Finding and Handling Bias in Clustered Randomized Trials

Bradford Chaney, National Academies of Sciences, Engineering, and Medicine

Disclaimer: The following presentation reflects my own research, and should not be interpreted as representing the positions of either the National Academies, where I am currently employed, or Westat, where much of the research was performed.

Sample Overstatement (journal article)

In the analyses that follow, we add the pretest covariate in order to increase the precision of the impact estimate. (But we recognize that, with or without this covariate, the impact estimate is unbiased as a result of the randomization.)

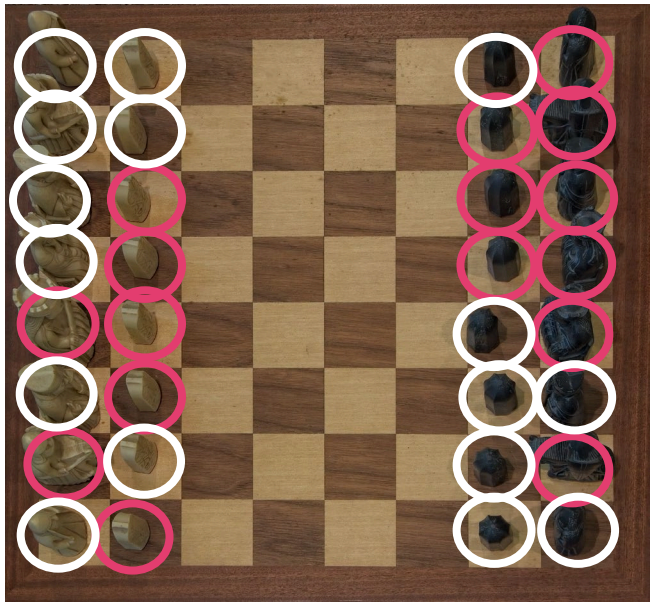
Another Sample Overstatement (email from journal editor)

The issue of model specification is actually much diminished as a challenge for experimental evaluations where only simple math (T outcome minus C outcome) is needed to produce an unbiased impact estimate (moving to a regression context improves precision in that estimate).

The Study Population (to divide into treatment and control groups)

