>
<td>.46</td>
<td>-26.27</td>
</tr>
<tr>
<td>School+Pre2+(Pre2-Pre1)+ poverty</td>
<td>-12.04</td>
<td>.46</td>
<td>-26.16</td>
</tr>
<tr>
<td>Hong and Raudenbush (model based)</td>
<td>-9.01</td>
<td>.68</td>
<td>-13.27</td>
</tr>
</tbody>
</table>

n=10,065, $R^2 = .40$

Note: 1 year’s growth is about 10 points, so retention effect > 1 year growth
Consider Alternate Sample (External Validity)

Define $\pi$ as the proportion of the sample that is replaced with an alternate sample.

$r_{xy}$ is correlation in unobserved data

$R_{xy}$ is combined correlation for observed and unobserved data:

$$R_{xy} = (1 - \pi)r_{xy} + \pi r_{xy}.$$  

Thresholds for Sample Replacement, $r_{xy}$

Set $R_{xy} = r^#$

$r^# = (1-\pi)r_{xy} + \pi r_{xy}$

and solve for $r_{xy}$:

$r_{xy} = [r^# + (\pi-1)r_{xy}] / \pi$

If half the sample is replaced ($\pi = .5$), original inference is invalid if $r_{xy} < 2r^# - r_{xy}$

Therefore, $2r^# - r_{xy}$ defines the threshold for replacement: $TR(\pi = .5)$

Assumes means and variances are constant across samples, alternative calculations available.
Thresholds for Sample Replacement, \( \pi \)

Set \( R = r^\# \)

\[ r^\# = (1 - \pi) r_{xy} + \pi r_{xy} \]

and solve for \( \pi \)

\[ \pi = \frac{r_{xy} - r^\#}{r_{xy} - r_{xy}} \]

if \( r_{xy} = 0 \), inference is altered if \( \pi > 1 - r^\#/r_{xy} \).

Therefore \( 1 - r^#/r_{xy} \) defines the threshold for replacement \( TR(r_{xy}) = 0 \).

Assumes means and variances are constant across samples, alternative calculations available.
Thresholds for Sample Replacement: Toy Example

If half the sample is replaced ($\pi=.5$), original inference is invalid if $r_{xy} < 2r^#-r_{xy}$

If $r^# = .4$ and $r_{xy} = .6$ then inference invalid if

$r_{xy} < .2$  \hspace{1cm} (2 \times .4-.6=.2)

If $r_{xy} = 0$, inference is altered if $\Pi > 1-r^#/r_{xy}$.  If $r^# = .4$ and $r_{xy} = .6$ then inference invalid if

$\Pi > .33$  \hspace{1cm} (1-.4/.6=.33)

Assumes means and variances are constant across samples, alternative calculations available.
Interpretation

Replacing cases with null hypothesis cases \(\Rightarrow\) % bias to invalidate

How much would you have to replace the data with flat line (null hypothesis) cases to invalidate?

How much would you have to disturb the data to invalidate the inference?
Application to Anything that is Regression Based in its Final Form

Propensity scores
   “Doubly Robust” (implies linear model at final stage)
Diff in Diff
   Done with a regression
Regression discontinuity
   Linear model, with proper functional form, in final stage
Comparative interrupted time series
   Confound be associated with pre vs post intervention or with treatment groups
Path Diagram is Fundamentally Sociological

the american journal of sociology

Volume 72 Number 1 July 1966

Path Analysis: Sociological Examples

Otis Dudley Duncan

Path Diagram
Comparison of Frameworks

Case replacement
- Good for experimental settings (treatments) or linear models
- Think in terms of cases (people or schools)
  - counterfactual or unsampled population
- Assume equal effect of replacing any case
- Or weighted cases with weighted replacements
- Good for comparing across studies

Correlational
- Uses causal diagrams (Pearl)
- Linear models only
- Think in terms of correlations, variables
- ITCV good for internal validity, not good for comparing between studies (different thresholds make comparison difficult)
Impact Threshold for a confounding variable

