

Causal generalizations: Building connections between science and policy

Elizabeth Tipton

Associate Professor of Statistics

Co-Director, Statistics for Evidence-Based Policy and Practice (STEPP) Center

Faculty Fellow, Institute for Policy Research

Northwestern University

Keynote at University of Nebraska-Lincoln, November 2021

Questions to begin

My background:

- PhD in Statistics: 2006 – 2011
- IES program Pre-Doctoral Fellow: 2007 – 2010
- Faculty, Teachers College, Columbia University: 2011 – 2018
- Faculty, Northwestern University: 2018 – now

As an education research, a human, and a citizen, my questions are driven by:

How can we do right by all of our children?

How can we make things better?

Education Sciences Reform Act (2002)

Through ESRA, the **Institute of Education Sciences** was formed.

This provided a new model for research in education, including:

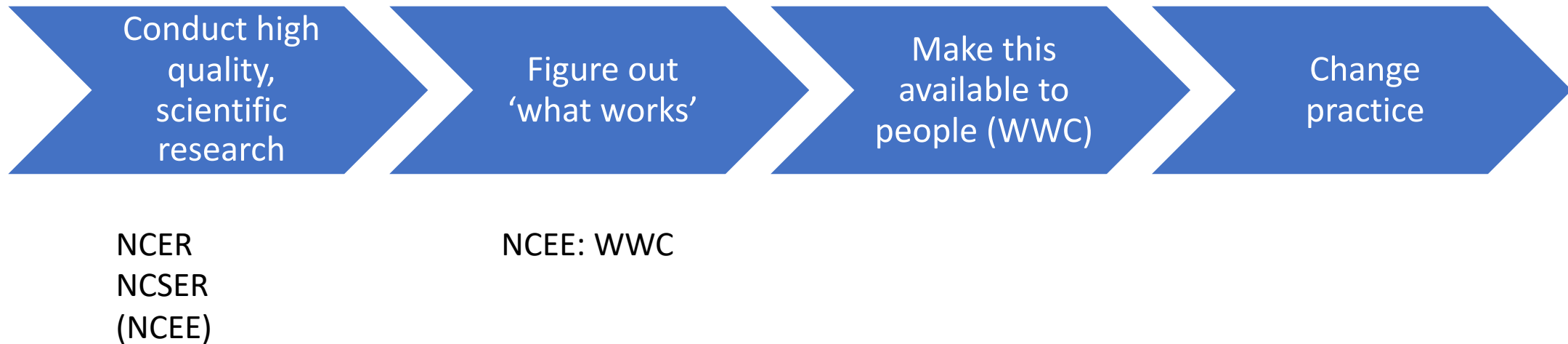
- National Center for Education Research (NCER)
- National Center for Special Education Research (NCSEER)
- National Center for Education Evaluation (NCEE)
- National Center for Education Statistics (NCES)

Mission: National Center for Education Research

- (1) To sponsor sustained research that will lead to the **accumulation of knowledge** and understanding of education,
 - (A) ensure that all children have access to a high- quality education;
 - (B) improve student academic achievement, including through the use of educational technology;
 - (C) close the achievement gap between high-performing and low-performing students through the improvement of teaching and learning of reading, writing, mathematics, science, and other academic subjects; and
 - (D) improve access to, and opportunity for, postsecondary education;
- (2) To **support the synthesis** and, as appropriate, the integration of education research;
- (3) To promote **quality and integrity** through the use of accepted practices of scientific inquiry to obtain knowledge and understanding of the validity of education theories, practices, or conditions; and
- (4) To promote **scientifically valid research findings** that can provide the basis for improving academic instruction and lifelong learning.

The underlying model

Undergirding this was a framework for change.



My area, as statistician

The term “scientifically valid education evaluation” means an evaluation that—

- (A) adheres to the highest possible standards of quality with respect to research design and statistical analysis;**
- (B) provides an adequate description of the programs evaluated and, to the extent possible, examines the relationship between program implementation and program impacts;
- (C) provides an analysis of the results achieved by the program with respect to its projected effects;
- (D) employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible; and**
- (E) may study program implementation through a combination of scientifically valid and reliable methods.

The clearinghouse model

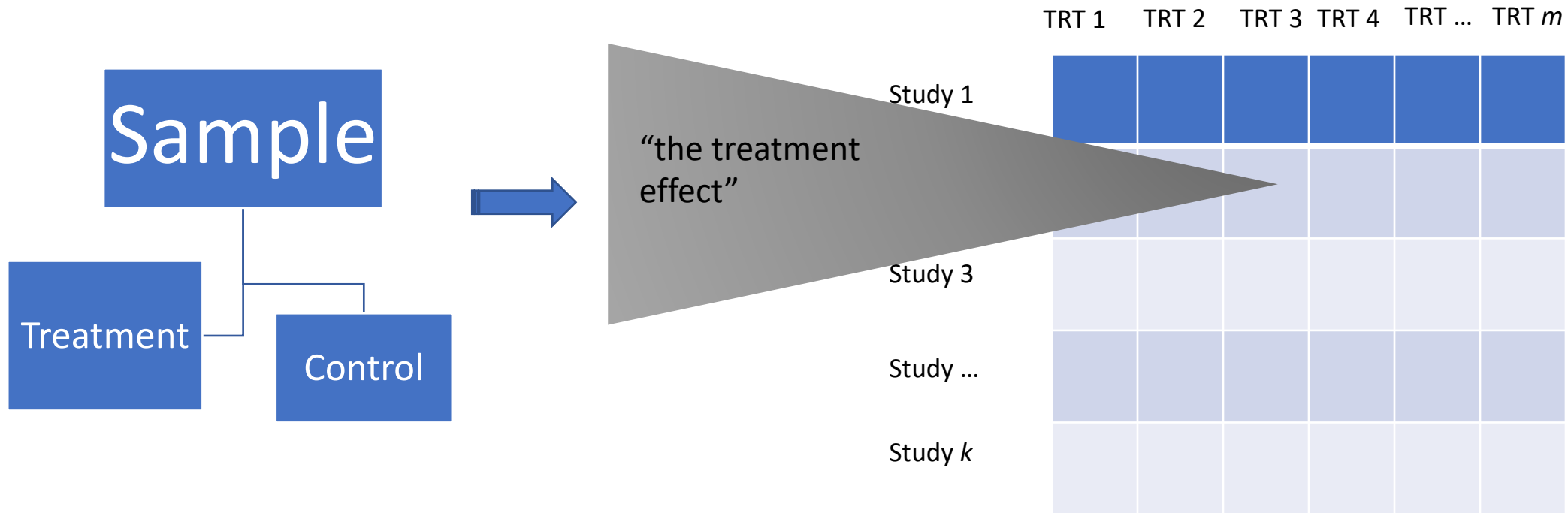
The end-goal of this new IES research = the WWC would provide a database.

Each cell would provide a treatment effect estimate from a high-quality study.

Teachers and principals would search for the column they needed, see the evidence, and use this to make a decision.