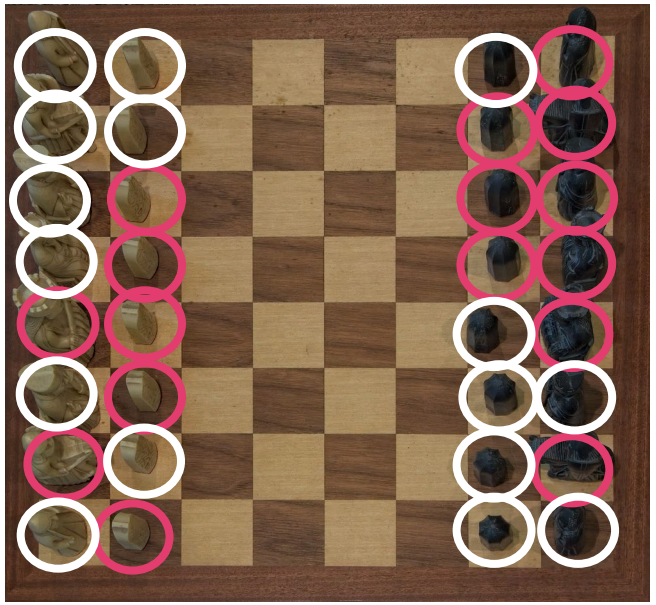
Randomized Controlled Trials

Sample Division



Randomized Controlled Trials

Sample Division

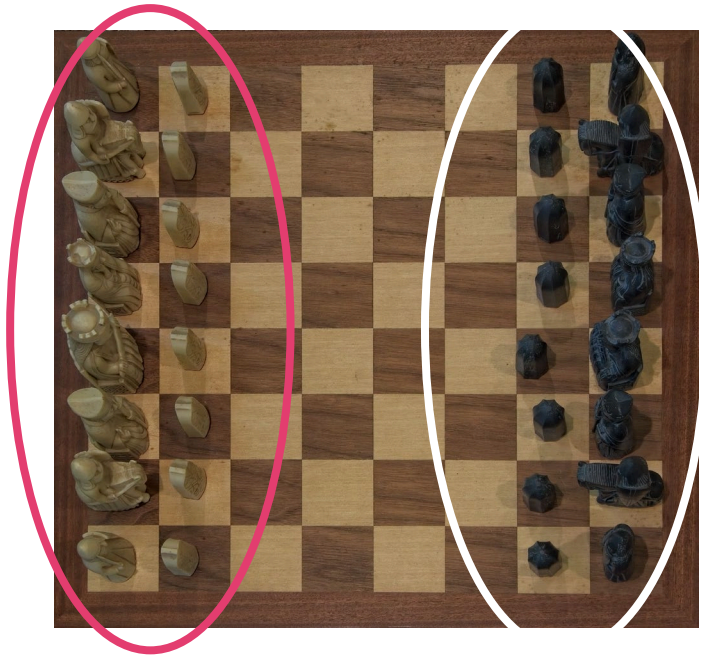


Comparison

Characteristic	Treatment	Control
Rook	1	3
Knight	3	1
Bishop	1	3
Queen	1	1
King	2	0
Pawn	8	8
Black	9	7
White	7	9

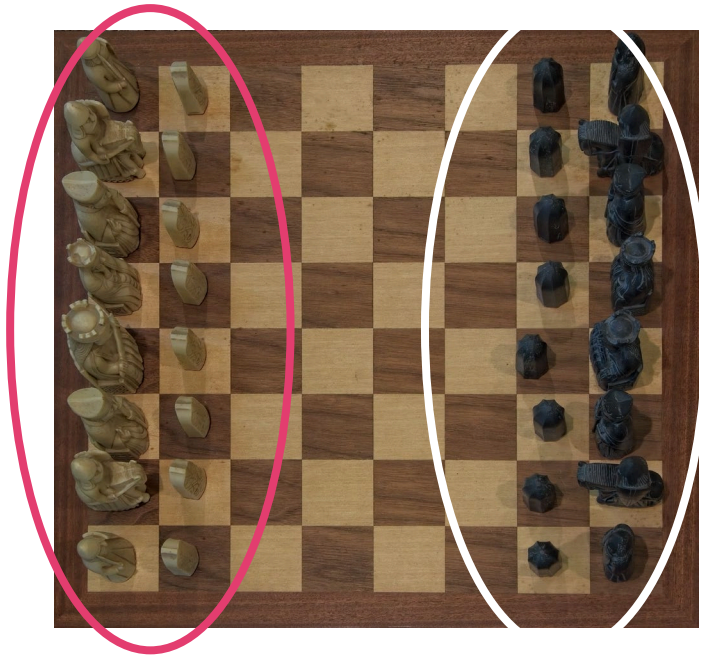
Clustered Randomized Trials

The Easier Way



Clustered Randomized Trials

The Easier Way

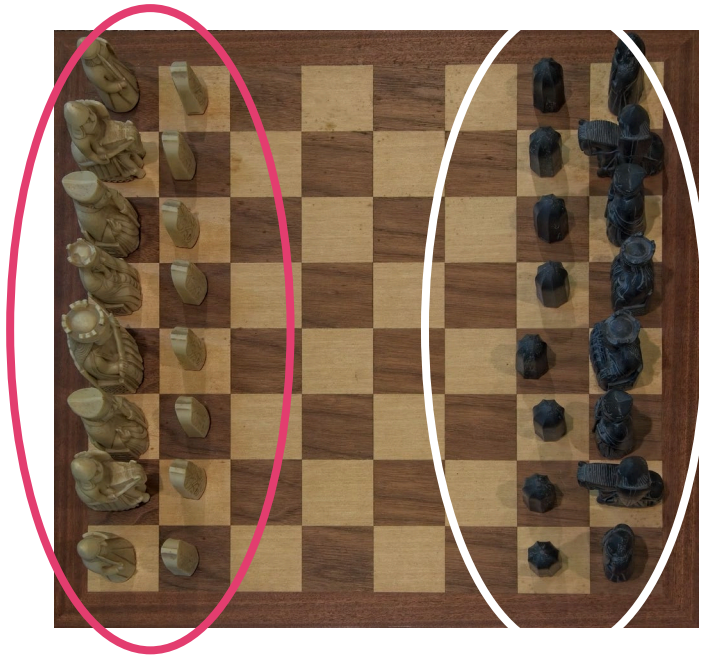


Advantages of CRTs

- Often easier to draw a sample
- Often less expensive
- Often easier to administer
- Helps to prevent contamination
- Statistical theory says this is also an unbiased sample

