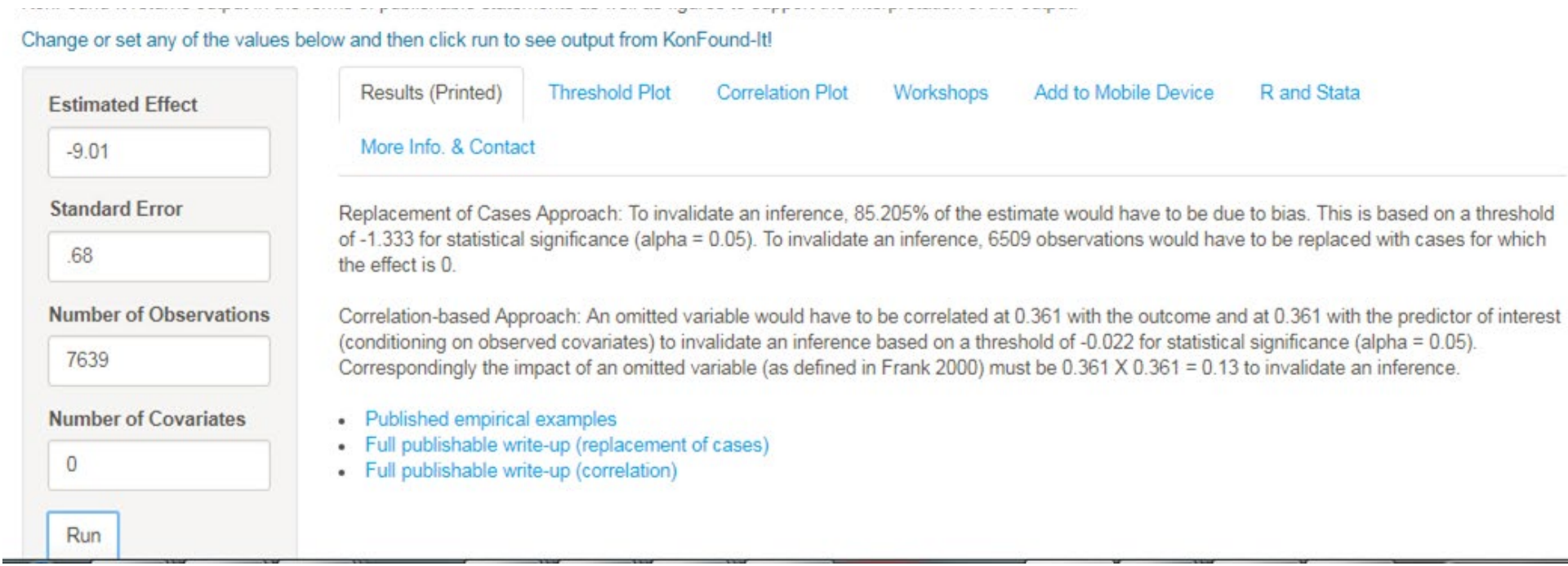
Discussing the Strength of Evidence in terms of Conditions and Context of People and Organizations: Special Applications for K-12 Education

Ken Frank (Michigan State) Qinyun Lin (University of Chicago)

November 2020

R Shiny app KonFound-it! (konfound-it.com/)



Abstract

•Pragmatic research informs the discrete choices of public policy, such as whether to implement a new policy or discontinue an old policy. This talk contributes to debates about policy by quantifying the strength of evidence relative to a threshold for making an inference. In particular, we will generate statements such as “xx% of the estimated effect must be due to bias to invalidate the inference.” Such statements can be interpreted using a potential outcomes framework: “One would have to replace xx% of the cases with null hypothesis cases to invalidate the inference.” This type of statement allows researchers and policymakers to discuss the strength of evidence in terms of the conditions and contexts of people and organizations. We will discuss specific applications to Value Added Models including potential spillover effects among students within the same classroom or school. The talk features a hands-on app: http://konfound-it.com. As well as commands in R and STATA available through the app.

Background Papers

Technical Papers:

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.

Frank, K.A., Gary Sykes, Dorothea Anagnostopoulos, Marisa Cannata, Linda Chard, Ann Krause, Raven McCrory. 2008. Extended Influence: National Board Certified Teachers as Help Providers. Education, Evaluation, and Policy Analysis. Vol 30(1): 3-30.

*Frank, K. A. and Min, K. 2007. Indices of Robustness for Sample Representation. *Sociological Methodology*. Vol 37, 349-392. * co first authors.

Pan, W., and Frank, K.A. 2004. "An Approximation to the Distribution of the Product of Two Dependent Correlation Coefficients." Journal of Statistical Computation and Simulation, 74, 419-443

Pan, W., and Frank, K.A., 2004. "A probability index of the robustness of a causal inference," Journal of Educational and Behavioral Statistics, 28, 315-337.

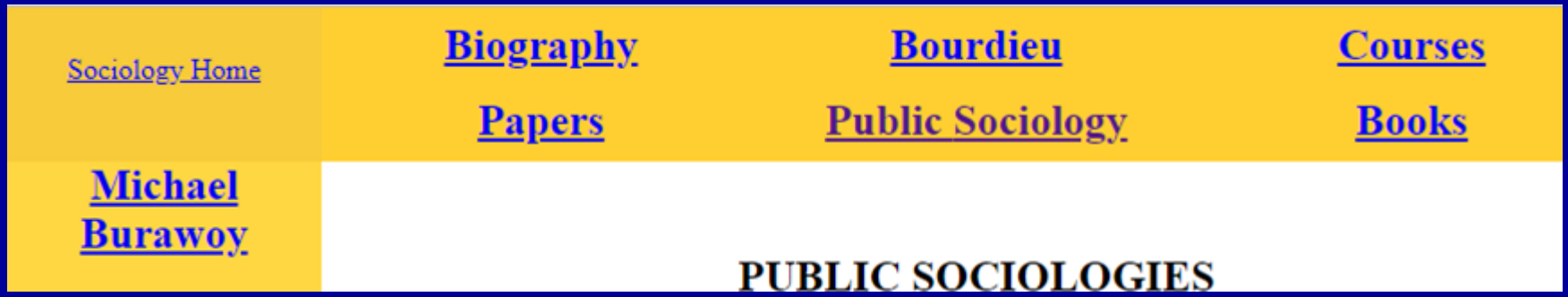
Frank, K. 2000. "Impact of a Confounding Variable on the Inference of a Regression Coefficient." Sociological Methods and Research, 29(2), 147-194

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.

Recent applications:

- 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.

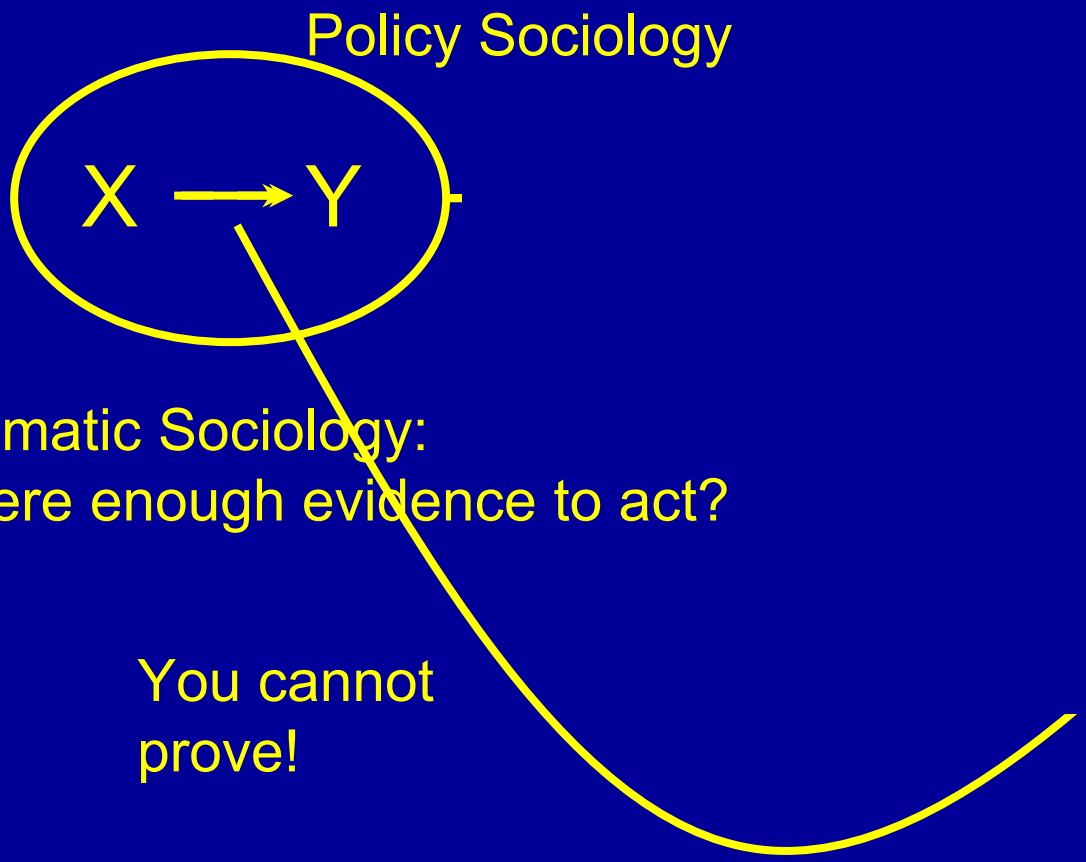
- **Frank, K.A**., Penuel, W.R. and Krause, A., 2015. What Is A "Good" Social Network For Policy Implementation? The Flow Of Know-How For Organizational Change. *Journal of Policy Analysis and*



Public Sociology endeavors to **bring sociology into dialogue with audiences beyond the academy**, an open dialogue in which both sides deepen their understanding of public issues. But what is its relation to the rest of sociology? It is the opposite of **Professional Sociology** – a scientific sociology created by and for sociologists – inspired by public sociology but, equally, without which public sociology would not exist. The relation between professional and public sociology is, thus, one of antagonistic interdependence. **Emphasis added**.

I say, robustness analysis can provide professional sociologists with a more precise and more interpretable language for pragmatic dialogue with the public.

Policy, Public, and Pragmatic Sociologies



Change 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?

Replacement of Cases Framework

How much bias must there be to invalidate an inference?

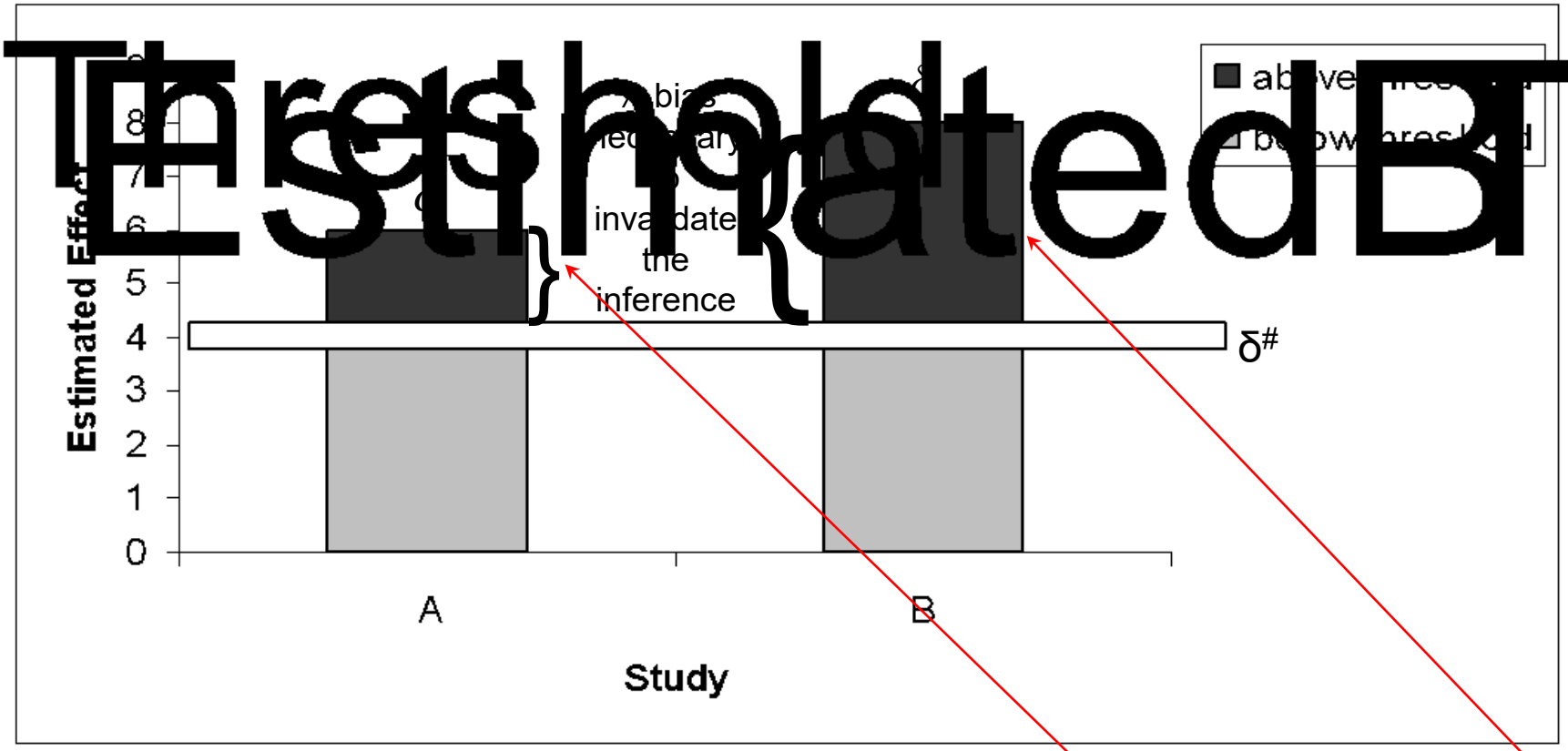
Concerns about Internal Validity

- 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?