	TRT 1	TRT 2	TRT 3	TRT 4	TRT ...	TRT m
Study 1						
Study 2						
Study 3						
Study ...						
Study k						

We just have to fill in the boxes



2002-2012: a lot of new statistical work

- Methods for generating evidence
 - Designs: Cluster-randomization, Multi-site randomization
 - Power analysis:
 - Non-centrality parameters
 - Design parameter values: ICCs, R^2 , effect sizes
 - Extensions: Differential Attrition, Measurement error, Baseline adjustments
- Methods for synthesizing evidence
 - Categories: Meets Standards (With, Without Reservations), DNMS
 - Rules for each: WWC Standards Guide

And now, 20 years later

We have a lot of evidence.

1. The WWC now includes over 10,000 studies.

2. IES has funded over 400 efficacy and effectiveness studies.

	TRT 1	TRT 2	TRT 3	TRT 4	TRT ...
Study 1					
Study 2					
Study 3					
Study ...					
Study k					

Are we there yet?

With this many studies, surely the WWC is nearly complete?

20 years in, we've learned a few things:

1. There isn't a single treatment effect to put in each cell.
 - Treatment effects vary. For a lot of reasons.
2. Synthesizing evidence is tricky.
 - What models do we use? Do we vote count? Do we meta-analyze?
3. Decision-makers don't turn to the WWC as much as we thought they would.
 - Now what?

1. Treatment effects vary

“The” treatment effect

In the original IES model, each cell provided an estimate of “the” treatment effect.

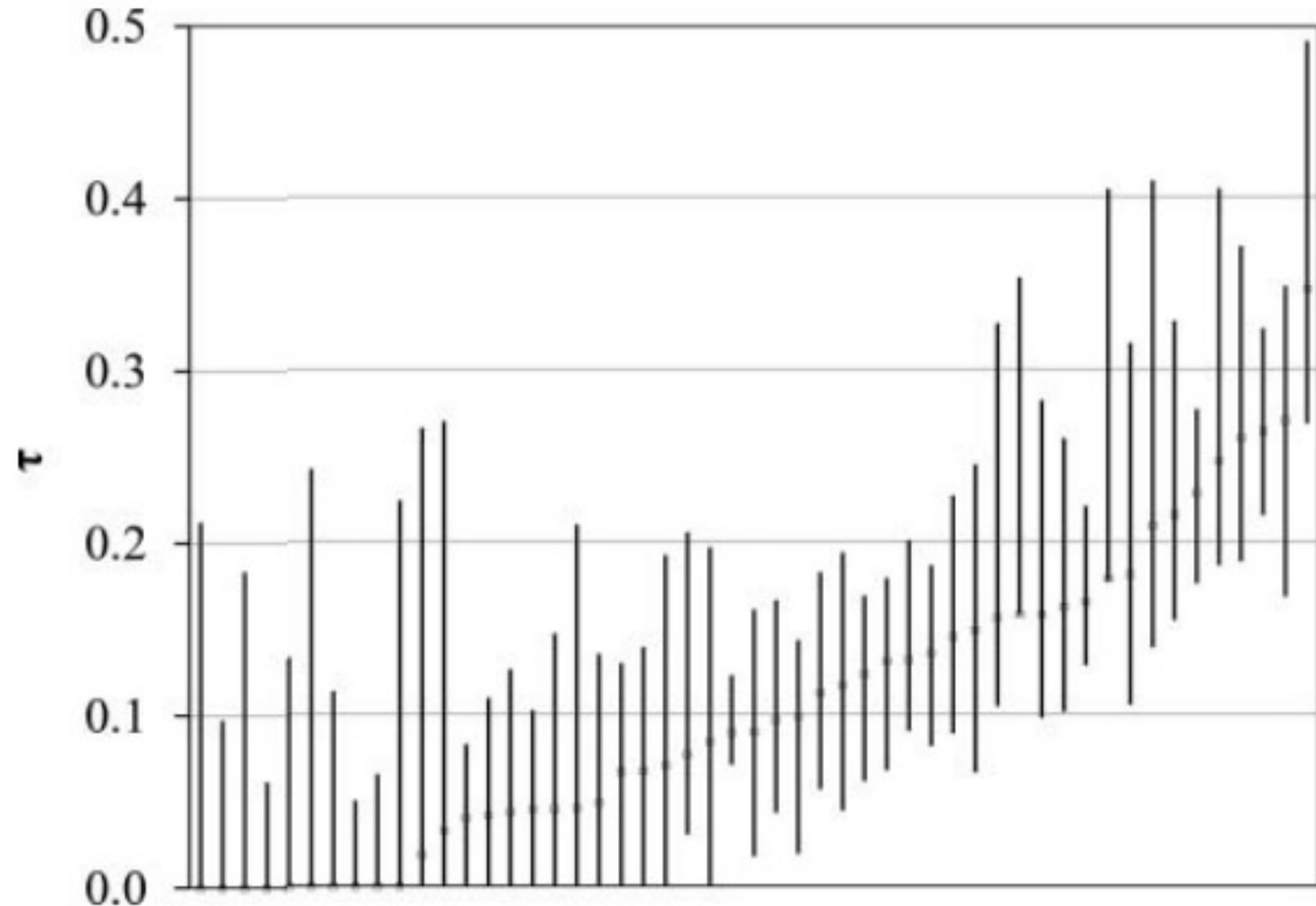
This idea can be found, too, in the language of ‘replication’:

- “Does it replicate?” makes sense only if there is one true effect

But this is an “average” treatment effect:

- It averages over individual treatment effects (which we cannot observe)
- These individual treatment effects include parts we can explain and parts we can't, parts having to do with students, teachers, and contexts, and so on.
- This average does not on its own communicate the variation in treatment effects (only certain designs make this possible to estimate)

Variation in effects



Michael J. Weiss, Howard S. Bloom, Natalya Verbitsky-Savitz, Himani Gupta, Alma E. Vigil & Daniel N. Cullinan (2017) How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials, *Journal of Research on Educational Effectiveness*, 10:4, 843-876, DOI: 10.1080/19345747.2017.1300719

A new assumption

There is enough evidence now to suggest that we should begin our trials by assuming that *treatment effects vary*.

This changes everything.