Clustered Randomized Trials

The Easier Way



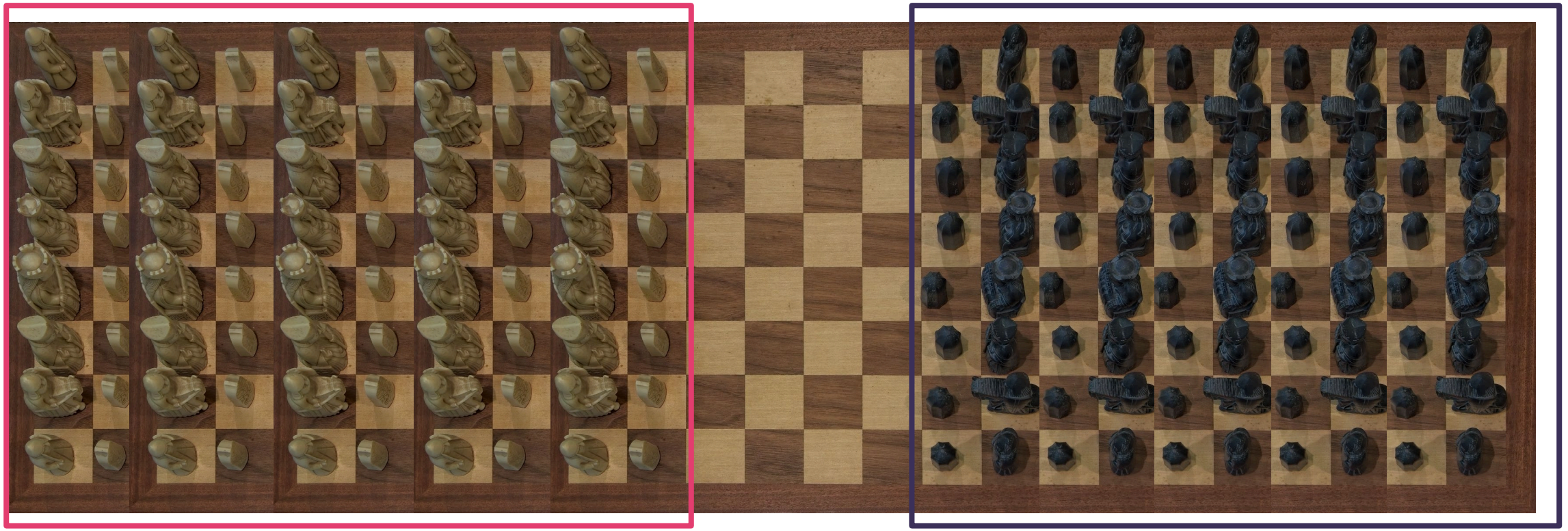
But is it really unbiased?

Characteristic	Treatment	Control
Rook	2	2
Knight	2	2
Bishop	2	2
Queen	1	1
King	1	1
Pawn	8	8
Black	16	0
White	0	16