Qinyun Lin
Equating the expressions, evaluating for % mixture of unobserved cases ($\pi$)

\[
\frac{r_{xy} - \text{impact}}{\sqrt{1-r_{y,cv}^2}} \sqrt{1-r_{x,cv}^2} = (1-\pi)r_{xy} + \pi r_{xy} = \pi (r_{xy} - r_{xy}) + r_{xy}
\]

\[
\frac{r_{xy} - \text{impact}}{\sqrt{1-r_{y,cv}^2}} \sqrt{1-r_{x,cv}^2} - r_{xy} = \pi (r_{xy} - r_{xy})
\]

\[
\frac{r_{xy} - \text{impact}}{\sqrt{1-r_{y,cv}^2}} \sqrt{1-r_{x,cv}^2} (r_{xy} - r_{xy}) - (r_{xy} - r_{xy}) = \frac{1}{(r_{xy} - r_{xy})} \left[ \frac{r_{xy} - \text{impact}}{\sqrt{1-r_{y,cv}^2}} \sqrt{1-r_{x,cv}^2} - r_{xy} \right] = \pi
\]

\[
\Rightarrow \pi = \frac{1}{(r_{xy} - r_{xy})} \left[ r_{xy} - \frac{r_{xy} - \text{impact}}{\sqrt{1-r_{y,cv}^2}} \sqrt{1-r_{x,cv}^2} \right]
\]

if impact is maximized:

\[
\Rightarrow \pi = \frac{1}{(r_{xy} - r_{xy})} \left[ r_{xy} - \frac{r_{xy} - \text{impact}}{1-\text{impact}} \right]
\]

if $r_{xy} = 0$

\[
\pi = 1 - \frac{r_{xy} - \text{impact}}{r_{xy} (1-\text{impact})}
\]
Can you *Prove* your Car Will Make it Through?

Is there enough evidence to act?
End here
% Bias Necessary to Invalidate the Inference and Confidence Interval

Lower bound of confidence interval “far from 0” estimate exceeds threshold by large amount

Closer to 0?
Case replacement with Spillover for SUTVA Violations

**Before replacement**

**Green:**
Original students that were replaced;

**Red:**
Replacement students that can trigger spillover effects;

**Black:**
Original students who stay in the class.

**Spillover effects**

**After replacement**
Definition of Replacement Cases as Counterfactual: Potential Outcomes

Definition of treatment effect for individual $i$: $\delta_i = Y_i^t - Y_i^c$

$Y_i^t = \text{value on outcome if unit received treatment}$

$Y_i^c = \text{value on outcome if unit received control}$

Fundamental problem of causal inference is that we cannot simultaneously observe $Y_i^t$ and $Y_i^c$
Symbolic: Fundamental Problem of Inference and Approximating the Counterfactual with Observed Data (Internal Validity)

But how well does the observed data approximate the counterfactual?
Fundamental Problem of Inference and Approximating the Counterfactual with Observed Data (Internal Validity)

But how well does the observed data approximate the counterfactual?

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>potential outcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>Control</td>
<td></td>
</tr>
<tr>
<td>Unit</td>
<td>s</td>
<td>Y\textsuperscript{t}</td>
<td>Y\textsuperscript{c}</td>
<td>Effect</td>
</tr>
<tr>
<td>1</td>
<td>t</td>
<td>9</td>
<td></td>
<td>6?</td>
</tr>
<tr>
<td>2</td>
<td>t</td>
<td>10</td>
<td></td>
<td>6?</td>
</tr>
<tr>
<td>3</td>
<td>t</td>
<td>11</td>
<td></td>
<td>6?</td>
</tr>
<tr>
<td>4</td>
<td>c</td>
<td>?</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>c</td>
<td>?</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>c</td>
<td>?</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>10.00</td>
<td>4.00</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>counterfactual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But how well does the observed data approximate the counterfactual?
**Kon-Found-it: Basics**

% Bias to Invalidate

Estimated effect \( \hat{\delta} = -9.01 \)

Standard error = 0.68

\[ \text{Covariates} = 223 \]

\[ n = 7168 + 471 = 7639 \]

---

<table>
<thead>
<tr>
<th>Calculated Values</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( df )</td>
<td>( t \text{ critical} )</td>
</tr>
<tr>
<td>7415</td>
<td>-1.960</td>
</tr>
</tbody>
</table>

- Default \( t \text{ critical} \) (2-tailed) = -1.960

For a 1-tailed test, double the size of \( \alpha \) in cell E4

To override cell C8, type in your own value

The default sign of \( t \text{ critical} \) is the same as the sign of the estimated effect.

---

**Replacement of Cases**

To invalidate the inference 85% of the cases (6508) would have to be replaced with cases in which there is zero effect at -.364 with \( x \) and at .364 with the outcome, conditional on covariates.

---

Specific calculations
Calculating the % Bias to Invalidate the Inference: Inside the Calculations

\[ \delta# = \text{the threshold for making an inference } = \text{se} \times t_{\text{critical, df}>230} = 0.68 \times -1.96 = -1.33 \]

[User can specify alternative threshold]

\[% \text{Bias necessary to invalidate inference} = \frac{1-\delta#}{\hat{\delta}} = \frac{1-1.33}{9.01} = 85\%\]

85% of the estimate must be due to bias to invalidate the inference.