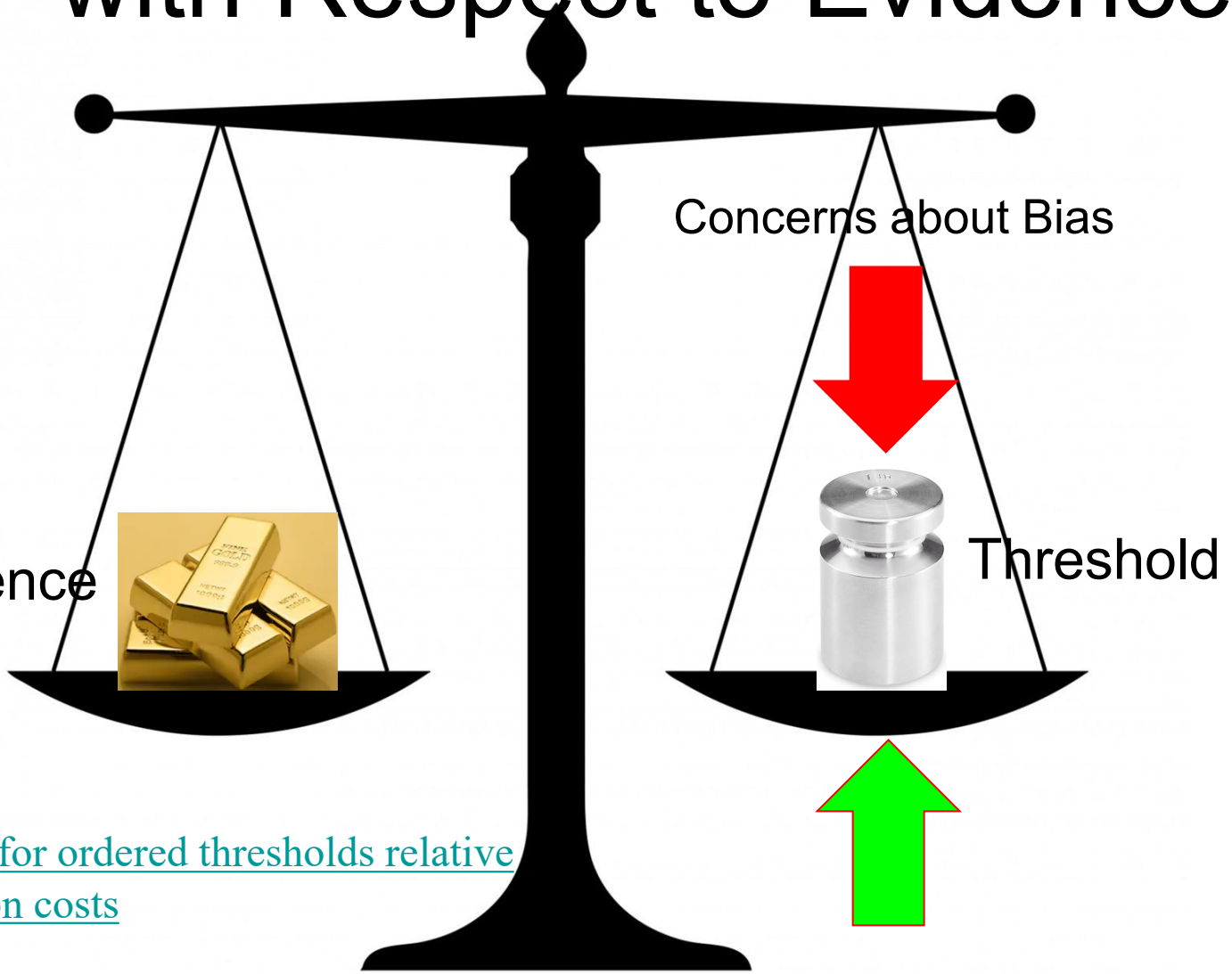
δ#



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

$$1 - \frac{4}{8} = \frac{1}{2} = 50\%$$

Threshold as Point of Indifference with Respect to Evidence



powerpoint for ordered thresholds relative to transaction costs

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

	A	B	C	D	E
1			potential outcome		
2			Treatment	Control	
3	Unit		Y^t	Y^c	Effect
4	1	t	9	8	1
5	2	t	10	9	1
6	3	t	11	10	1
7	4	c	?	3	
8	5	c	?	4	
9	6	c	?	5	
10	Mean		10.00	9	6
11					
12	counterfactual				
13	Observed				

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

Fundamental problem of causal inference is that we cannot simultaneously observe Y_i^t and Y_i^c

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945_70. *(25-40)*

Using the Counterfactual to Interpret % Bias to Invalidate the Inference: Replacement with Average Values

	A	B	C	D	E
1			potential outcome		
2			Treatment	Control	
3	Unit		Y^t	Y^c	Effect
4	1	t	9	7	0
5	2	t	10	7	0
6	3	t	11	7	0
7	4	c	7	3	######
8	5	c	7	4	######
9	6	c	7	5	######
10	Mean		9	5	4
11					
12	counterfactual				
13	Observed				

How many cases would you have to replace with zero effect counterfactuals to change the inference?
Assume threshold is 4 ($\delta^{\#}$ =4):
1- $\delta^{\#}$ / $\hat{\delta}$

=1-4/6=.33 =(1/3)

The inference would be invalid if you replaced 33% (or 2 cases) with counterfactuals for which there was no treatment effect.
New estimate=(1-% replaced)$\hat{\delta}$ +%replaced(no effect)=
(1-%replaced) $\hat{\delta}$ =(1-.33)6=.66(6)=4

Using the Counterfactual to Interpret % Bias to Invalidate the Inference: Replacement with Average Values

	A	B	C	D	E
1			potential outcome		
2			Treatment	Control	
3	Unit		Y^t	Y^c	Effect
4	1	t	9	3	0
5	2	t	10	4	0
6	3	t	11	5	0
7	4	c	?	3	######
8	5	c	?	4	######
9	6	c	?	5	######
10	Mean		9	5	4
11					
12	counterfactual				
13	Observed				

How many cases would you have to replace with zero effect counterfactuals to change the inference?
Assume threshold is 4 ($\delta^{\#}$ =4):
1- $\delta^{\#}$ / $\hat{\delta}$

=1-4/6=.33 =(1/3)

The inference would be invalid if you replaced 33% (or 2 cases) with counterfactuals for which there was no treatment effect.
New estimate=(1-% replaced)$\hat{\delta}$ +%replaced(no effect)=
(1-%replaced) $\hat{\delta}$ =(1-.33)6=.66(6)=4

Alternative to caser replacement: rendering effect for a given case null.

	A	B	C	D	E
1			potential outcome		
2			Treatment	Control	
3	Unit		Y^t	Y^c	Effect
4	1	t	9	3	0
5	2	t	4	4	0
6	3	t	11	5	0
7	4	c	?	3	######
8	5	c	?	4	######
9	6	c	?	5	######
10	Mean		8	4	4
11					
12	counterfactual				
13	Observed				

How many cases would you have to replace with zero effect counterfactuals to change the inference?
Assume threshold is 4 ($\delta^{\#}$ =4):
1- $\delta^{\#}$ / $\hat{\delta}$

=1-4/6=.33 =(1/3)

The inference would be invalid if you replaced 33% (or 2 cases) with counterfactuals for which there was no treatment effect.
New estimate=(1-% replaced)$\hat{\delta}$ +%replaced(no effect)=
(1-%replaced) $\hat{\delta}$ =(1-.33)6=.66(6)=4

Which Cases to Replace?

- Expectation: if you randomly replaced 1/3 of the cases, and repeated 1,000 times, on average the new estimate would be 4
- Assumes constant treatment effect
- Conditioning on covariates and interactions in model
- Assumes cases carry equal weight
- Extensions include selective replacement, spillover, weighted observations, logistic, “causal” designs (e.g., RD)

$\delta^{\#}$

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

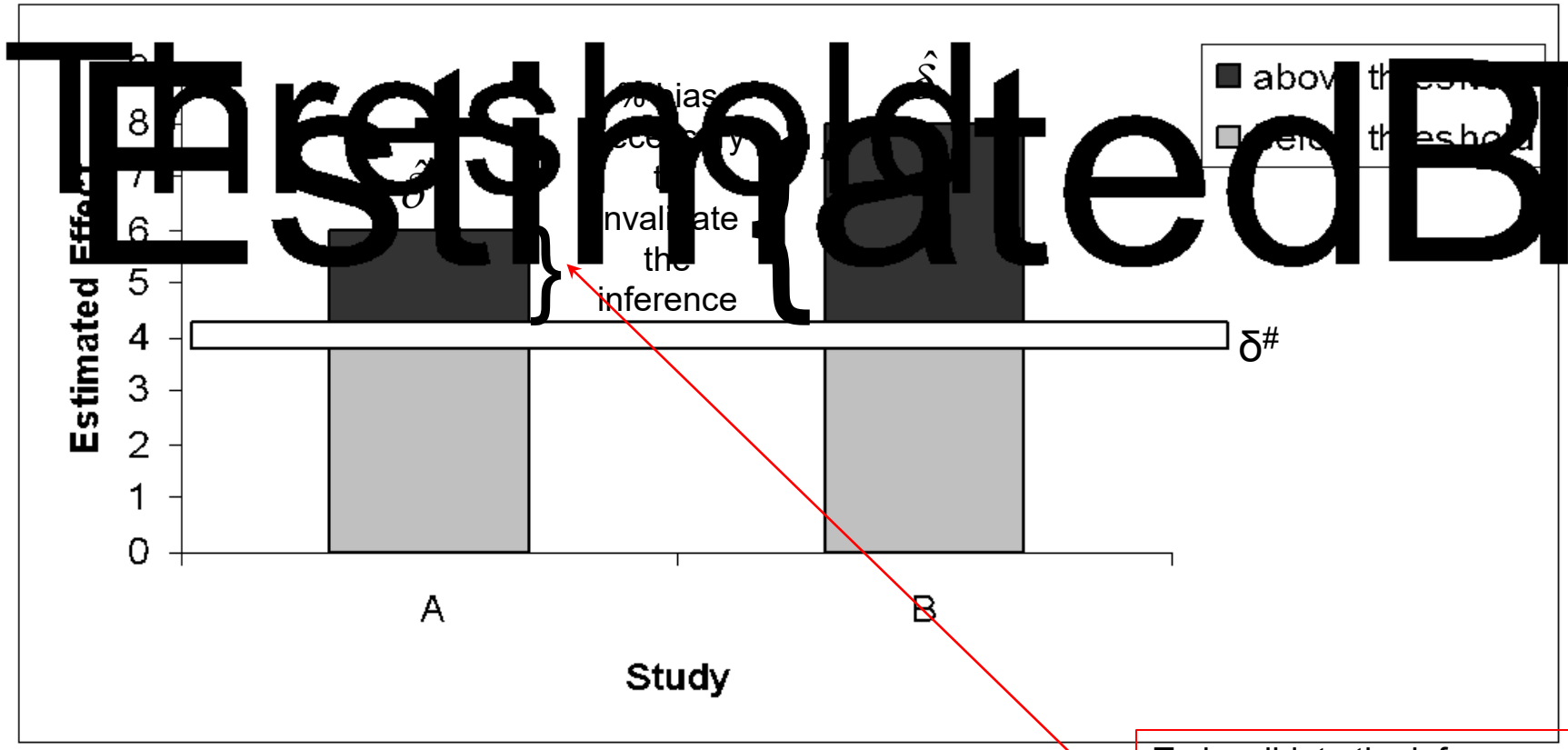
To invalidate the inference, replace 33% of cases with counterfactual cases with zero effect

