The logic of generalization:
From systematic reviews and meta-analyses to diverse policy and practice contexts

Julia H. Littell, PhD, Professor Emerita
Graduate School of Social Work and Social Research
Bryn Mawr College, Bryn Mawr, PA, USA

AERA SRMA SIG online seminar
20 October 2023

Introduction

- **Central questions**: Can we use results of systematic reviews and meta-analyses (SRMAs) to make inferences about wider policy/practice contexts?
- Focus on SRMAs of research on *intervention effects*
- **Generalizability**, external validity: extrapolation beyond the data at hand (Shadish, Cook, & Campbell, 2002)
  - To what Populations/Problems, Interventions, Comparisons, Outcomes, Times, and Settings (PICOTS) are results of our SRMA likely to apply?
- **Applicability**: relevance for specific target context(s)
  - From a policymaker's or practitioner's viewpoint: What are the likely effects of this intervention in my context (with my PICOTS)?

Generalization: From one set of multi-attribute contexts to others


Campbell Collaboration

Plain Language Summary
Social Welfare

2021

Effects of Multisystemic Therapy® are inconsistent within and across studies

Twenty-three randomised controlled trials provide evidence of effects of Multisystemic Therapy®

UPDATED SYSTEMATIC REVIEW

Campbell Collaboration WILEY

Multisystemic Therapy® for social, emotional, and behavioural problems in youth age 10 to 17: An updated systematic review and meta-analysis

Julia H. Littell[1] | Therese D. Pigott[2] | Karianne H. Nilsen[3] | Stacy J. Green[4] | Olga L. K. Montgomery[5]

Abstract

Background: Multisystemic Therapy® (MST®) is an intensive, home-based intervention for families of youth with social, emotional, and behavioural problems. MST therapists engage family members in identifying and changing individual, family, and environmental factors thought to contribute to problem behaviour. Intervention may include efforts to improve communication, parenting skills, peer relations, school performance, and social networks. MST is widely considered to be a well-established, evidence-based programme.

Objectives: We assessed (1) impacts of MST on out-of-home placements, crime and delinquency, and other behavioural and psychosocial outcomes for youth and families; (2) consistency of effects across studies; and (3) potential moderators of effects including study location, evaluator independence, and risks of bias.

Search Methods: Searches were performed in 2003, 2010, and March to April 2020. We searched PsycINFO, MEDLINE, ERIC, NCJRS Abstracts, ProQuest and WorldCAT dissertations and theses, and 10 other databases, along with government and professional websites. Reference lists of included articles and research reviews were examined. Between April and August 2020 we contacted 22 experts in search of missing data on 16 MST trials.

Selection Criteria: Eligible studies included youth (ages 10 to 17) with social, emotional, and/or behavioural problems who were randomly assigned to licensed MST programmes or other conditions. There were no restrictions on publication status, language, or geographic location.

Data Collection and Analysis: Two reviewers independently screened 1802 titles and abstracts, read all available study reports, assessed study eligibility, and extracted data onto structured electronic forms. We assessed risks of bias (ROB) using modified versions of the Cochrane ROB tool and What Works Clearinghouse standards.

?

Study designs, participants, interventions, comparisons, outcome measures, endpoints

Generalizing from… versus applying to…

Generalizing from SRMA results …

Which PICOTS?

Which (if any) SRMA results?

… apply to specific context(s)

Known PICOTS

Generalizability assessment

- How do/can we assess the generalizability/applicability of results of a SRMA?
  - Is generalizability assessment a thing?
- Parallel to **evaluability assessment,** a process used to determine whether/when interventions are ready for rigorous outcome/impact evaluation (Rossi et al., 2019).
    - Pre-requisites:
      - Logic model or theory of change
      - Descriptive data on participants, intervention processes, outcomes
      - Running smoothly for at least one year
      - “Proud”

Generalizability assessment - 2

- Developing generalizability assessments -- by muddling through
  - **Two case studies**: SRMAs of effects of two “evidence-based” programs
  - **Three frameworks for generalization**
    1. Probability theory and sampling methods
    2. Principles of generalized causal inference (Shadish, Cook, & Campbell, 2002)
    3. Common rubrics and rhetoric of generalization