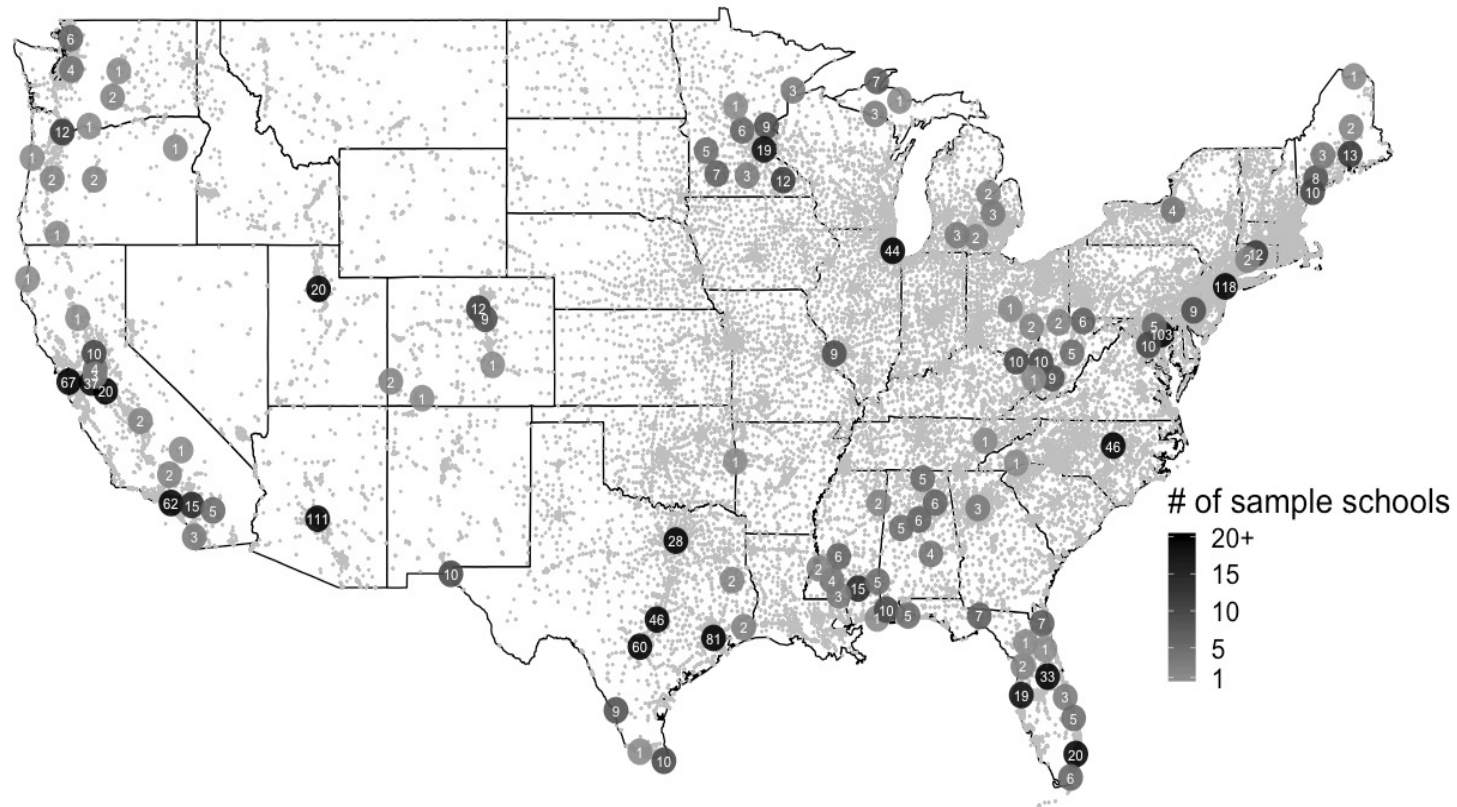


Layer 1. What samples are we including in our studies? What populations do these come from?

This is a question about current, baseline practice. What are we doing already in education studies?

Samples differ from target populations

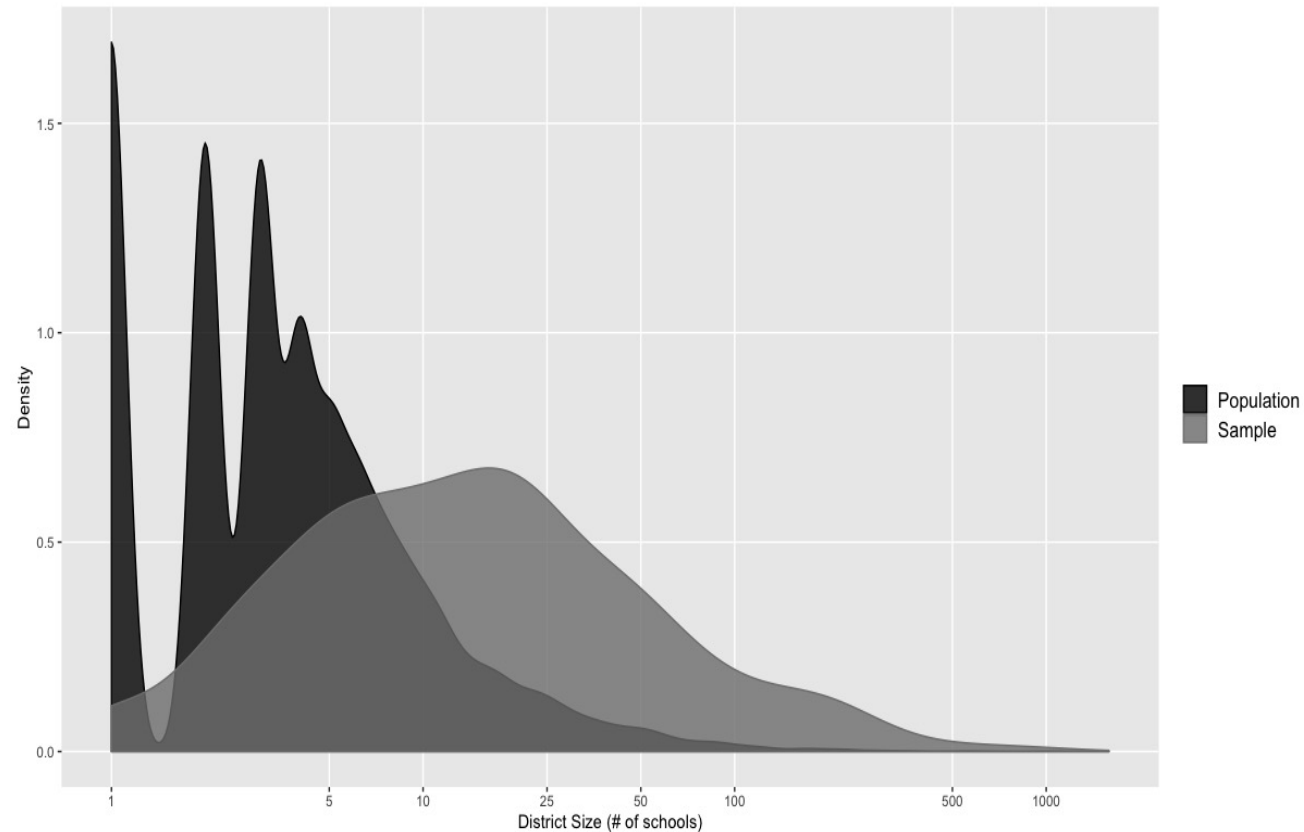
Locations of schools in 34 RCTs funded by IES between 2011-2015



Tipton, Spybrook, Fitzgerald, Zhang, & Davidson (2020)

We often include 'easier' to recruit sites

Size of school districts in 34 RCTs funded by IES between 2011-2015



Researchers prefer large school districts.

Large districts tend to bring with them more schools.

They are more often urban.

They have very different resources and students.