Using an Unsampled Population to Invalidate the Inference (External Validity)

	A	B	C	D	E
1			potential population		
2			sampled		not sampled
3	Unit	s	Y	combined	Y
4	1	t	9		
5	2	t	10		
6	3	t	11		
7	4	c	3		
8	5	c	4		
9	6	c	5		
10	7	t			7
11	8	t			7
12	9	t			7
13	10	c			7
14	11	c			7
15	12	c			7
16	effect		6	4	0

How many cases would you have to replace with cases with zero effect to change the inference?
Assume threshold is:
$\delta^{\#} = 4$:
$1 - \delta^{\#} / \hat{\delta}$

=1-4/6=.33 =(1/3)



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

To invalidate the inference, replace 33% of cases with cases from unsampled population with zero effect

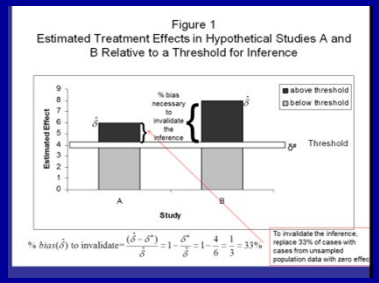
Review & Reflection (in breakout rooms)

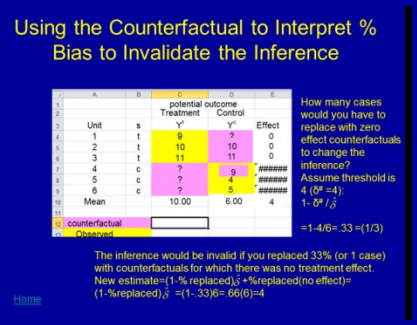
Review of Framework

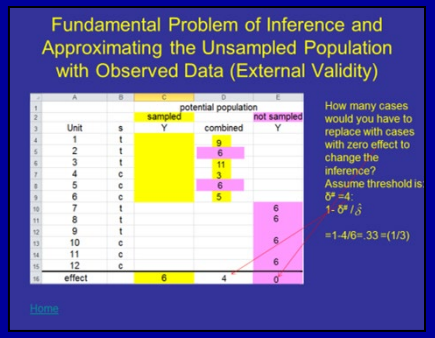
- Pragmatism thresholds
- How much does an estimate exceed the threshold % bias to invalidate the inference
- Interpretation:
- % bias to invalidate number of cases that must be replaced with counterfactual cases (for which there is no effect)

Reflect

- Which part is most confusing to you?
- Is there more than one interpretation?
- Discuss with members of breakout room







Example of Calculating the % Bias to Invalidate and Inference in an 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?

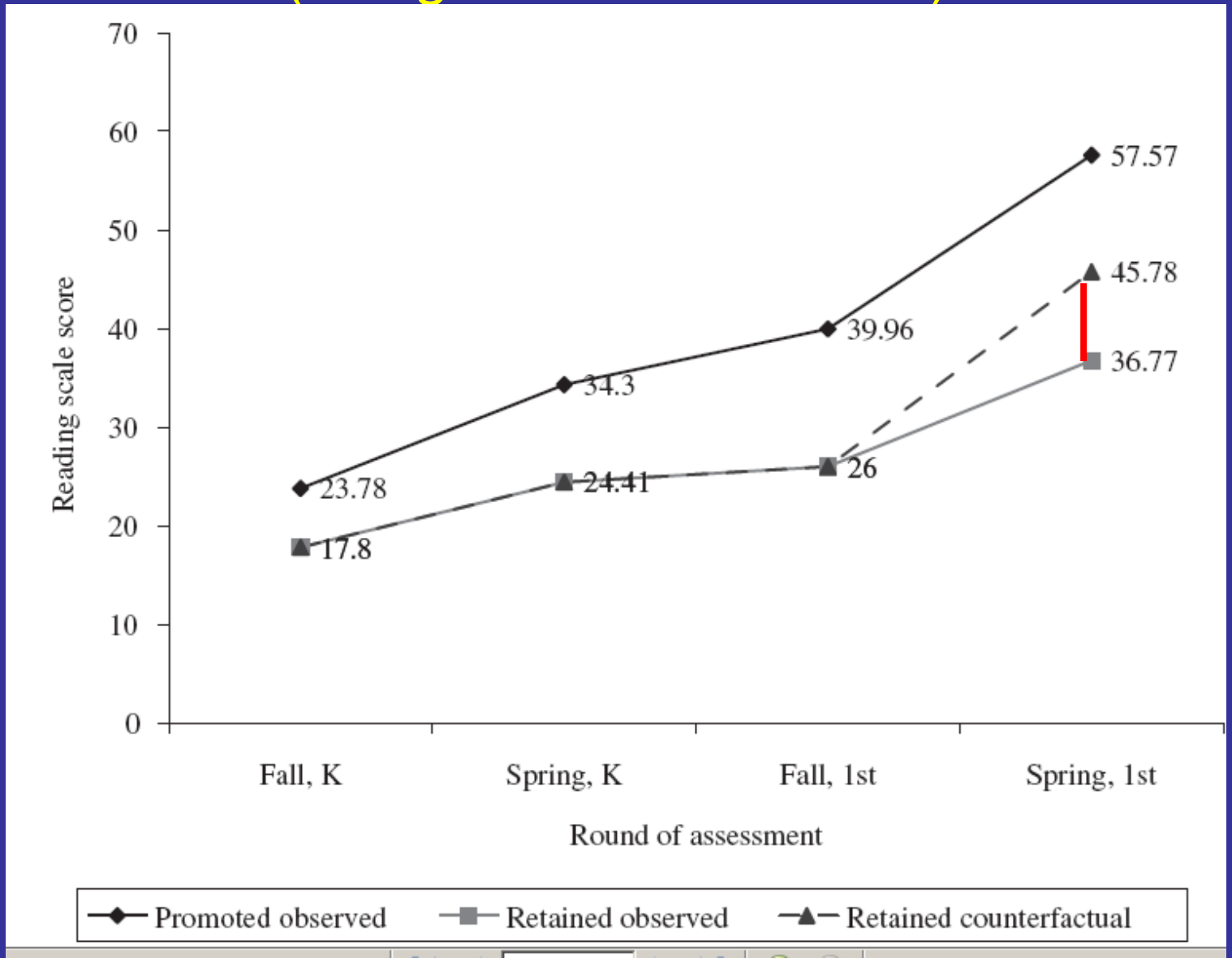
Use principal stratification.

Hong, G. and Raudenbush, S. (2005). Effects of Kindergarten Retention Policy on Children's Cognitive Growth in Reading and Mathematics. *Educational Evaluation and Policy Analysis. Vol. 27, No. 3, pp. 205–224*

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
  - 10,255 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

Obtain df, Estimated Effect and Standard Error

TABLE 11

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

Fixed effect	Coefficient	*SE*
Retention school promoted at-risk kid intercept, γ	53.99	0.28
Retention effect in retention schools, δ_z	−9.01	0.68

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.

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

Standard error=.68

n=7168+471=7639; df > 500,

$t_{critical}$=-1.96

Page 215

From: Hong, G. and Raudenbush, S. (2005). Effects of Kindergarten Retention Policy on Children's Cognitive Growth in Reading and Mathematics. *Educational Evaluation and Policy Analysis. Vol. 27, No. 3, pp. 205–224.*

Pop Quiz

How is the assumption of “no omitted variable” like Alexander Hamilton?

They will never be satisfied!

Calculating % Bias to Invalidate the Inference

1) Calculate threshold $\delta^{\#}$

Estimated effect is statistically significant if:

|Estimated effect| /standard error > $|t_{critical}|$

- |Estimated effect| > $|t_{critical}|$ x standard error= $\delta^{\#}$
- |Estimated effect| > 1.96 x .68 = 1.33= $\delta^{\#}$

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

3) % bias to invalidate the inference is

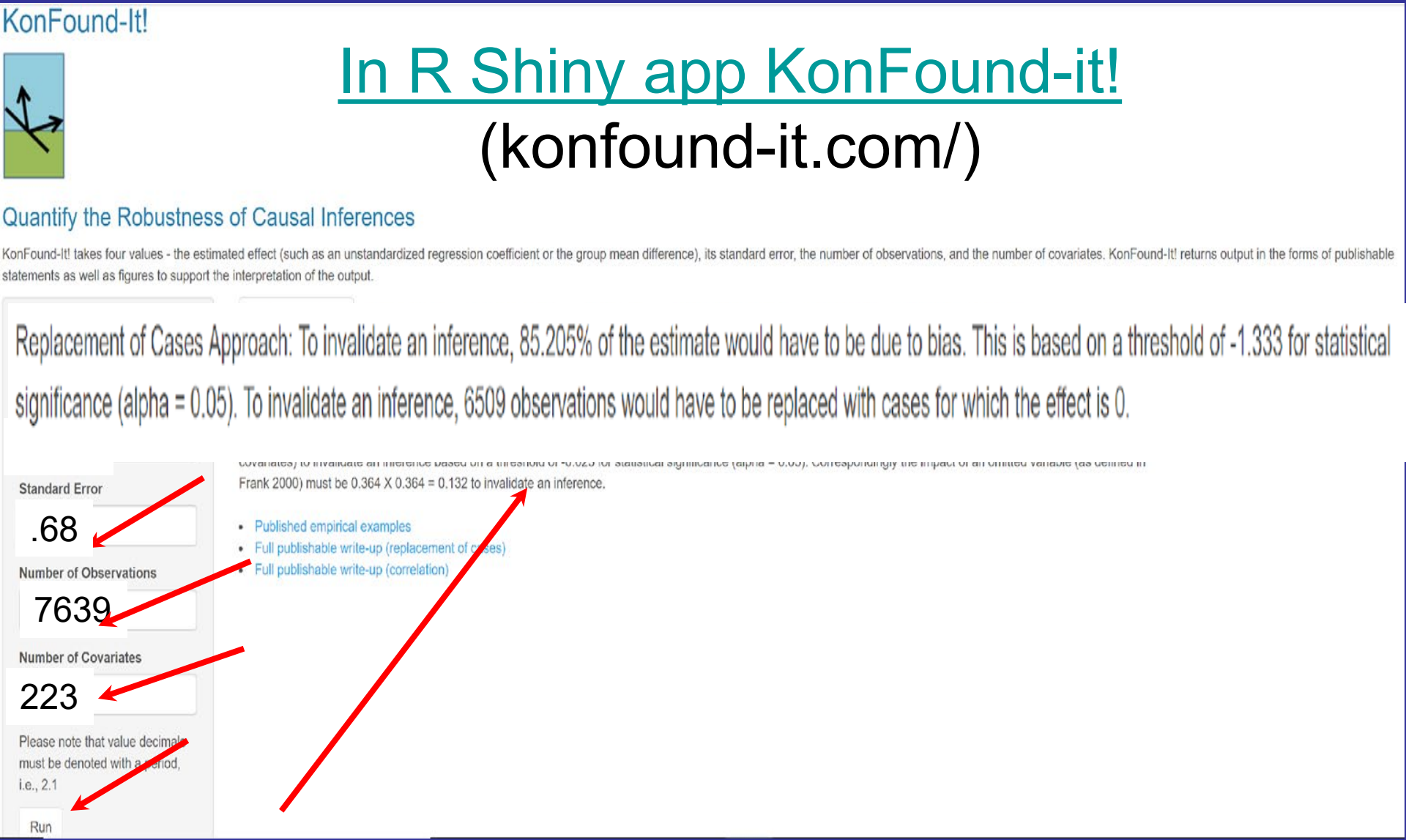
1- $\delta^{\#}/\hat{\delta}$ =1-.1.33/9.01=.85

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/)



Take out your phone and try it!!!