Two case studies: SRMAs of effects of…

- Multisystemic Therapy (MST) (Littell, Pigott, et al., 2021)
- Functional Family Therapy (FFT) (Littell, Pigott, et al., 2023)
- Prominent, “evidence-based” psychosocial treatments
  - For families of youth with social, emotional, and/or behavioral (SEB) problems
  - Short-term (3-6 months), home- and community-based treatment
  - Use techniques from various cognitive-behavioral and family therapies
  - Involve social networks and social service systems
  - “Branded interventions” require training and licensing by companies (LLCs) founded by program developers
  - Strong assumptions about “proven effectiveness” and generalizability

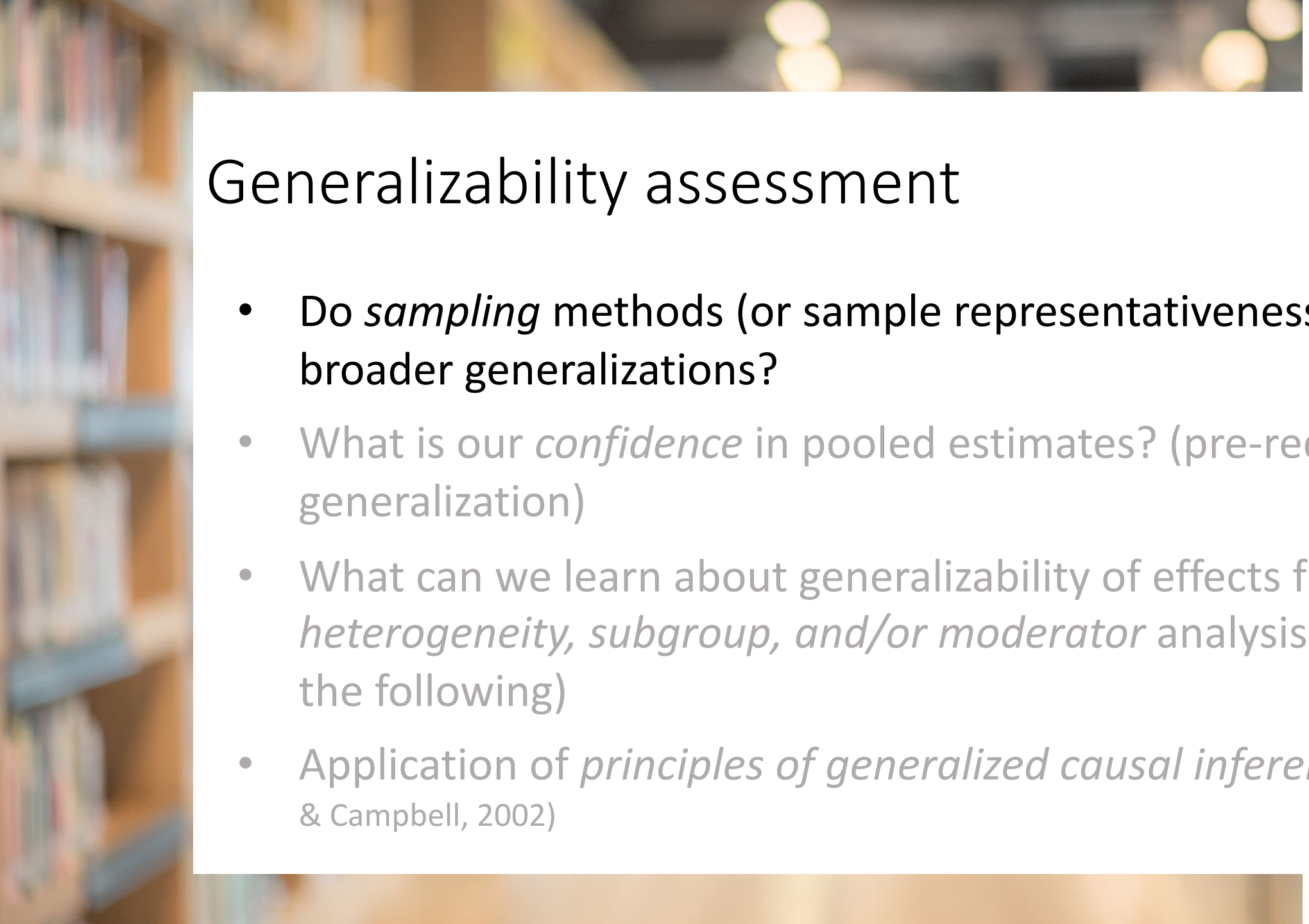
Generalizability claims

Multisystemic Therapy (MST)

- Effectiveness of MST has been demonstrated "across **problems**, **therapists**, and **settings**… [showing] that the treatment and methods of decision making can be extended and that treatment effects are reliable" (Kazdin & Weisz, 1998, p. 28).
- "MST is superior in reducing [**outcomes**] delinquency, drug use, and emotional and behavioral problems and increasing school attendance and family functioning, in **comparison to [a variety of] other procedures**, including 'usual services,'…individual counseling, and community-based eclectic treatment" (Kazdin, 2015, p. 150).