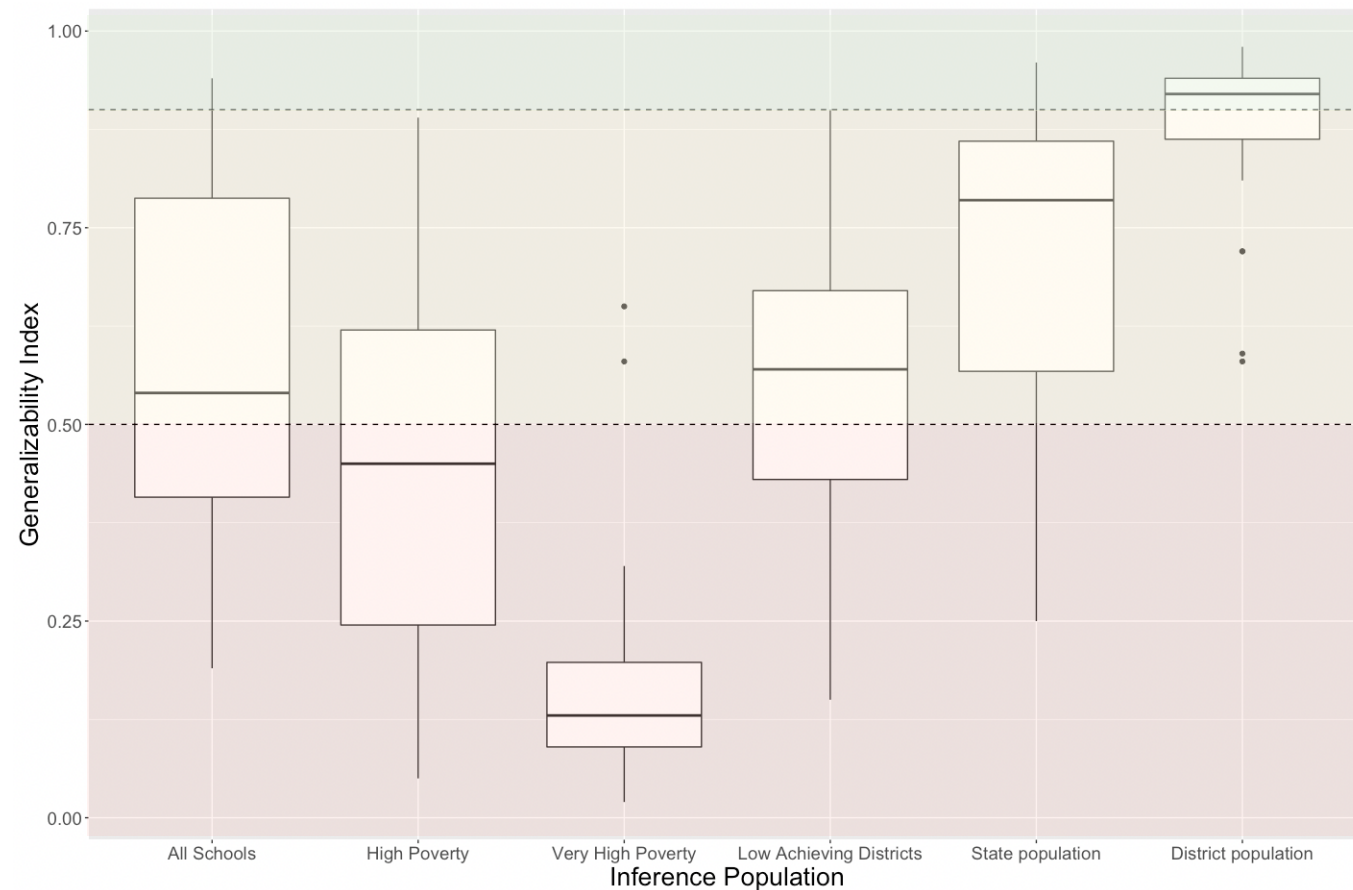
Layer 2. Are these the right ones? Can we use the data we have to estimate the ATE for the ‘right’ population?

In some cases, we can adjust ATE estimates based upon population information, using :

- Propensity score post-stratification¹²
- Propensity score inverse probability weighting³
- Maximum entropy weighting⁴
- Use of bounding approaches⁵

Layer 3. What is the 'right' population?

Comparisons of study samples to 6 populations in each of the 34 RCTs funded by IES between 2011-2015