What to publish

In Methods:

We recognize there may be concerns about bias in estimated effects due to unobserved or omitted confounding variables. This bias could have lead to invalid inferences. Our first response is to leverage our data and design as much as possible. In particular, we controlled for pre-measures of achievement, student background (including mother's education), and teacher evaluations. Nonetheless, even after employing these controls there may still be concerns about omitted variables. Therefore we drew on Frank et al (2013) as in the Konfound-it app and quantified how much bias there would have to be due to omitted variables or any other source to invalidate any inferences we made.

In results

Recognizing concerns about potential omitted variables invalidating the inference of an effect of kindergarten retention on achievement, our first response was to leverage our data and design as much as possible. In particular, we controlled for pre-measures of achievement, student background (including mother's education), and teacher evaluations. Nonetheless, even after employing these controls there may still be concerns about omitted variables affecting our estimates and inference. Therefore we drew on Frank et al (2013) as in the http://Konfound-it app and quantified how much bias there would have to be due to omitted variables or any other source to invalidate our inference. This analysis indicates that 85% of the estimated effect of kindergarten retention on achievement would have to be due to bias to invalidate the inference of an effect of retention on achievement. Correspondingly, to invalidate the inference one would have to replace 85% of the observed data with null hypothesis cases of no effect of retention.

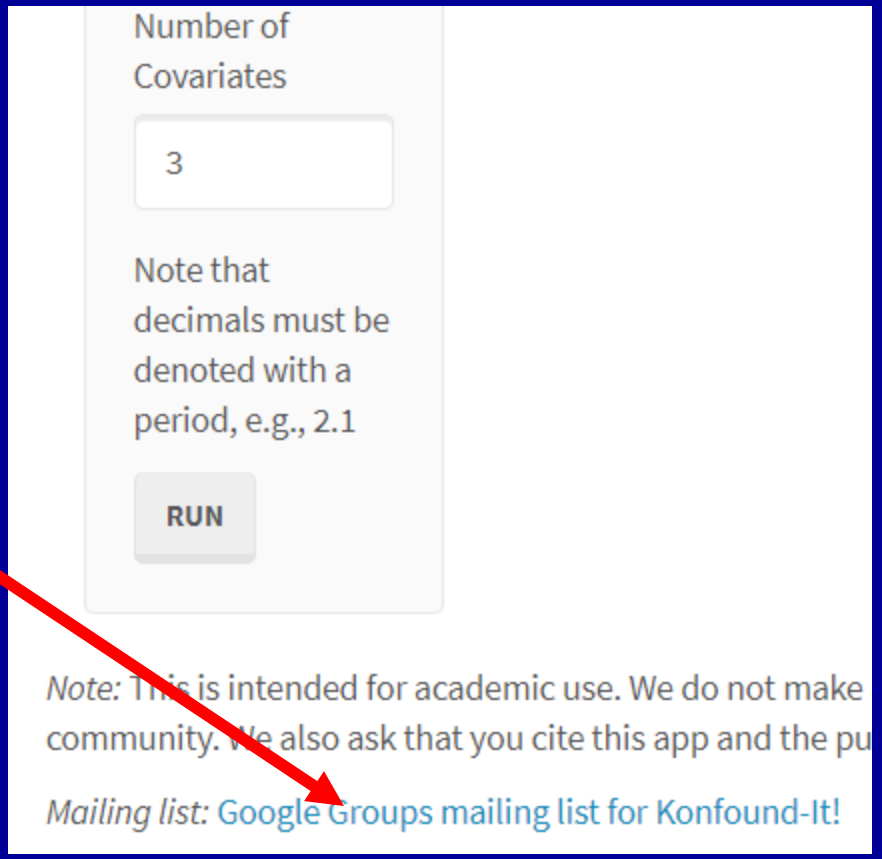
Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437-460.

Rosenberg, J. M., Xu, R., & Frank, K. A. (2018). KonFound-It!: Quantify the robustness of causal inferences.

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.

Join the KonFound-it Google Group

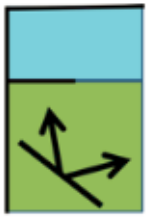
- Software updates
- Workshop notifications
- User community



Please Tweet

Konfound-it.com

KonFound-It!



R and STATA

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 numbe
forms 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!

Estimated Effect

2

Standard Error

0.4

Number of Observations

100

Number of Covariates

3

Run

Results (Printed) | Threshold Plot | Correlation Plot | Workshops | Add to Mobile Device | R and Stata | More Info. & Contact

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:

Konfound-it.com

KonFound-It!

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 numbe
forms 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!

Estimated Effect

2

Standard Error

0.4

Number of Observations

100

Number of Covariates

3

Run

Results (Printed) | Threshold Plot | Correlation Plot | Workshops | Add to Mobile Device | R and Stata | More Info. & Contact