RCTs are unbiased **in expectation**

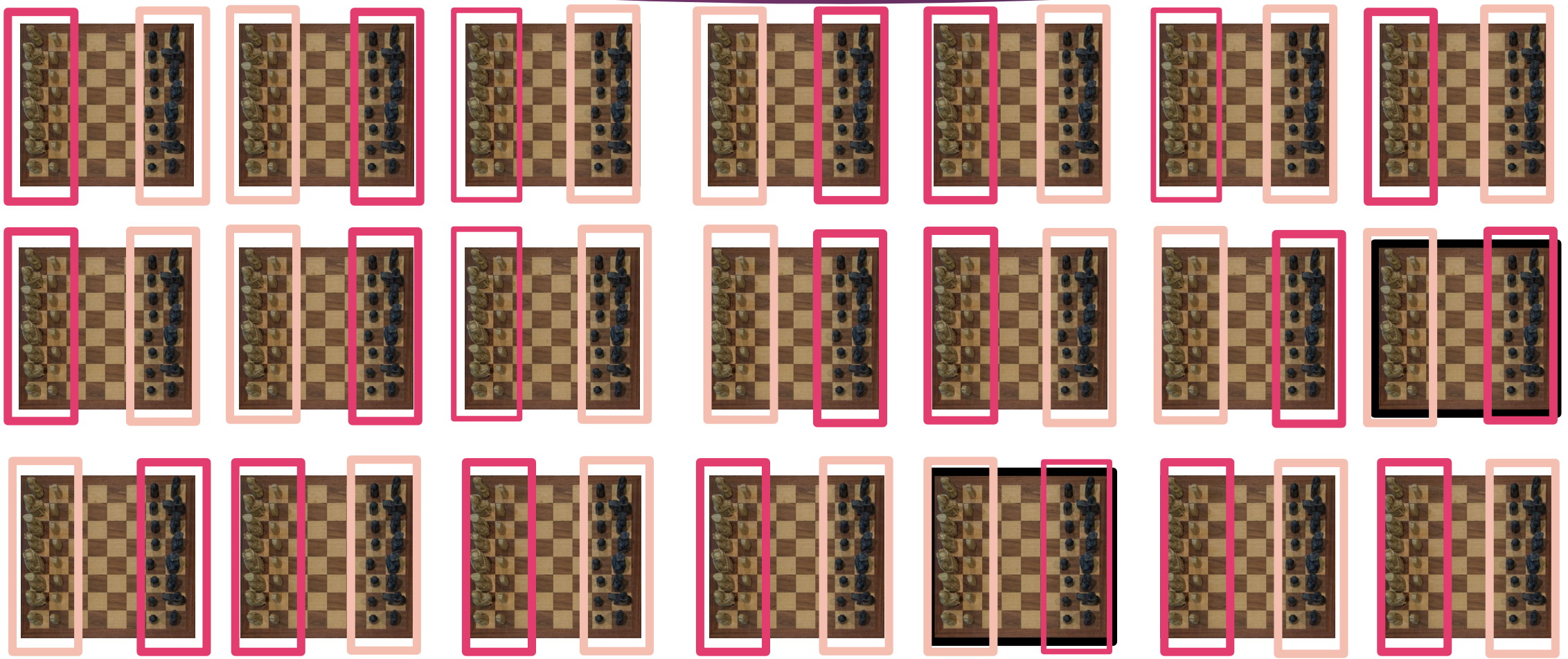
Over 1,000 repetitions, every individual trial will be biased, but collectively they are unbiased.

Alternatively, you an increase the sample size



Increasing the N by itself makes no difference.

It's not the total N but the number of clusters that is important



Relative Frequency of CRTs and RCTs

Topic area	CRTs	RCTs	Other
Knowledge is Power Program	0	1	3
Teach for America	0	3	4
Cognitive Tutor	4	1	2
Pre-K Mathematics	2	0	0
Building Blocks for Math	2	0	0
Lindamood Phoneme Sequencing	1	1	0
Green Dot Public Schools	0	0	1
University of Chicago Mathematics Project	0	0	3
eMINT Comprehensive Program	1	0	1
Odyssey Math	2	0	1
Totals	12	6	15

How many clusters are typical?