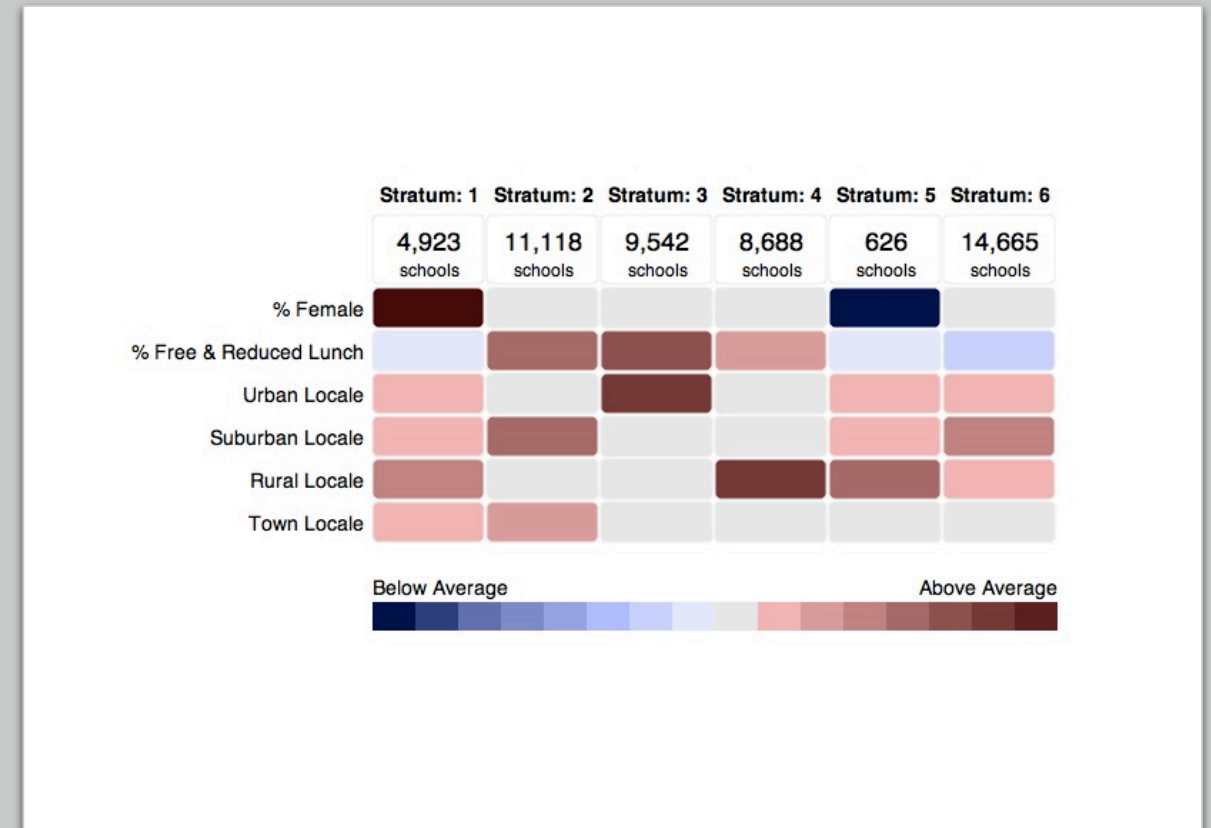
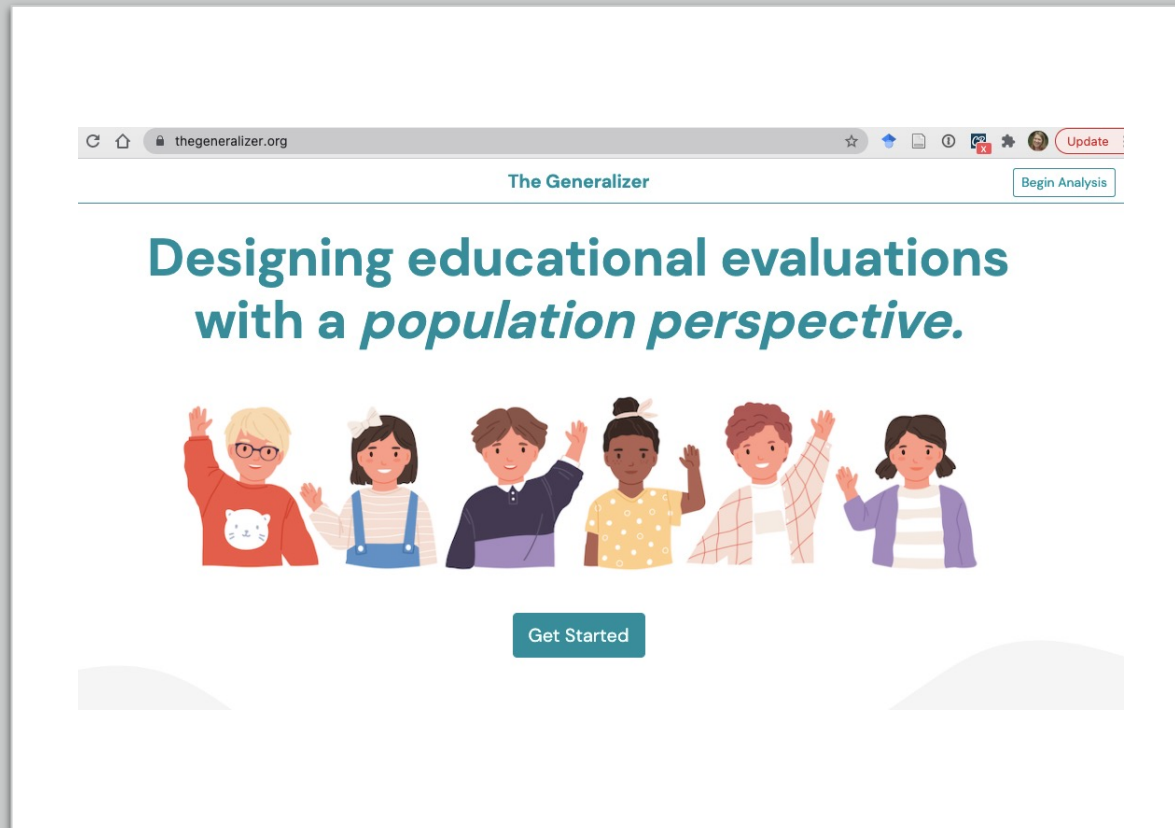


1. Studies did not report clearly what their target populations were.

2. Most studies did **not** represent policy populations well.

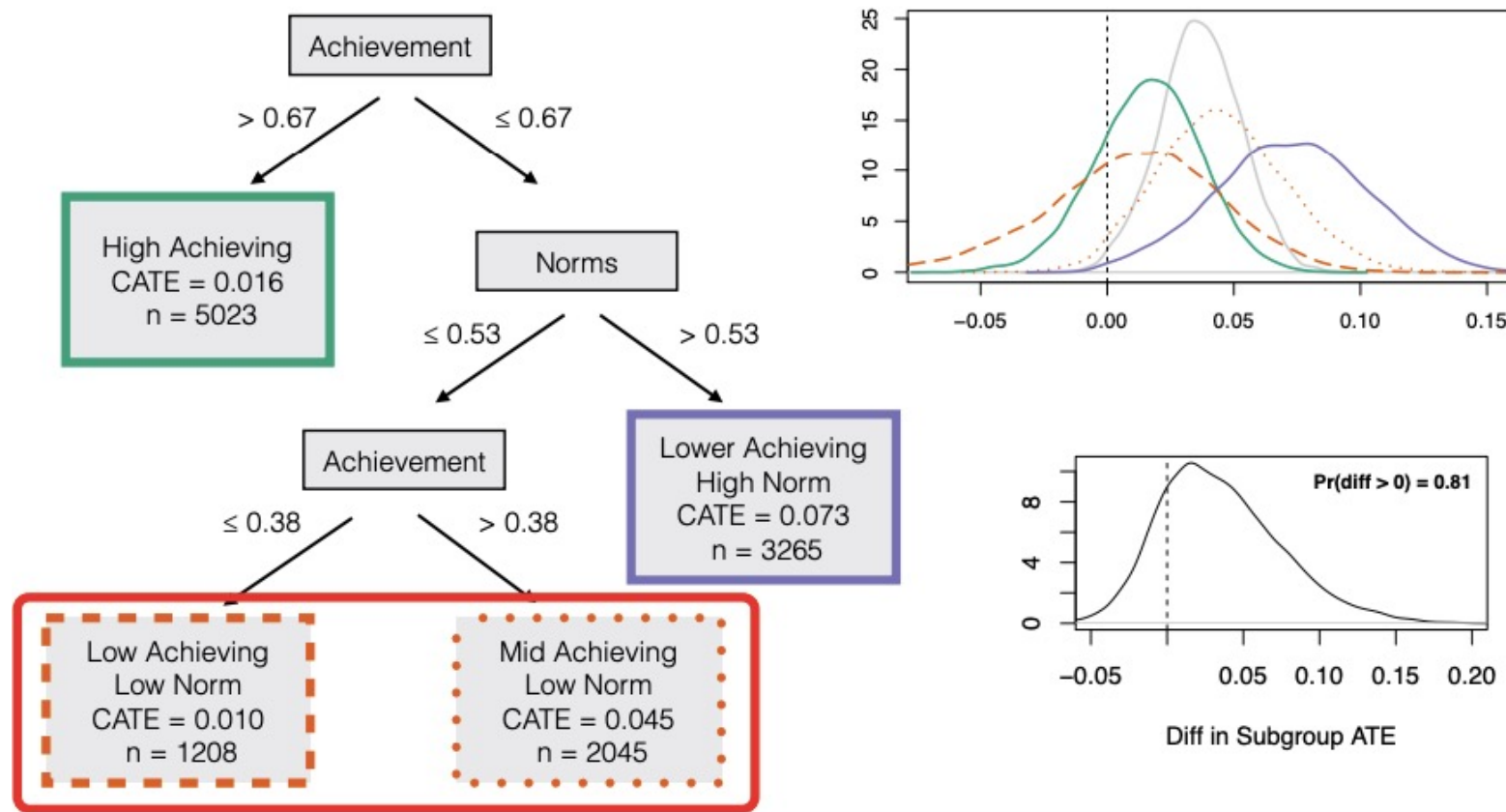
3. They did best at representing the districts they were in.