Functional Family Therapy (FFT)

- FFT outcome studies demonstrate effectiveness "with a wide variety of adolescent related **problems** including youth violence, drug abuse, and other delinquency related behaviors. The positive outcomes of FFT remain relatively stable [over **time**] even after a five-year follow-up" (Sexton & Turner, 2010, p. 339).
- FFT is said to be effective across presenting problems, **populations** (gender, race/ethnicity), and **outcome** measures (Robbins, Alexander, Turner, & Hollimon, 2016).
- Kazdin (1998) claimed that FFT is more effective than "**various control conditions**" including family groups, youth groups, family therapy, and no treatment controls.



Generalizability assessment

- Do *sampling* methods (or sample representativeness) support broader generalizations?
- What is our *confidence* in pooled estimates? (pre-requisite for generalization)
- What can we learn about generalizability of effects from *heterogeneity, subgroup, and/or moderator* analysis? (to inform the following)
- Application of *principles of generalized causal inference* (Shadish, Cook, & Campbell, 2002)

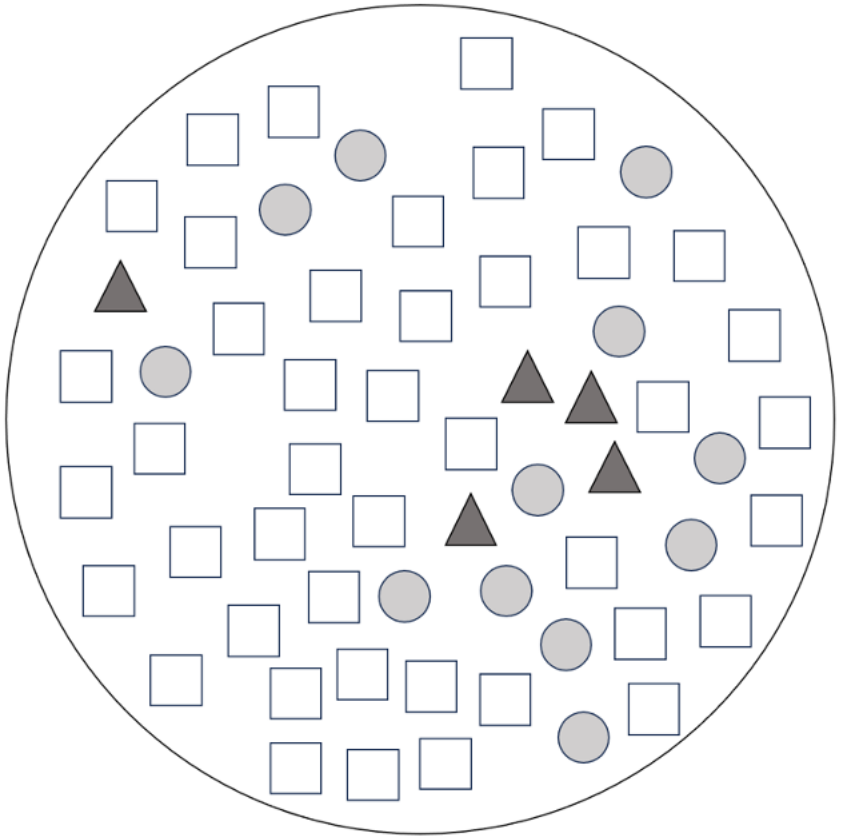
Probability theory and sampling methods

- Probability samples are the “gold standard” for generalizing from sample data to a larger population
  - Support use of inferential statistics to estimate population parameters.
- Types of **studies** often included in SRMAs of intervention effects—i.e., randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) **rarely use probability samples**.
- **SRMAs themselves do not use probability samples of studies** from some larger population of studies.