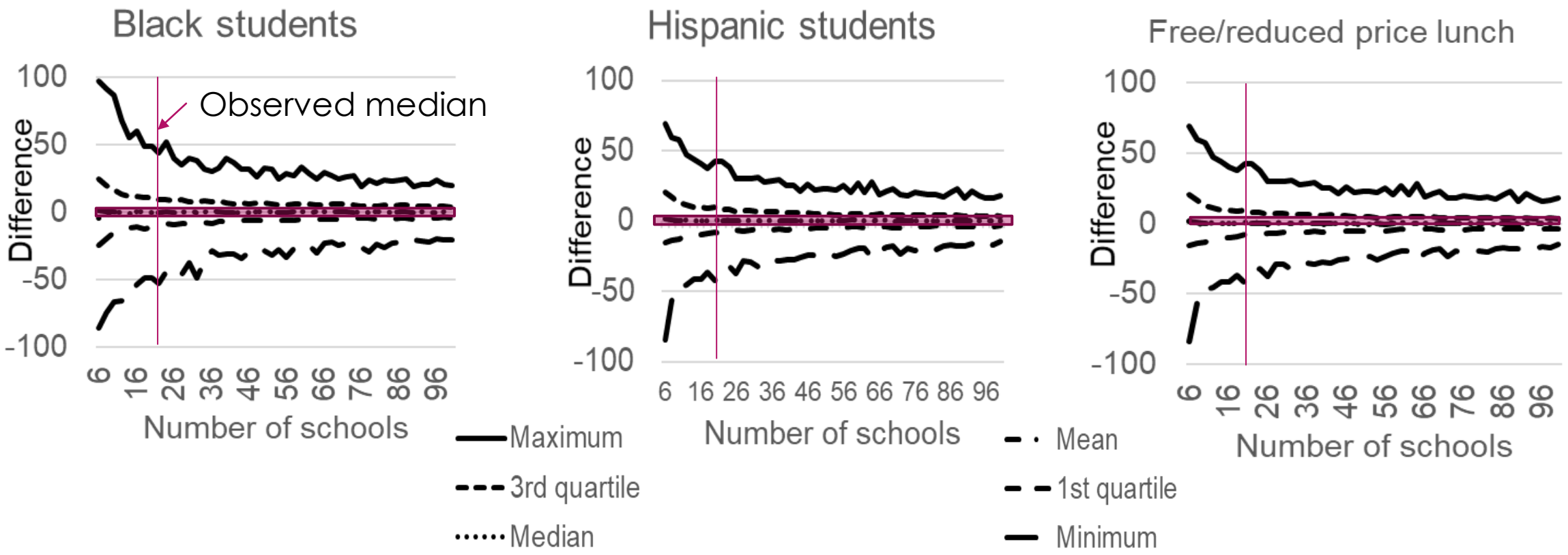
Topic area	CRTs
Cognitive Tutor	6, 9, 22, 73
Pre-K Mathematics	40, 40
Building Blocks for Math	4, 20
Lindamood Phoneme Sequencing	2
eMINT Comprehensive Program	20
Odyssey Math	13, 124

Half have 20 or fewer clusters

Method for Estimating Differences Across Treatment and Control Groups

- Draw 1,000 samples of schools at each of 48 sample sizes (2, 4, …, 98, 100)
  - Source: Common Core of Data, National Center for Education Statistics
- Randomly assign schools to treatment and control groups, with equal numbers of schools in each group
- Calculate the percentage in each group with a particular characteristic (e.g., the percentage who were black)
- Calculate the difference between the treatment and control groups

How Greatly do Treatment and Control Groups Vary Based on Number of Clusters?



Scenario for testing

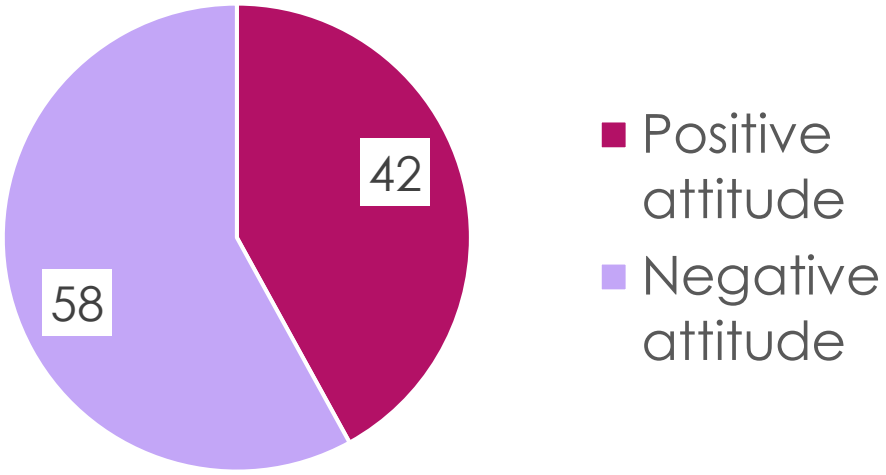
- A new mathematics education program needs to be evaluated.
- The impact of the program will depend on two things:
  - Students' motivation
  - Students' self-efficacy
- Define a treatment effect: Impact = 0.2 x (self-efficacy) + 0.3 x (student motivation)
  - Test once assuming there are no other sources of change
    - Posttest score = test score time 2002 + treatment effect
  - Test again while allowing for change over two years
    - Posttest score = test score time 2004 + treatment effect