Layer 4. Maybe this would be easier if we recruited samples from the 'right' population?



- 1. Tipton (2014); 2. Tipton et al., 2014; 3. Tipton & Miller (2015)

Layer 5. If they vary, don't we want to know why / how?

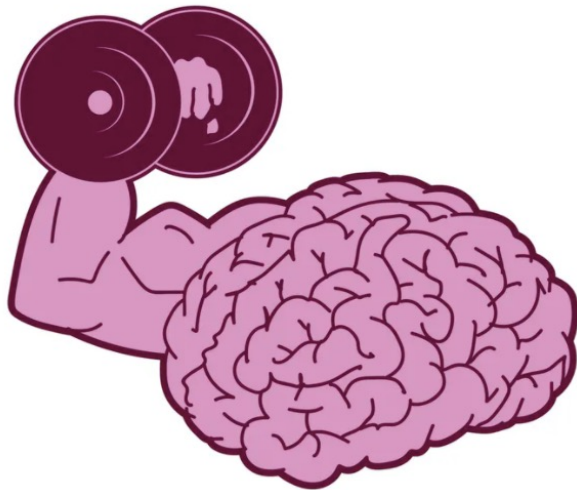


Layer 6. Shouldn't we design studies to do understand variation?

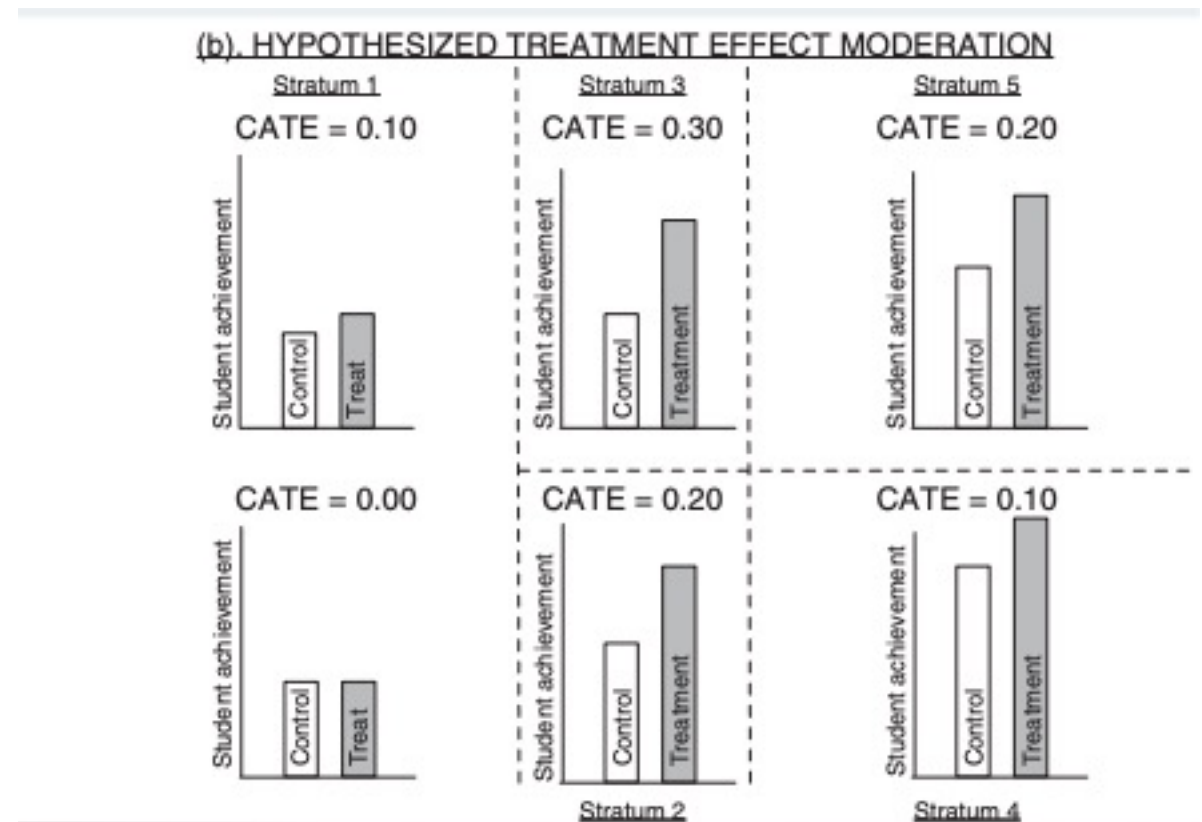
EDUCATION RESEARCH

New Study Shows Where 'Growth Mindset' Training Works (And Where It Doesn't)

By Jeffrey R. Young Aug 7, 2019

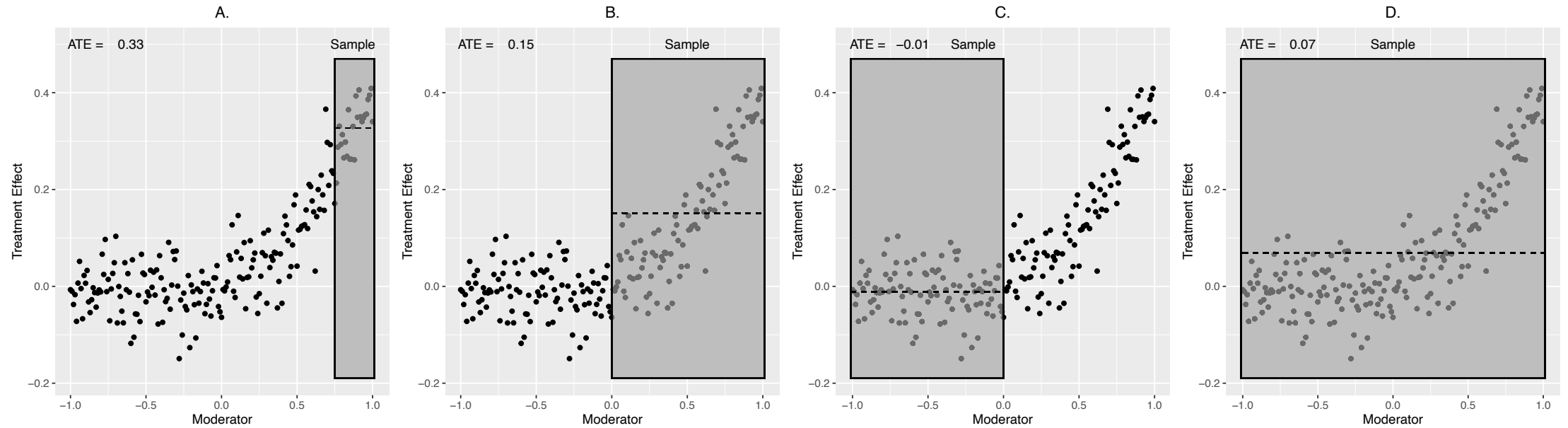


Yeager et al, 2018. *Nature*.



Tipton, Yeager, Schneider, Iachan (2019)

Layer 7. We need to think big.