Sampling problem

Consider a hypothetical universe of all relevant interventions

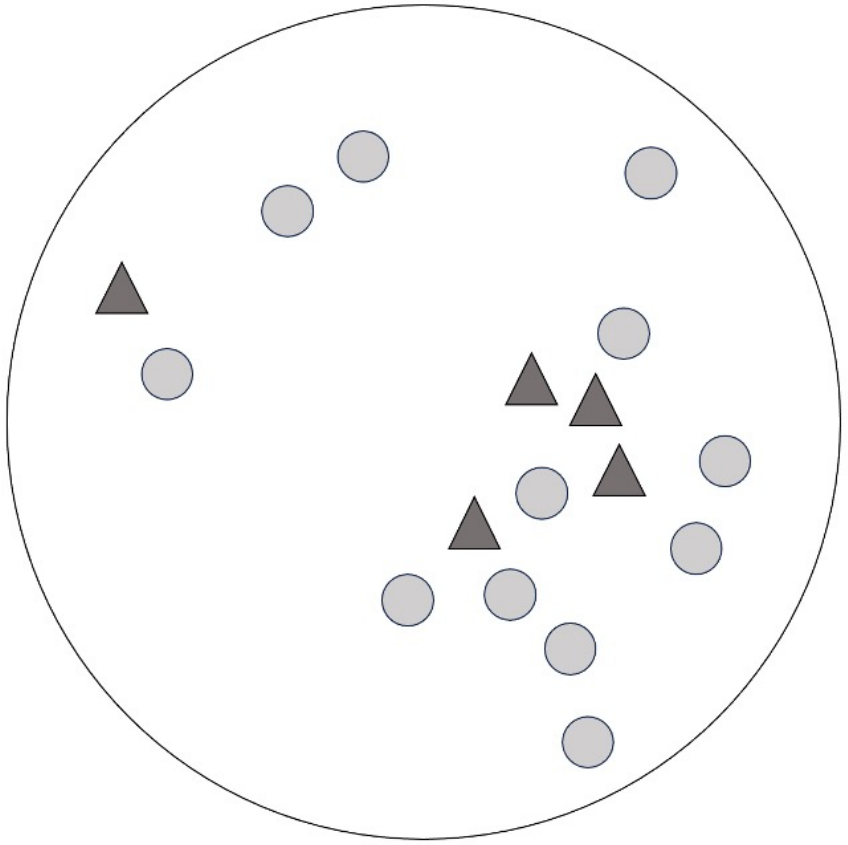
- □ Many have no impact evaluations
- ○ Some are evaluated with QEDs
- △ Few are evaluated with RCTs



Sample of studies for SRMA

A sample of available impact evaluations (RCTs and QEDs)

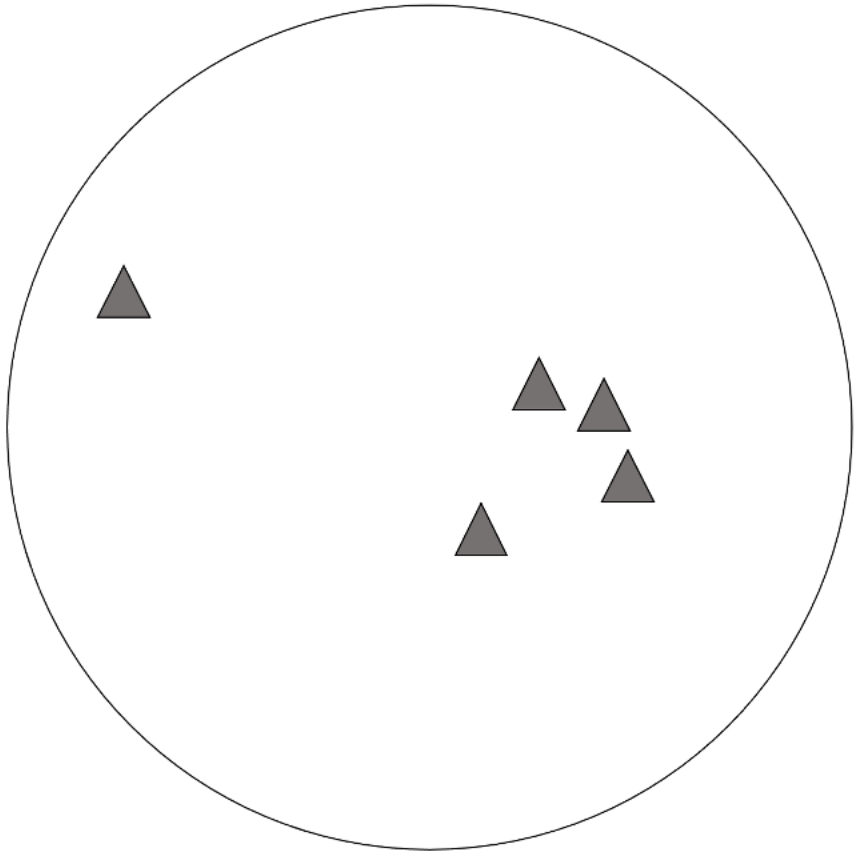
Programs that have been evaluated are not a representative sample of all relevant programs.