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)
```

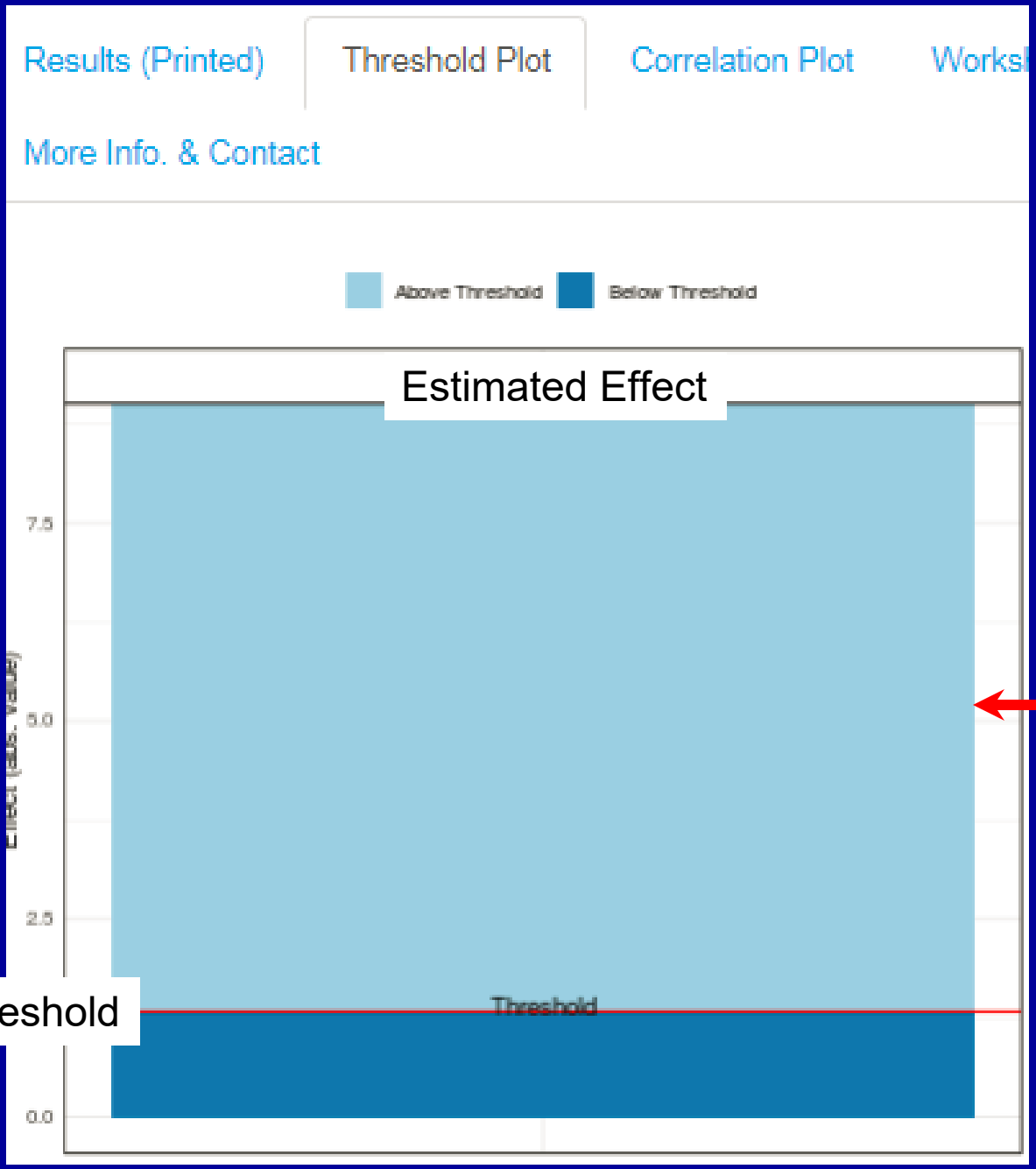
Then you use the following functions for already-published studies, models (including mixed effects models) fit in R, and meta-analyses, respectively:

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

STATA

More information on the STATA module can be found here

For STATA, issue the following commands:



δ# Threshold

% Bias necessary to invalidate inference = 1-δ#/ =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

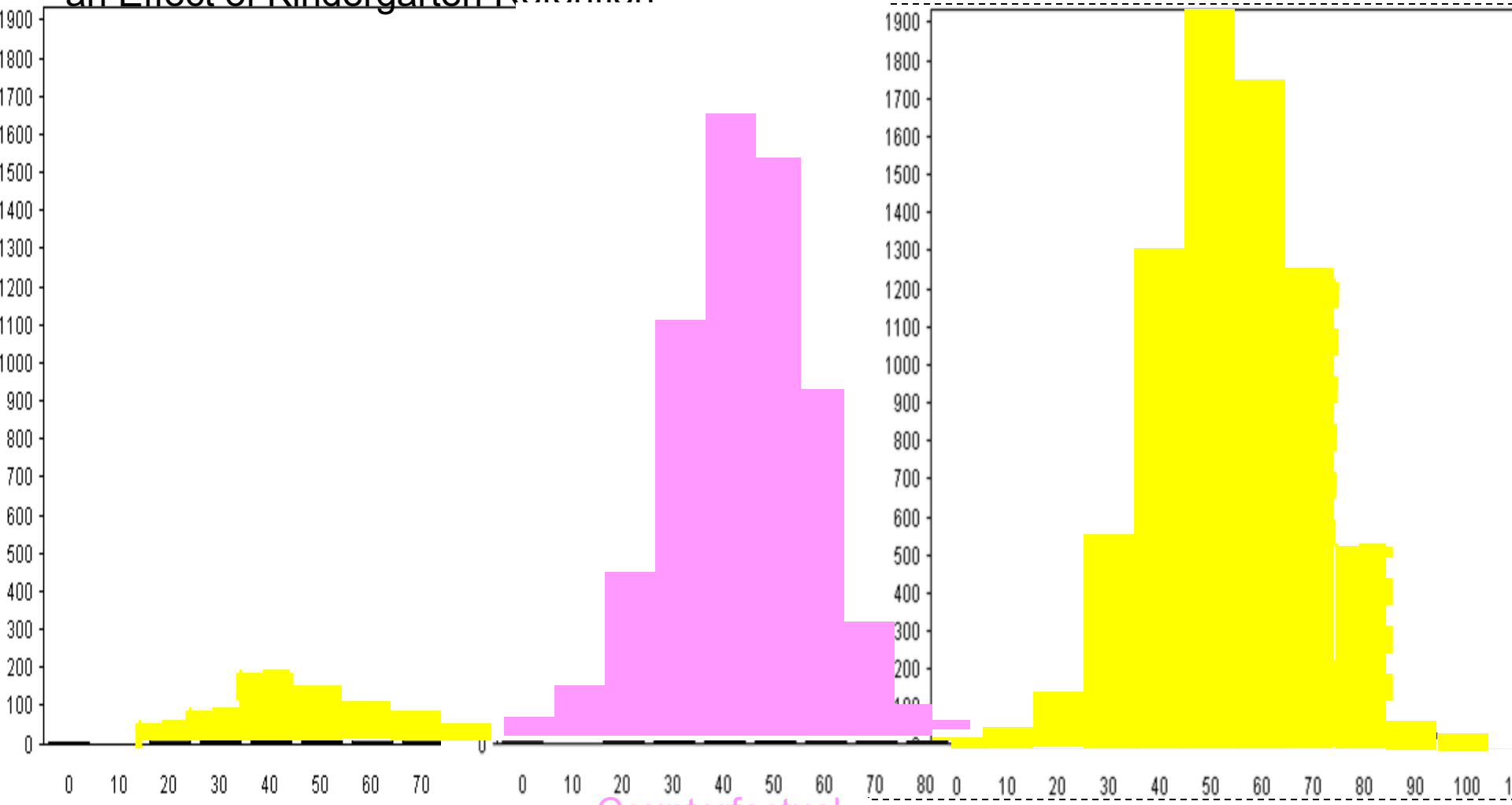
	A	B	C	D	E
1			potential outcome		
2			Treatment	Control	
3	Unit		Y^t	Y^c	Effect
4	1	t	9	7	0
5	2	t	10	7	0
6	3	t	11	7	0
7	4	c	7	3	######
8	5	c	7	4	######
9	6	c	7	5	######
10	Mean		9	5	4
11					
12	counterfactual				
13	Observed				

How many cases would you have to replace with zero effect counterfactuals to change the inference?
Assume threshold is 4 ($\delta^{\#}$ =4):
1- $\delta^{\#}$ / $\hat{\delta}$

=1-4/6=.33 =(1/3)

The inference would be invalid if you replaced 33% (or 2 cases) with counterfactuals for which there was no treatment effect.
New estimate=(1-% replaced)$\hat{\delta}$ +%replaced(no effect)=
(1-%replaced) $\hat{\delta}$ =(1-.33)6=.66(6)=4

Example Replacement of Cases with Counterfactual Data to Invalidate Inference of an Effect of Kindergarten Retention



Original cases that were not replaced
Replacement counterfactual cases with zero effect
Original distribution

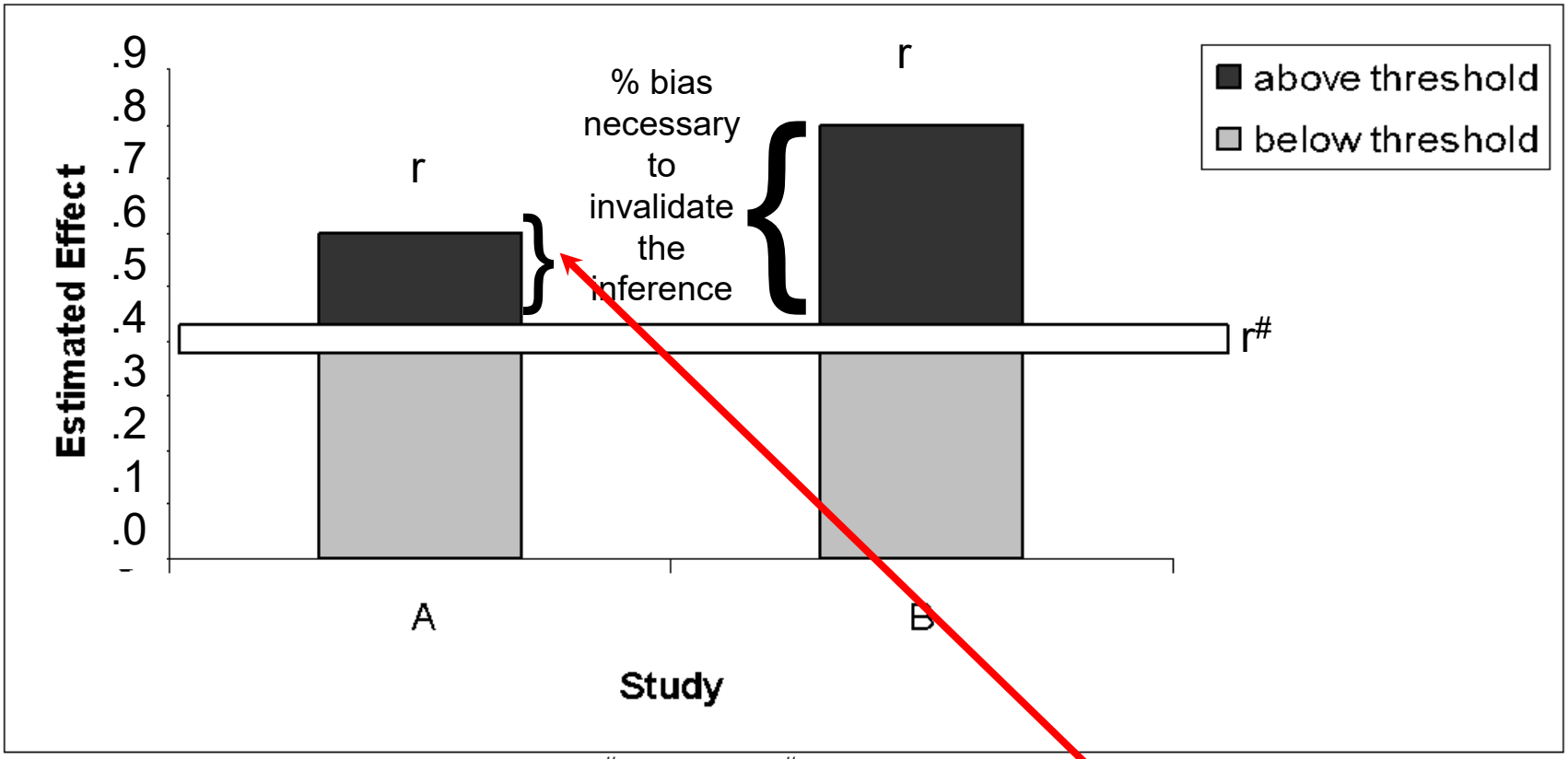
Evaluation of % Bias Necessary to Invalidate Inference

- 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



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

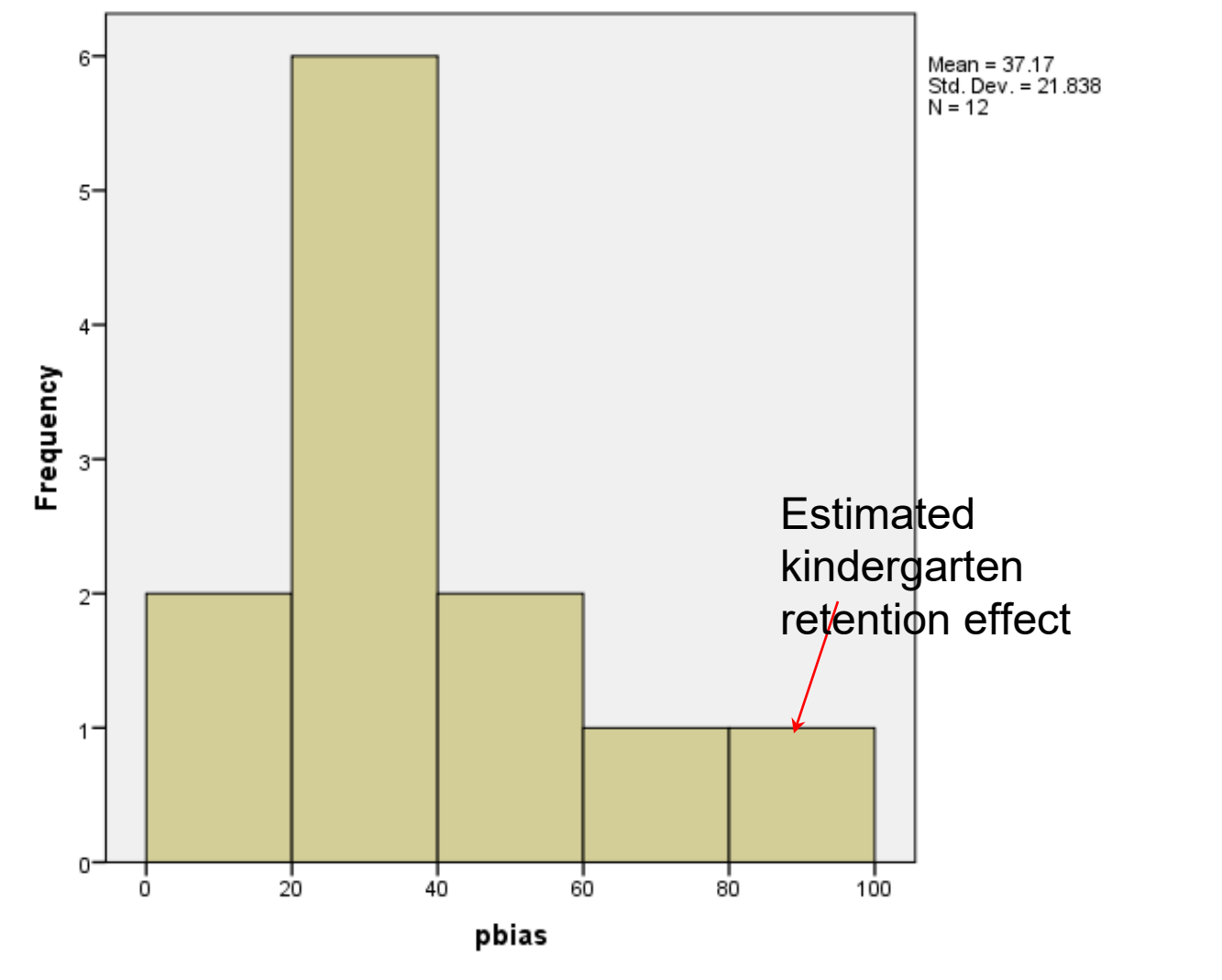
Compare with Bias other Observational Studies

Table 2

Quantifying the Robustness of Inferences from Observational Studies

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

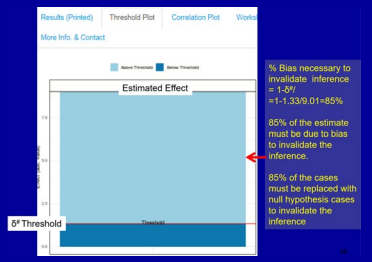
% Bias to Invalidate Inference for Observational Studies on-line EEPA July 24-Nov 15 2012



Review & Reflection (in breakout rooms)

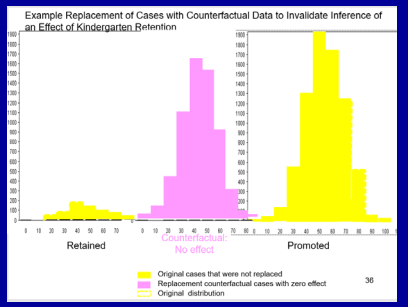
Review of applications

Enter Data and run app



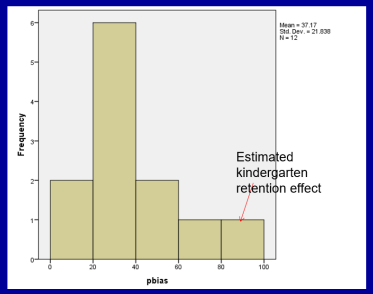
Kindergarten retention (Hong and Raudenbush)

- 85% of cases must be replaced counterfactual data (with no effect) to invalidate the inference of a negative effect of retention on reading achievement



Compare with other studies

% Bias to Invalidate Inference for Observational Studies on-line EEPA July 24-Nov 15 2012



What to publish

In Methods:

We recognize there may be concerns about bias in estimated effects due to unobserved or omitted confounding variables. This bias could have lead to invalid inferences. Our first response is to leverage our data and design as much as possible. In particular, we controlled for pre-measures of achievement, student background (including mother's education), and teacher evaluations. Nonetheless, even after employing these controls there may still be concerns about omitted variables. Therefore we drew on Frank et al (2013) as in the Konfound-it app and quantified how much bias there would have to be due to omitted variables or any other source to invalidate any inferences we made.

In results

Recognizing concerns about potential omitted variables invalidating the inference of an effect of kindergarten retention on achievement, our first response was to leverage our data and design as much as possible. In particular, we controlled for pre-measures of achievement, student background (including mother's education), and teacher evaluations. Nonetheless, even after employing these controls there may still be concerns about omitted variables affecting our estimates and inference. Therefore we drew on Frank et al (2013) as in the http://Konfound-it app and quantified how much bias there would have to be due to omitted variables or any other source to invalidate our inference. This analysis indicates that 85% of the estimated effect of kindergarten retention on achievement would have to be due to bias to invalidate the inference of an effect of retention on achievement. Correspondingly, to invalidate the inference one would have to replace 85% of the observed data with null hypothesis cases of no effect of retention.

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (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(4), 437-460.

Rosenberg, J. M., Xu, R., & Frank, K. A. (2018). KonFound-It!: Quantify the robustness of causal inferences.

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.

Reflect

Which part is most confusing to you?

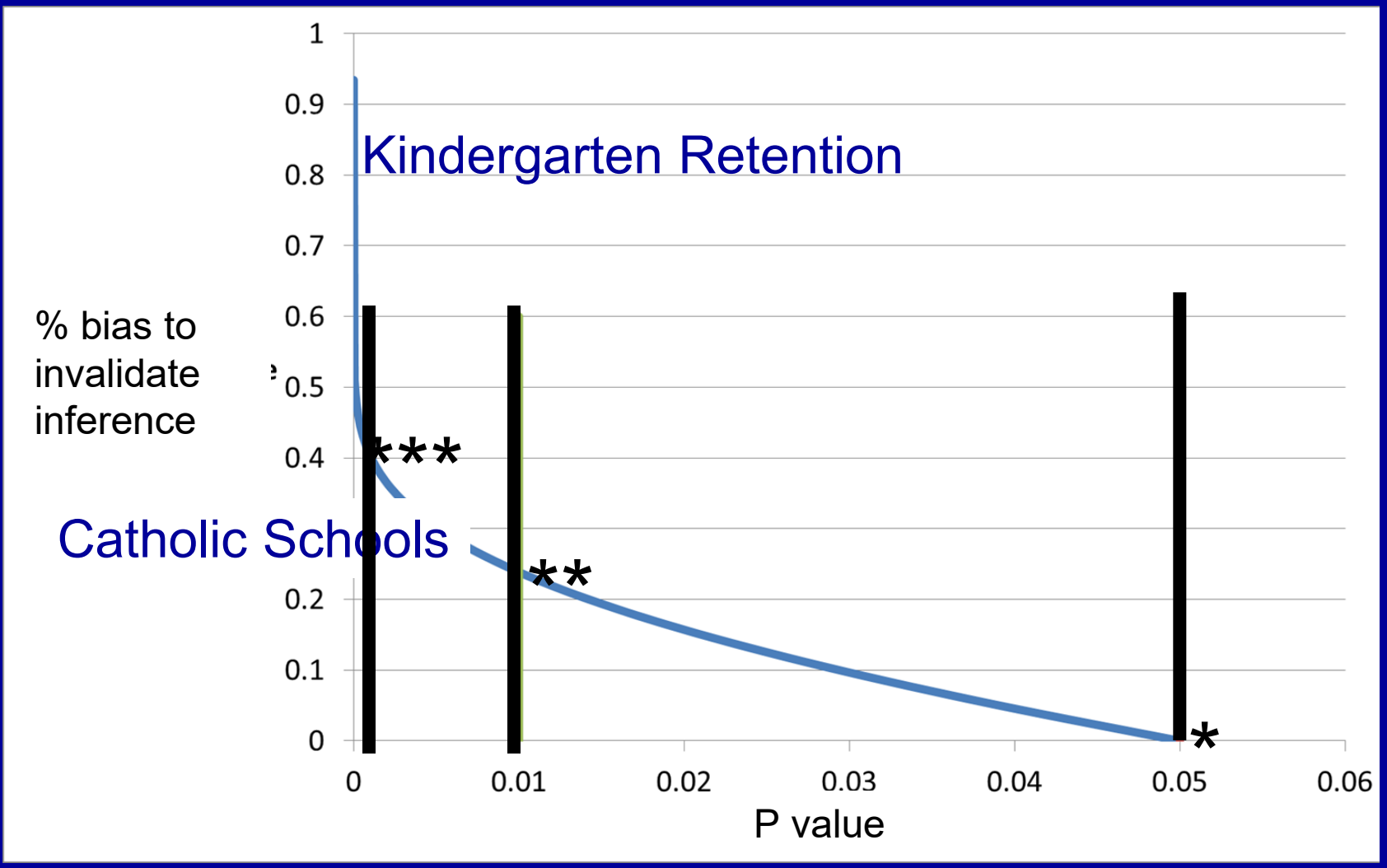
Is there more than one interpretation?

Discuss with members of your group

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
    - Better than “highly significant”
  - Information along a continuous distribution

% 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

Konfound-it.com

KonFound-It!

R and STATA

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 numbe
forms 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!

Estimated Effect

2

Standard Error

0.4

Number of Observations

100

Number of Covariates

3

Run