Method for Examining how Models Performed in the Presence of Bias

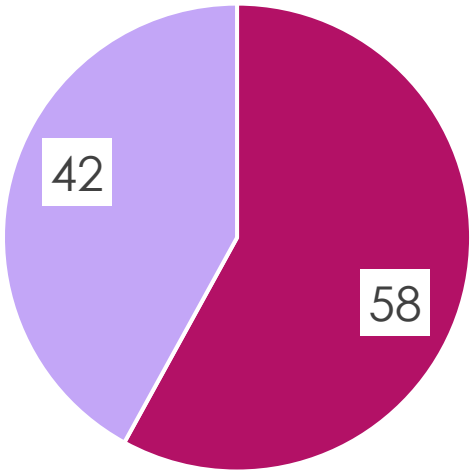
- Draw 1,000 samples of students, each with 2,000 students (frame had 8,590 students)
  - Source: Education Longitudinal Study of 2002, National Center for Education Statistics
- Randomly assign students to treatment and control groups
- Applied bias to random assignment
  - Attitude = Motivation + Self-efficacy

To create bias, summed motivation and self-efficacy

Treatment



Control



Sample distribution based on bias formula

Status	School	Number of students	Math base year scores	Socio-economic status composite	Percent black	Percent white
			Mean	Mean	Mean	Mean
Total	Total	8,590	46.06	0.15	0.10	0.63
Treatment	Total	4,279	45.35	0.13	0.09	0.64
Control	Total	4,311	46.77	0.17	0.10	0.63
Treatment	A	1,564	30.79	-0.19	0.17	0.51
Control	B	1,407	30.79	-0.16	0.19	0.49
Treatment	C	1,470	47.06	0.18	0.06	0.70
Control	D	1,449	46.90	0.17	0.09	0.67
Treatment	E	1,245	61.61	0.46	0.02	0.72
Control	F	1,455	62.09	0.50	0.03	0.73

Sample distribution based on bias formula

Status	School	Number of students	Math base year scores	Socio-economic status composite	Percent black	Percent white
			Mean	Mean	Mean	Mean
Total	Total	8,590	46.06	0.15	0.10	0.63
Treatment	Total	4,279	45.35	0.13	0.09	0.64
Control	Total	4,311	46.77	0.17	0.10	0.63
Treatment	A	1,564	30.79	-0.19	0.17	0.51
Control	B	1,407	30.79	-0.16	0.19	0.49
Treatment	C	1,470	47.06	0.18	0.06	0.70
Control	D	1,449	46.90	0.17	0.09	0.67
Treatment	E	1,245	61.61	0.46	0.02	0.72
Control	F	1,455	62.09	0.50	0.03	0.73

Sample distribution based on bias formula

Status	School	Number of students	Math base year scores	Socio-economic status composite	Percent black	Percent white
			Mean	Mean	Mean	Mean
Total	Total	8,590	46.06	0.15	0.10	0.63
Treatment	Total	4,279	45.35	0.13	0.09	0.64
Control	Total	4,311	46.77	0.17	0.10	0.63
Treatment	A	1,564	30.79	-0.19	0.17	0.51
Control	B	1,407	30.79	-0.16	0.19	0.49
Treatment	C	1,470	47.06	0.18	0.06	0.70
Control	D	1,449	46.90	0.17	0.09	0.67
Treatment	E	1,245	61.61	0.46	0.02	0.72
Control	F	1,455	62.09	0.50	0.03	0.73