Bryan, Tipton, Yeager (2021) *Nature Human Behavior*.

2. How to determine 'what works'

We have data, now what?

How do we categorize evidence? [orange column]

For example, the WWC rates findings (cells) as:

- *Meets Standards (without reservations)*
- *Meets Standards (with reservations)*
- *Does not meet standards*

	TRT 1	TRT 2	TRT 3	TRT 4	TRT ...
Study 1					
Study 2					
Study 3					
Study ...					
Study k					

Reviewed Research















Early Childhood
Education

July 2010

 EVIDENCE SNAPSHOT

 INTERVENTION REPORT (587 KB)

 REVIEW PROTOCOL

Outcome domain 	Effectiveness rating 	Studies meeting standards 	Grades examined 	Students 	Improvement index 
Cognition	--  --++	<u>1 study meets standards</u>	PK	722	--
General Mathematics Achievement	--  --++	<u>1 study meets standards</u>	PK	185	--
Oral language	--  ++	<u>3 studies meet standards</u>	PK	1,002	
Phonological processing	--  ++	<u>3 studies meet standards</u>	PK	1,004	
Print knowledge	--  ++	<u>3 studies meet standards</u>	PK	999	

We have data, now what?

Options include:

1. Count how many studies have a statistically significant effect?
 - Categorize how?
2. Combine the effect sizes using meta-analysis.
 - Which model? Fixed? Random?
 - Categorize how?
3. What if treatment effects vary within studies? What then?

Case study

Oral language	<div><div>--</div><div>-</div><div>0</div><div>+</div><div>++</div></div>	<u>3 studies meet standards</u>	PK	1,002	
		Farver, J. A. M., Lonigan, C. J., & Eppe, S. (2009)	PK	94	
		Lonigan, C. J., Farver, J. M., Clancy-Menchetti, J., & Phillips, B. M. (2005, April)	PK	722	
		Preschool Curriculum Evaluation Research (PCER) Consortium. (2008)	PK	186	--

Vote count

Should we count how many studies have a statistically significant effect?

This is 'vote-counting' – a practice we know has poor properties (from meta-analysis research).

Why? Type II errors.

- Imagine we take one large, well powered study with a statistically significant average treatment effect.
- If we randomly divide this into many small studies, eventually even though the treatment effect is the same, none of the studies will be statistically significant.

Meta-analytic approach

Should we combine the effect sizes using meta-analysis?

Yes! But using which model?

- Are the studies all replicates of the same study? Great – let's use a '**fixed effects**' model that assumes they are all estimating the same average treatment effect.

Open questions

But they probably aren't all estimating the same thing – we know treatment effects vary.

Let's use a '**random effects**' model instead, that assumes each study has a different average treatment effect.

- But can we estimate this model well? We only have < 5 studies?
[Probably not]
- If we can estimate it, how do we summarize the result? Do we focus on the average? Or an interval?

Is there another approach?

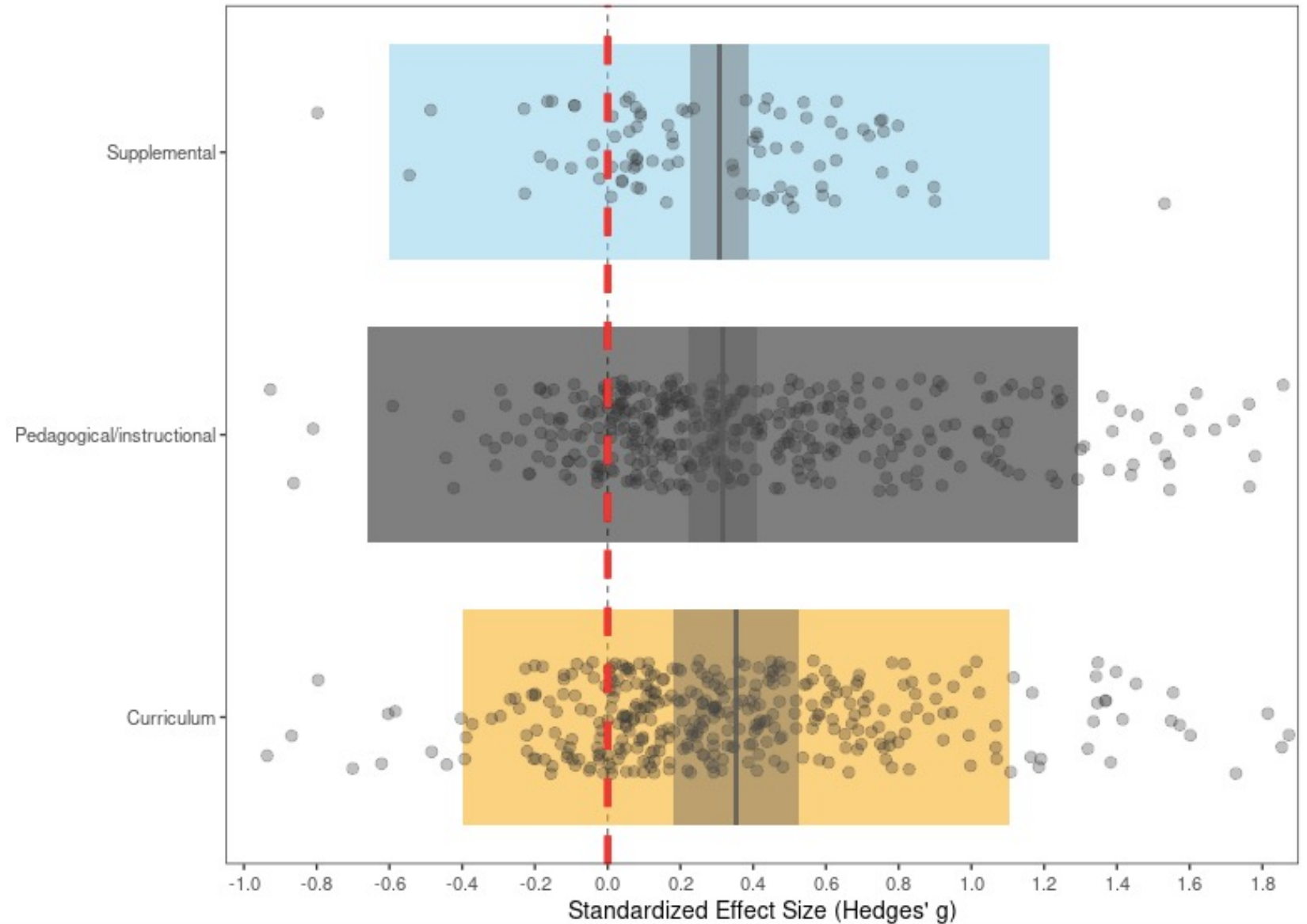
Could we look across several columns at once using meta-analysis?

Yes! We can.

- This requires careful modeling of the data.
- It allows us to examine some moderators.

	TRT 1	TRT 2	TRT 3	TRT 4	TRT ...
Study 1					
Study 2					
Study 3					
Study ...					
Study k					

Example



Citkowicz, Williams, and Lindsay. See here: <https://www.air.org/centers/mosaic/mosaic-db>

Large meta-analyses are different

These aren't your "Meta Analysis 101" studies.