Results (Printed) | Threshold Plot | Correlation Plot | Workshops | Add to Mobile Device | R and Stata | More Info. & Contact

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:

The Stata Journal (2019)
19, Number 3, pp. 1–28 DOI: !!

konfound: Command to quantify robustness of causal inferences

Ran Xu
Department of Industrial and Systems Engineering
Virginia Tech
Falls Church, VA
ranxu@vt.edu

Kenneth A. Frank
Department of Counseling, Educational Psychology and Special Education
Michigan State University
East Lansing, MI
kenfrank@msu.edu

Spiro J. Maroulis
School of Public Affairs
Arizona State University
Tempe, AR
Spiro.Maroulis@asu.edu

Joshua M. Rosenberg
College of Education, Health, and Human Sciences
University of Tennessee
Knoxville, TN
jmrosenberg@utk.edu

Abstract. Statistical methods that quantify the discourse about causal inferences in terms of possible sources of biases are becoming increasingly important to many social-science fields such as public policy, sociology, and education. These methods are also known as "robustness or sensitivity analyses". A series of recent works (Frank [2000, *Sociological Methods and Research* 29: 147–194]; Pan and Frank [2003, *Journal of Educational and Behavioral Statistics* 28: 315–337]; Frank and Min [2007, *Sociological Methodology* 37: 349–392]; and Frank et al. [2013, *Educational Evaluation and Policy Analysis* 35: 437–460]) on robustness analysis extends earlier methods. We implement these recent developments in Stata. In particular, we provide commands to quantify the percent bias necessary to invalidate an inference from a Rubin causal model framework and the robustness of causal inferences in terms of correlations associated with unobserved variables.

Keywords: st00!!, konfound, mkonfound, pkonfound, causal inferences, bias, confounding, robustness or sensitivity analyses

Apply RIR to Quantifying Strength of Evidence for Inferences in *Value Added Measures*, Accounting for *Spillover Effects*

VAM: teachers are evaluated based on their student achievement.

Sources of bias in VAM:

- Violation of the conditional random assignment assumption
- Spillover: students' achievements are affected by other students
  - Different from baseline peer effects
  - Independent of the teacher

Strength of Evidence in VAMs

- Estimated VAMs are compared to a threshold
  - whether the teacher effect is effective/ineffective
- A threshold a point at which evidence would make one indifferent to the teacher evaluation (Frank et al., 2013)
- But the comparison between the estimated effect and the threshold represents the strength of evidence that supports the inference that directly links to the personnel decision
- calculations of standard errors can be controversial in VAMs arbitrary thresholds not based on statistical significance

RIR for VAM

- Direct application of RIR
  - The estimated effect - VAM
  - Null hypothesis - mean student gain score (an average teacher effect experienced by an average student)
  - Replace observed students (taught by this teacher) by counterfactual of other students if they were taught by this teacher
- For a teacher below the threshold:
  - How many students need to be replaced by average students to invalidate the evaluation?
  - $\pi = \frac{Threshold - VAM}{g_t - VAM}$
- For a teacher above the threshold:
  - How many average level students need to replaced by threshold level students to invalidate the evaluation?
  - $\pi = \frac{VAM - Threshold}{g_t - Threshold}$

Illustrative Example of the Study of Class Size Effect in STAR Project

- 268 teachers from 54 schools
- Math achievement in Grade1
- Three-level HLM model
- EB residuals as VAM

Level 1 (student i)

$$Y_{ijk} = \beta_{0ij} + \beta_{1jk}Pretest_{ijk} + \beta_{2jk}Female_{ijk} + \beta_{3jk}FRL_{ijk} + \beta_{4jk}Minority_{ijk} + \varepsilon_{ijk}$$

Level 2 (teacher/classroom j)

$$\beta_{0jk} = \pi_{00k} + \pi_{01k}Small_{jk} + \pi_{01k}Aide_{jk} + r_{0jk}$$

VAM for teacher j in school k

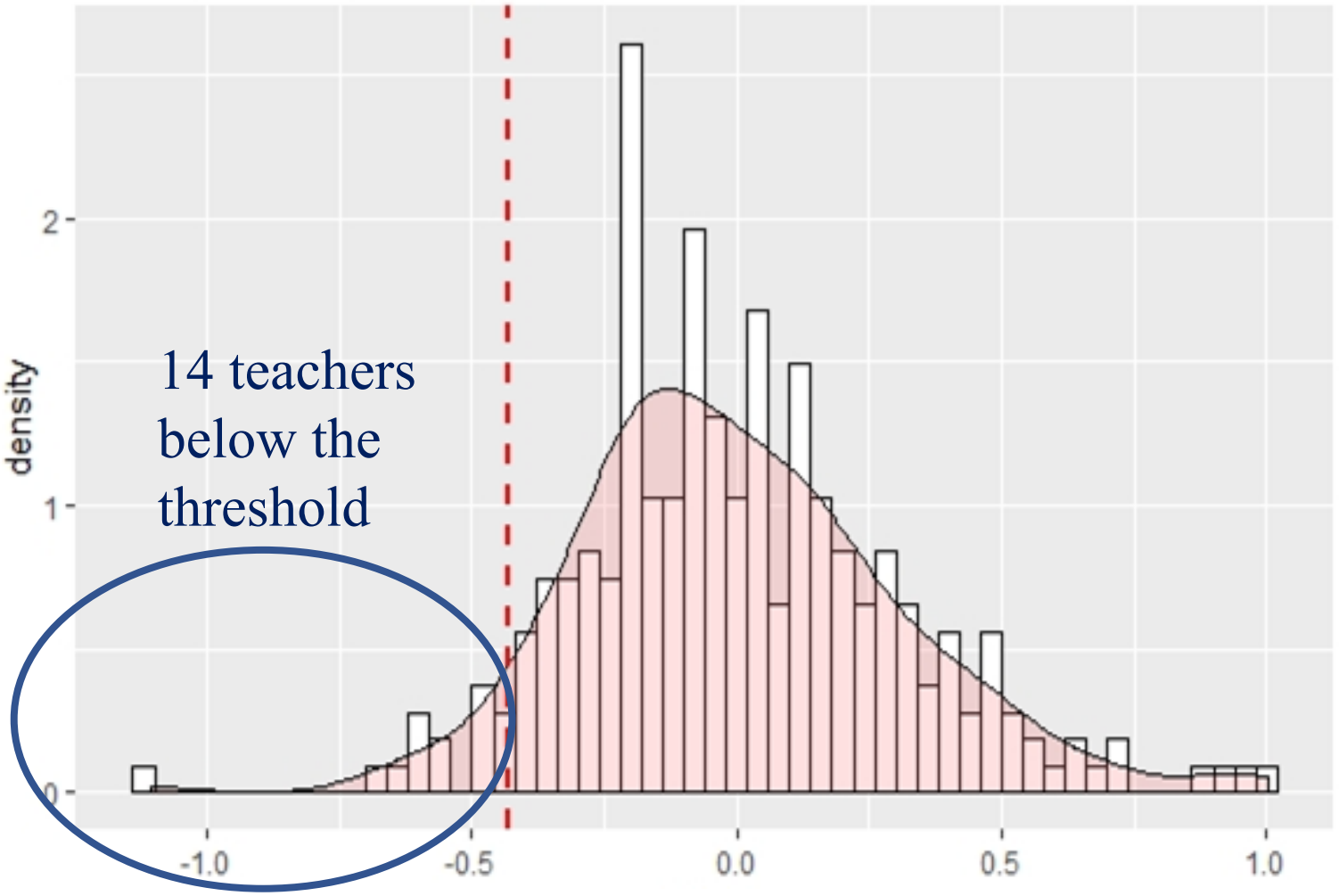
$$\beta_{1jk} = \pi_{10k}\,, \beta_{2jk} = \pi_{20k}, \beta_{3jk} = \pi_{30k}, \beta_{4jk} = \pi_{40k}$$

Level 3 (school k)

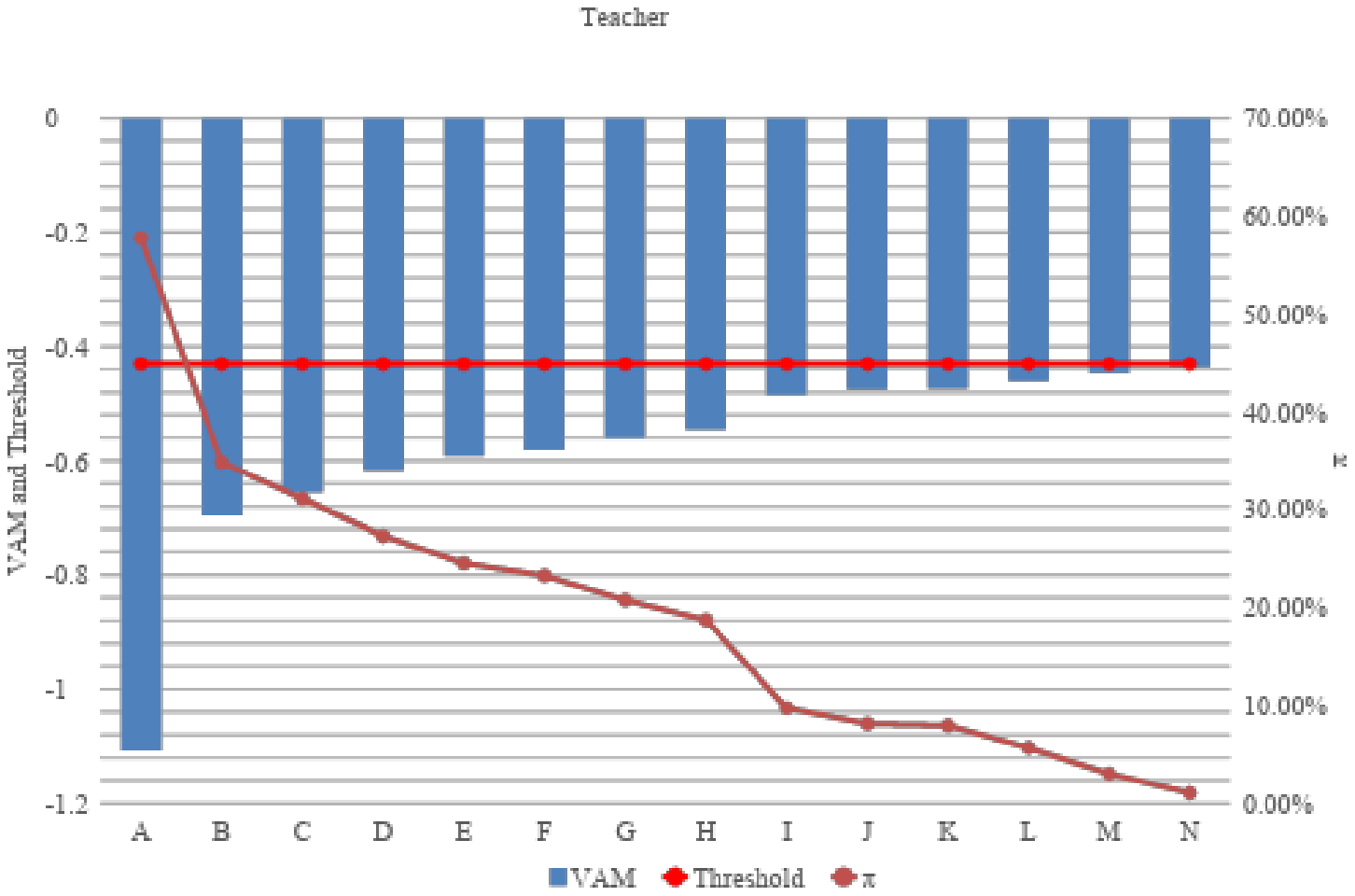
$$\pi_{00k} = \gamma_{000} + \eta_{00k}$$

$$\pi_{10k} = \gamma_{100}\,, \pi_{20k} = \gamma_{200}, \pi_{30k} = \gamma_{300}, \pi_{40k} = \gamma_{400}$$

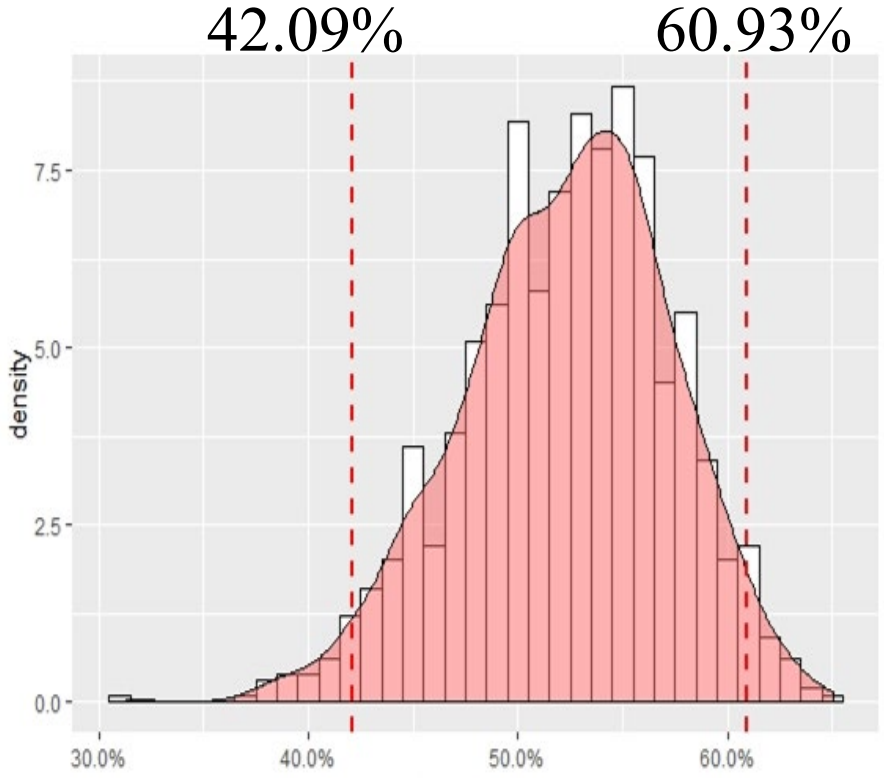
VAMs for 268 teachers



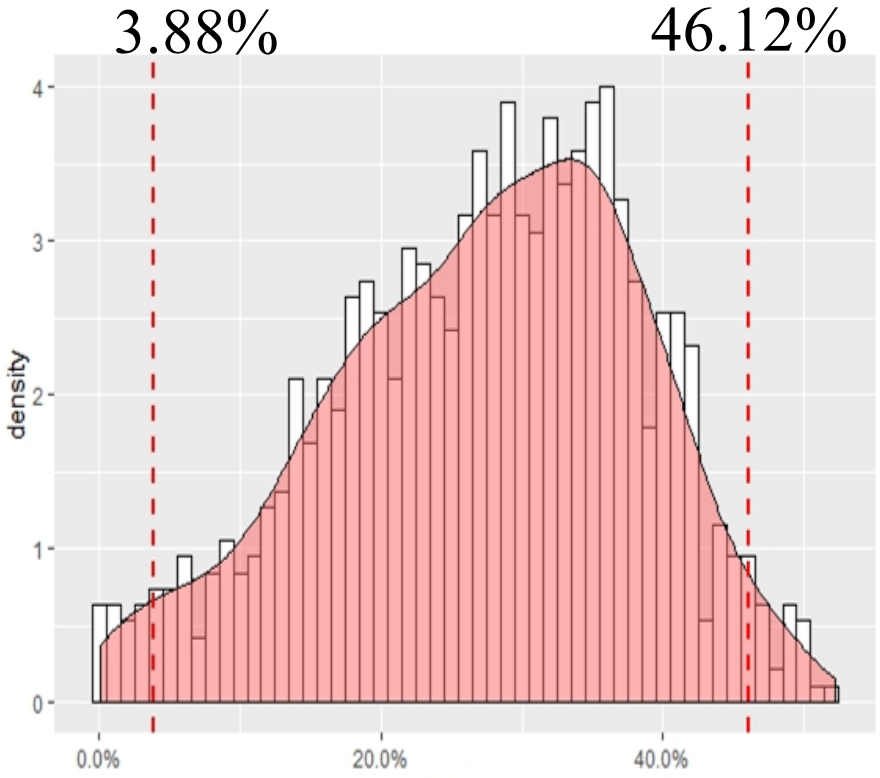
VAMs for 14 ineffective teachers and their robustness in terms of percent of students need to be replaced (π).



Bootstrap (within teachers) to account for sampling variability



Teacher A ($\pi = 57.74\%$)



Teacher B ($\pi = 34.82\%$)

The evaluation for Teacher A is robust, even after accounting for sampling variability.

RIR for VAM – negative spillover effects

- PID: peer-initiated distraction effect
- SID: self-initiated distraction effect

Individuals within one class		Distract others		Distracted by others	
		A	*B*	*C*	*D*
Distract others	*A*	*NA*	*B* □ *A* = *0*	*C* □ *A* = *SID*	*D* □ *A* = *SID*
	B	*A* □ *B* = *0*	*NA*	*C* □ *B* = *SID*	*D* □ *B* = *SID*
Distracted by others	*C*	*A* □ *C* = *PID*	*B* □ *C* = *PID*	*NA*	*D* □ *C* = *0*
			B □ *D* =		

Different ways of interpretation