Distribution of the First of 1,000 Samples

Group	Number	Test score – no time change	Test score with time change
Students with positive attitude			
Treatment	436	58.25	63.26
Control	555	57.83	62.85
Students with negative attitude			
Treatment	539	35.62	41.32
Control	470	34.40	40.04
All students combined			
Treatment	975	45.74	51.13
Control	1,025	47.09	52.39

Distribution of the First of 1,000 Samples

Group	Number	Test score – no time change	Test score with time change
Students with positive attitude			
Treatment	436	58.25	63.26
Control	555	57.83	62.85
Students with negative attitude			
Treatment	539	35.62	41.32
Control	470	34.40	40.04
All students combined			
Treatment	975	45.74	51.13
Control	1,025	47.09	52.39

Distribution of the First of 1,000 Samples

Group	Number	Test score – no time change	Test score with time change
Students with positive attitude			
Treatment	436	58.25	63.26
Control	555	57.83	62.85
Students with negative attitude			
Treatment	539	35.62	41.32
Control	470	34.40	40.04
All students combined			
Treatment	975	45.74	51.13
Control	1,025	47.09	52.39

Distribution of the First of 1,000 Samples

Group	Number	Test score – no time change	Test score with time change
Students with positive attitude			
Treatment	436	58.25	63.26
Control	555	57.83	62.85
Students with negative attitude			
Treatment	539	35.62	41.32
Control	470	34.40	40.04
All students combined			
Treatment	975	45.74	51.13
Control	1,025	47.09	52.39

This is an incidence of Simpson's paradox

- The outcome for the totals is the opposite direction from the outcome for the subgroups.
- This is because of bias.
  - The treatment group has more students with negative attitudes.
  - Because of their greater number, they can pull down the overall means.
- So what does this say about being able to trust a simple comparison of the means?

Assuming no other sources of change

Model	Student in treatment group	Self-efficacy score (treatment group only)	Motivation score (treatment group only)
	Mean (p-value)	Mean (p-value)	Mean (p-value)
Truth	0.0	0.20	0.30
No baseline data			
(A1) Demographic characteristics	-0.08 (3)	.	.
(A2) Attitudinal measure added	0.97 (45)	.	.
Add baseline test score as predictor			
(A3) Demographic characteristics	0.95 (100)	.	.
(A4) Attitudinal measure added	1.00 (100)	.	.
(A5) Key variables, no interaction terms	1.00 (100)	.	.
(A6) Key variables with interaction terms	0.00 (100)	0.20 (100)	0.30 (100)
(A7) Key variables as interaction terms only	0.00 (100)	0.20 (100)	0.30 (100)
(A8) Drop separate treatment status	.	0.20 (100)	0.30 (100)

Assuming no other sources of change

Model	Student in treatment group	Self-efficacy score (treatment group only)	Motivation score (treatment group only)
	Mean (p-value)	Mean (p-value)	Mean (p-value)
Truth	0.0	0.20	0.30
No baseline data			
(A1) Demographic characteristics	-0.08 (3)	.	.
(A2) Attitudinal measure added	0.97 (45)	.	.
Add baseline test score as predictor			
(A3) Demographic characteristics	0.95 (100)	.	.
(A4) Attitudinal measure added	1.00 (100)	.	.
(A5) Key variables, no interaction terms	1.00 (100)	.	.
(A6) Key variables with interaction terms	0.00 (100)	0.20 (100)	0.30 (100)
(A7) Key variables as interaction terms only	0.00 (100)	0.20 (100)	0.30 (100)
(A8) Drop separate treatment status	.	0.20 (100)	0.30 (100)

Assuming no other sources of change

Model	Student in treatment group	Self-efficacy score (treatment group only)	Motivation score (treatment group only)
	Mean (p-value)	Mean (p-value)	Mean (p-value)
Truth	0.0	0.20	0.30
No baseline data			
(A1) Demographic characteristics	-0.08 (3)	.	.
(A2) Attitudinal measure added	0.97 (45)	.	.
Add baseline test score as predictor			
(A3) Demographic characteristics	0.95 (100)	.	.
(A4) Attitudinal measure added	1.00 (100)	.	.
(A5) Key variables, no interaction terms	1.00 (100)	.	.
(A6) Key variables with interaction terms	0.00 (100)	0.20 (100)	0.30 (100)
(A7) Key variables as interaction terms only	0.00 (100)	0.20 (100)	0.30 (100)
(A8) Drop separate treatment status	.	0.20 (100)	0.30 (100)