They require:

- More complex models
- A focus on exploring heterogeneity, not simply reporting the average
- Attempts at explaining the heterogeneity using moderators
- Controlling for methodological confounds
- Dealing with missing data
- Careful approaches to conveying findings

3. We built it, but they didn't come?

Have we changed practice?

A survey suggests that very few decision makers use the WWC:

- 18% of District officials use it a lot
- 56% have *never* used it

Why ? Or Why not?

Is the system:	For the decision encountered:
Usable?	It is possible to determine if there is evidence in the system.
Available?	There is evidence in the system that is relevant.
Accurate?	The evidence provided is statistically sound (unbiased, precise).
Interpretable?	The evidence is appropriately interpreted by the decision-maker.
Useful?	An appropriate intervention is selected, based on the evidence.
Helpful?	The selected intervention improves the individual outcome.

Available? Useful?

- **Wrong questions:**

- The *interventions* studied are not what districts are looking for evidence regarding
- The interventions they want to know about aren't studied or included
- Districts don't want packaged interventions, they want concepts, theories, and principles

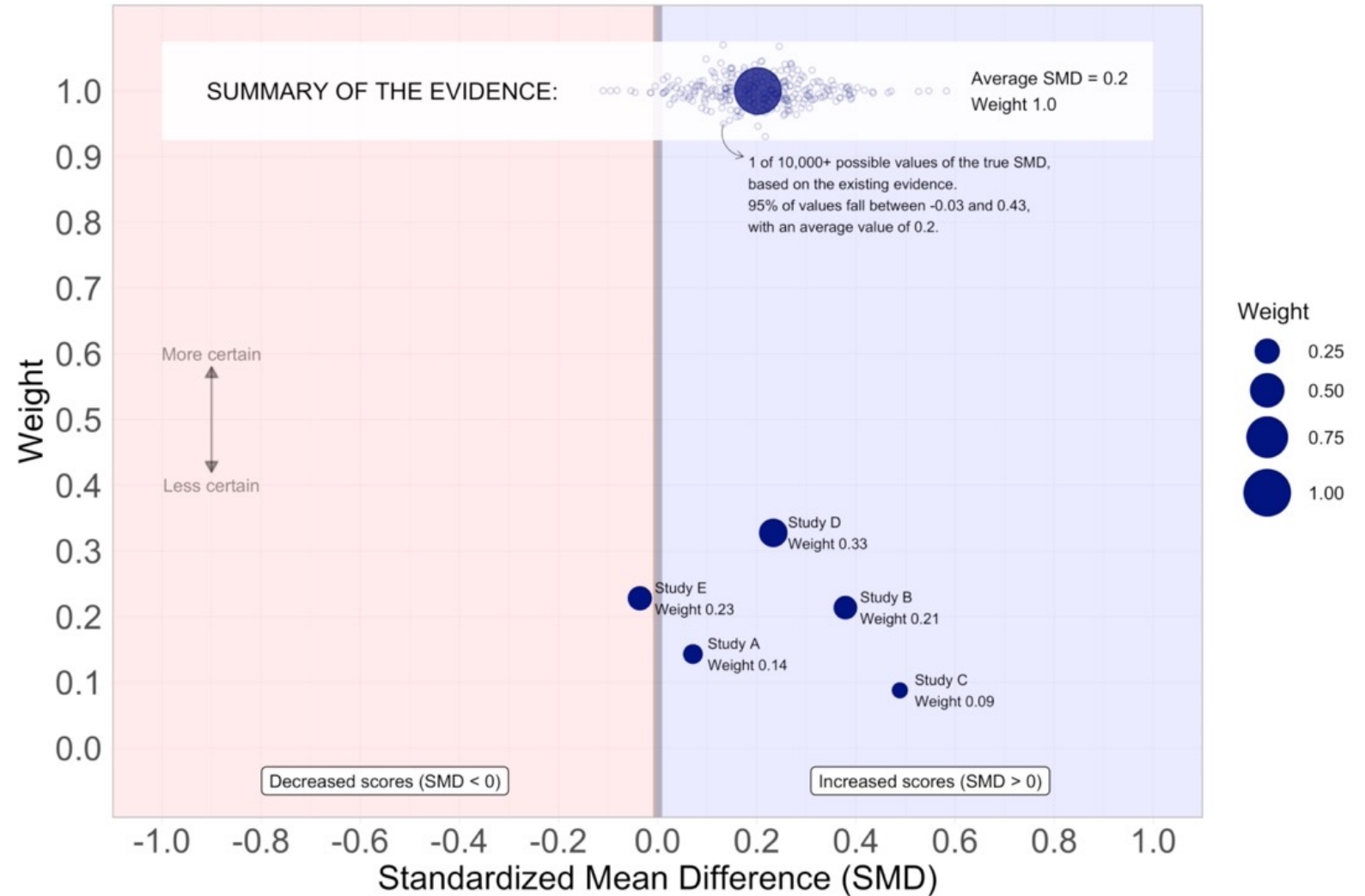
- **Context matters:**

- The scientific focus is on internal validity, but users care about external validity
- The ratings / categories/ interface don't take this into account

Usable? Interpretable?

- **Do users even know where to look?**
 - Where are they looking for evidence when decision making?
 - Why aren't they looking in the WWC (or other places)?
- **Are they able to clearly interpret the results?**
 - Do they 'receive' the message that the WWC intends to send?
 - Are these categories meaningful to them?
 - Are they interpretable?

Translation
can be
studied



This figure displays SMD estimates from 5 studies and a corresponding weighted average that summarizes the evidence. Other plausible values for the true SMD - simulated based on the existing evidence and uncertainty - are also displayed.

SMD > 0 indicates that the new curriculum increased student scores, SMD < 0 indicates it decreased scores, and SMD = 0 indicates it had no effect on scores.

Fitzgerald & Tipton, in press.

What do we need?

As researchers, we know that if we value something, we study it.

What do we need to study?

- We need indicators and measures of use.
- Decision-making processes
- Best practices for translation
- And so on...

We can't just relegate this to anecdotes and dissemination plans.

Concluding thoughts

Take-aways

20 years ago, the promise of IES was that we'd figure out what works, provide it to decision-makers, and improve education.

We've made a lot of progress.

But it's more complicated than we thought.

1. Treatment effects vary. This changes everything.
2. Categorizing evidence is tricky. We need to be careful.
3. Simply providing evidence isn't enough to change the system.

Thank you!

Elizabeth Tipton

tipton@northwestern.edu

<https://steppcenter.northwestern.edu>

@stats_tipton

Abstract

In science, we are often interested in knowing if an intervention or treatment 'causes' an outcome to change. Teasing apart causality requires the use of research designs that have high internal validity - e.g., randomized experiments or strong quasi-experiments. Outside of basic science, the results of these studies are often intended to inform policies and practice for individuals and organizations. This use calls to question the external validity of these designs.

In this talk, I reflect on my work as a statistician developing methods to improve the external validity of these high internal validity designs. This includes work on the design and analysis of individual field trials, as well as the collection of evidence across trials using meta-analysis.