1. Given R (unit of peer effect, e.g., $NE = PID + SID$):

a larger $\pi \rightarrow$ more students teaching/distracting others $\rightarrow$ larger difference between VAM and Thr, more robust inference

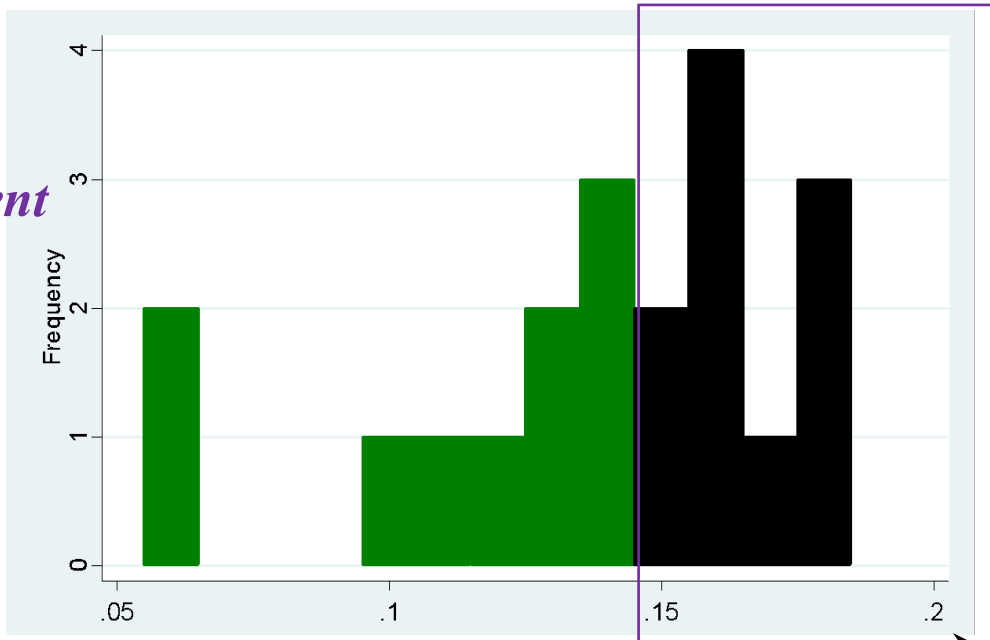
2. Given π:

a larger $R \rightarrow$ larger unit of peer effect (stronger distraction) $\rightarrow$ stronger evidence

3. Counterfactual – use π to answer:

replacing students with other students who experience the same teacher effect but no spillover effects

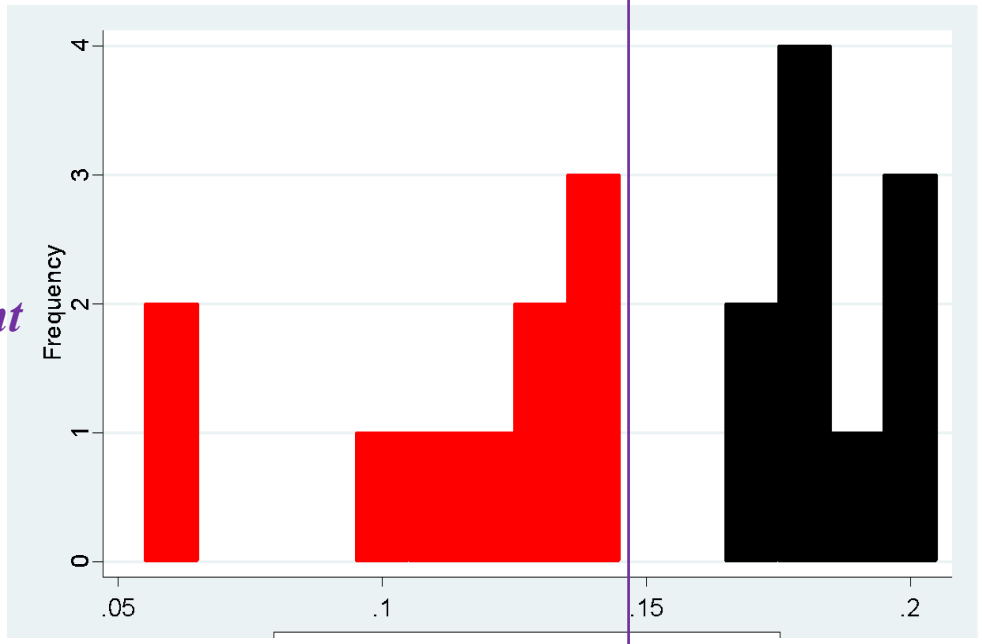
Before replacement



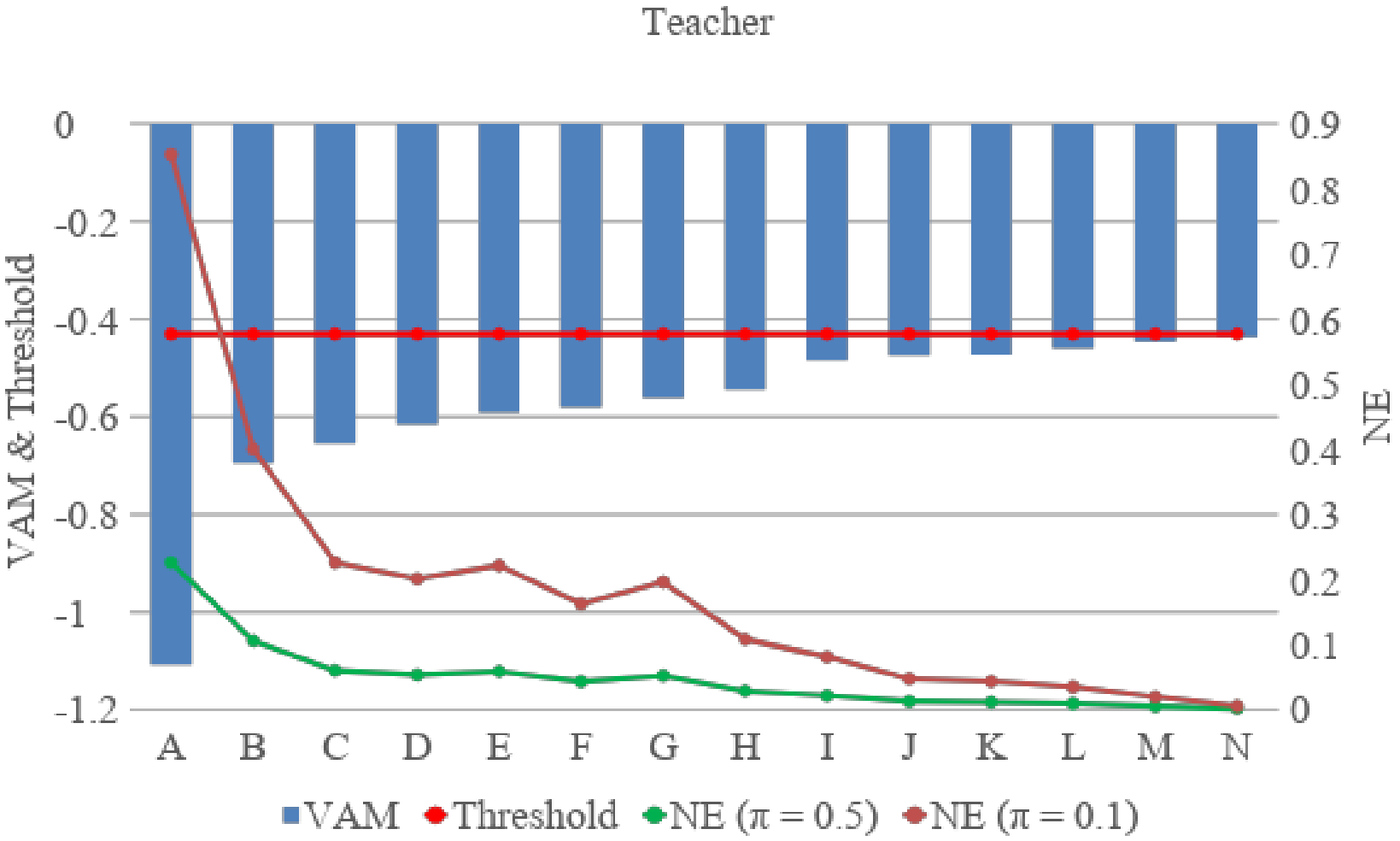
Green:
Original students that were replaced;
Red:
Replacement/hypothetical students that can only bring baseline peer effects, no distraction;
Black:
Original students who stay in the class.

Disappearance of peer-initiated distraction effects (NE=PID+SID)

After replacement



VAMs for 14 ineffective teachers and their robustness to potential peer effects as bias.



In Donald Rubin's words

"Nothing is wrong with making assumptions; on the contrary, such assumptions are the strands that join the field of statistics to scientific disciplines. The quality of these assumptions and their precise explication, not their existence, is the issue"(Rubin, 2004, page 345).

Ken adds: and we should talk about the inferences in terms of the sensitivity to the assumptions

Paul Holland and Don Rubin

"This [the use of randomization to alleviate the problems of untestable assumptions] should not be interpreted as meaning that randomization is necessary for drawing causal inferences. In many cases, appropriate untestable, assumptions will be well supported by intuition, theory or past evidence. In such cases we should not avoid drawing causal inferences and hide behind the cover of uninteresting descriptive statements. Rather we should make causal statements that explicate the underlying assumptions and justify them as well as possible."

The Taboo Against Explicit Causal Inference in Nonexperimental Psychology

Michael P. Grosza , Julia M. Rohrerb,c, and Felix Thoemmesd

Causal inference is a central goal of research. However, most psychologists refrain from explicitly addressing causal research questions and avoid drawing causal inference on the basis of nonexperimental evidence. We argue that this taboo against causal inference in nonexperimental psychology impairs study design and data analysis, holds back cumulative research, leads to a disconnect between original findings and how they are interpreted in subsequent work, and limits the relevance of nonexperimental psychology for policy making. At the same time, the taboo does not prevent researchers from interpreting findings as causal effects—the inference is simply made implicitly, and assumptions remain unarticulated. Thus, we recommend that nonexperimental psychologists begin to talk openly about causal assumptions and causal effects. Only then can researchers take advantage of recent methodological advances in causal reasoning and analysis and develop a solid understanding of the underlying causal mechanisms that can inform future research, theory, and policy makers.

The Debate

You say: I infer X has an effect on Y

They say: you did this from a regression. Regression is not a strong basis for a causal inference. It's not "causal."

You say: regression works if you have the right variables in the model, especially pretests

They say. Yes, but you may not have the right variables. You must do something fancier like propensity score matching or instrumental variables.

You say: propensity scores are no better than the variables that go in them (Heckman, 2005; Morgan & Harding, 2006, page 40; Rosenbaum, 2002, page 297; Shadish et al., 2002, page 164);

Propensity=f(covariates).

Instrumental variables assume there is a variable that predicts the treatment but not directly the outcome, and that treatments can be strongly manipulated (Larcker and Rusticus). Instead, lets talk about what it would take to change my inference…

Quantifying What it Would Take to Change an Inference: Toward a Pragmatic Sociology

Ken Frank (kenfrank@msu.edu)

ASA August 10 2020

R Shiny app KonFound-it! (konfound-it.com/)

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

Estimated Effect

-9.01

Standard Error

.68

Number of Observations

7639

Number of Covariates

0

Run

Results (Printed) | Threshold Plot | Correlation Plot | Workshops | Add to Mobile Device | R and Stata

More Info. & Contact

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.

Correlation-based Approach: An omitted variable would have to be correlated at 0.361 with the outcome and at 0.361 with the predictor of interest (conditioning on observed covariates) to invalidate an inference based on a threshold of -0.022 for statistical significance (alpha = 0.05). Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be 0.361 X 0.361 = 0.13 to invalidate an inference.

- Published empirical examples
- Full publishable write-up (replacement of cases)
- Full publishable write-up (correlation)

Can you *Prove* your Car Will Make it Through?

Pragmatism: Is there enough evidence to act?

End Here

In STATA

. pkonfound -9.01 .68 7639 221
. /* pkonfound estimate standard_error n number_of_covariates */