Assuming no other sources of change

Model	Student in treatment group	Self-efficacy score (treatment group only)	Motivation score (treatment group only)
	Mean (p-value)	Mean (p-value)	Mean (p-value)
Truth	0.0	0.20	0.30
No baseline data			
(A1) Demographic characteristics	-0.08 (3)	.	.
(A2) Attitudinal measure added	0.97 (45)	.	.
Add baseline test score as predictor			
(A3) Demographic characteristics	0.95 (100)	.	.
(A4) Attitudinal measure added	1.00 (100)	.	.
(A5) Key variables, no interaction terms	1.00 (100)	.	.
(A6) Key variables with interaction terms	0.00 (100)	0.20 (100)	0.30 (100)
(A7) Key variables as interaction terms only	0.00 (100)	0.20 (100)	0.30 (100)
(A8) Drop separate treatment status	.	0.20 (100)	0.30 (100)

Including actual change over time

Model	Student in treatment group	Self-efficacy score (treatment group only)	Motivation score (treatment group only)
	Mean (% sig)	Mean (% sig)	Mean (% sig)
Truth	0.0	0.20	0.30
No baseline data			
(A1) Demographic characteristics	-0.15 (4)	.	.
(A2) Attitudinal measure added	0.99 (43)	.	.
Add baseline test score as predictor			
(A3) Demographic characteristics	0.83 (89)	.	.
(A4) Attitudinal measure added	1.02 (98)	.	.
(A5) Key variables, no interaction terms	1.02 (97)	.	.
(A6) Key variables with interaction terms	0.14 (6)	0.15 (6)	0.29 (17)
(A7) Key variables as interaction terms only	-1.4 (76)	0.71 (93)	0.45 (57)
(A8) Drop separate treatment status	.	0.47 (70)	0.17 (14)

Including actual change over time

Model	Student in treatment group	Self-efficacy score (treatment group only)	Motivation score (treatment group only)
	Mean (% sig)	Mean (% sig)	Mean (% sig)
Truth	0.0	0.20	0.30
No baseline data			
(A1) Demographic characteristics	-0.15 (4)	.	.
(A2) Attitudinal measure added	0.99 (43)	.	.
Add baseline test score as predictor			
(A3) Demographic characteristics	0.83 (89)	.	.
(A4) Attitudinal measure added	1.02 (98)	.	.
(A5) Key variables, no interaction terms	1.02 (97)	.	.
(A6) Key variables with interaction terms	0.14 (6)	0.15 (6)	0.29 (17)
(A7) Key variables as interaction terms only	-1.4 (76)	0.71 (93)	0.45 (57)
(A8) Drop separate treatment status	.	0.47 (70)	0.17 (14)

Including actual change over time

Model	Student in treatment group	Self-efficacy score (treatment group only)	Motivation score (treatment group only)
	Mean (% sig)	Mean (% sig)	Mean (% sig)
Truth	0.0	0.20	0.30
No baseline data			
(A1) Demographic characteristics	-0.15 (4)	.	.
(A2) Attitudinal measure added	0.99 (43)	.	.
Add baseline test score as predictor			
(A3) Demographic characteristics	0.83 (89)	.	.
(A4) Attitudinal measure added	1.02 (98)	.	.
(A5) Key variables, no interaction terms	1.02 (97)	.	.
(A6) Key variables with interaction terms	0.14 (6)	0.15 (6)	0.29 (17)
(A7) Key variables as interaction terms only	-1.4 (76)	0.71 (93)	0.45 (57)
(A8) Drop separate treatment status	.	0.47 (70)	0.17 (14)

Conclusions

- Even with randomization, bias can occur.
- A simple comparison of means may produce misleading results.
- Simply adding variables that measure demographic differences may be insufficient.
- Ways to handle bias:
  - Have a large number of clusters
  - Get as much data as you can, perhaps especially including pretest scores
  - Correct model specification is important