```
. . pkonfound -9.01 .68 7639 223
------------------
Impact Threshold for Omitted Variable

An omitted variable would have to be correlated at 0.364 with the outcome and at -.364 with the predictor
of interest (conditioning on observed covariates. Signs are interchangeable) to invalidate an inference.
Correspondingly the impact of an omitted variable (as defined in Frank 2000) must be
0.364 x -.364=-0.1324 to invalidate an inference.
------------------
The Threshold for % Bias to Invalidate/Sustain the Inference

To invalidate the inference 85.21% of the estimate would have to be due to bias; to invalidate the
inference 85.21% (6509) cases would have to be replaced with cases for which there is an effect of 0.
```

IGNORE FOR NOW

USE THIS

STATA Commands: Regression

```
. use http://stats.idre.ucla.edu/stat/stata/examples/rwg/concord1
(Hamilton (1983))
. reg water81 water80 income educat retire peop80
      Source |       SS           df       MS      Number of obs   =       496
-------------+----------------------------------   F(5, 490)       =    194.82
       Model |   727354309         5   145470862   Prob > F        =    0.0000
    Residual |   365884401       490   746702.858   R-squared       =    0.6653
-------------+----------------------------------   Adj R-squared   =    0.6619
       Total |  1.0932e+09       495  2208563.05   Root MSE        =    864.12
------------------------------------------------------------------------------
     water81 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     water80 |   .4943149   .0268001    18.44   0.000     .4416577    .5469722
      income |   22.60311   3.502279     6.45   0.000     15.72177    29.48445
      educat |  -44.25776   13.43811    -3.29   0.001     -70.6612   -17.85433
      retire |   155.4727   96.33892     1.61   0.107    -33.81568    344.761
      peop80 |   225.1984   28.70482     7.85   0.000     168.7987    281.5981
       _cons |   299.7437   210.0136     1.43   0.154    -112.8947    712.3821
------------------------------------------------------------------------------

. konfound peop80
```

STATA Command: Konfound

```
. konfound peop80
------------------
The Threshold for % Bias to Invalidate/Sustain the Inference

For peop80:
To invalidate the inference 74.96% of the estimate would have to be due to bias; to invalidate the
inference 74.96% (372) cases would have to be replaced with cases for which there is an effect of
0.
------------------
Impact Threshold for Omitted Variable
For peop80:
An omitted variable would have to be correlated at 0.519 with the outcome and at 0.519 with the
predictor
of interest (conditioning on observed covariates) to invalidate an inference.
```

In R: Pkonfound (published example)

install.packages("konfound")

library(konfound)

pkonfound(est_eff = -9.01, std_err = .68, n_obs = 7639, n_covariates = 221)

```
> 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 statist
ical 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 -0.023 for statistical significance (alpha = 0.05).
Correspondingly, the impact of an omitted variable (as defined in Frank 2000) must be 0.364 X 0.364 = 0.132 to invalidate an inferenc
e.
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().
```

USE THIS

IGNORE FOR NOW

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)
```

```
> model

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

Coefficients:
(Intercept)           X1           X4
    4.33291      0.45073     -0.09873
```

USE THIS

```
> 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 pred
ictor of interest (conditioning on observed covariates) to invalidate an inference based on a thresh
old 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.
```

IGNORE FOR NOW

Exercise 1 : % Bias necessary to Invalidate an Inference for Internal Validity [in breakout rooms]

As a group:

Take an example from an observational study.
Calculate the % bias necessary to invalidate the inference

[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

Discuss with members of your group.