Sample of studies for SRMA

A sample of available RCTs.

Programs evaluated with RCTs are not representative of all programs that have been evaluated.



Licensed MST programs and completed trials

- 558 MST programs in 15 countries
- 23 MST trials in 6 countries
- Most in the USA
- Some countries not represented at all

Country/State	Licensed MST programs [a]		MST trials [b]	
	N	Column %	k	Column %
Australia	17	3.0		
Belgium	1	0.2		
Canada	2	0.4	1	4.3
Chile	1	0.2		
England	6	1.1	3	13.0
France	1	0.2		
Germany	2	0.4		
Iceland	3	0.5		
Ireland	3	0.5		
Netherlands	41	7.3	1	4.3
New Zealand	8	1.4		
Norway	24	4.3	1	4.3
Sweden	6	1.1	1	4.3
Switzerland	3	0.5		
USA	440	78.9	16	69.6
Total	558		23	

(Continued)

Licensed MST programs and completed trials: USA

- 440 MST programs in 34 U.S. states
- 16 MST trials in 7 states

USA states	N	Column %	k	Column %
Delaware	2	0.5	1	6.3
Hawaii	6	1.4	1	6.3
Illinois	6	1.4	1	6.3
Missouri	1	0.2	4	25.0
Ohio	11	2.5	1	6.3
South Carolina	3	0.7	6	37.6
Tennessee	8	1.8	2	12.5
27 other states	403	91.6	0	
Remaining 16 states	0		NA	
Subtotal	440		16	

[a] Licensed MST organizations not including adaptations of MST which target different populations and/or include services other than MST (49 programs). Accessed August 27, 2023 at: https://www.mstservices.com/licensed-organizations

[b] Randomized controlled trials of licensed MST programs for social, emotional, and behavioral problems among youth ages 10-17, as of April 2020 (Littell et al., 2021).

Licensed MST programs and completed trials: USA

2 states contain

- < 1% of MST programs in USA
- 63% of MST trials in USA
- 43% of all MST trials in the world

USA states	N	Column %	k	Column %
Delaware	2	0.5	1	6.3
Hawaii	6	1.4	1	6.3
Illinois	6	1.4	1	6.3
Missouri	1	0.2	4	25.0
Ohio	11	2.5	1	6.3
South Carolina	3	0.7	6	37.6
Tennessee	8	1.8	2	12.5
27 other states	403	91.6	0	
Remaining 16 states	0		NA	
Subtotal	440		16	

[a] Licensed MST organizations not including adaptations of MST which target different populations and/or include services other than MST (49 programs). Accessed August 27, 2023 at: https://www.mstservices.com/licensed-organizations

[b] Randomized controlled trials of licensed MST programs for social, emotional, and behavioral problems among youth ages 10-17, as of April 2020 (Littell et al., 2021).

Licensed FFT programs and completed trials

- 334 FFT programs in 12 countries
- 20 FFT evaluations in 6 countries

- Most in the USA:
- 275 programs in 38 states + DC
- 15 evaluations in 8 states

Country/State	Licensed FFT programs [a]		FFT RCTs/QEDs [b]	
	N	Column %	k	Column %
Australia	21	6.3		
Canada	2	0.6		
Denmark	8	2.4		
England	9	2.7	1	5.0
Ireland [c]	1	0.3	1	5.0
Netherlands	3	0.9	1	5.0
New Zealand	6	1.8		
Norway		0.0	1	5.0
Sweden		0.0	1	5.0
Scotland	7	2.1		
Singapore	2	0.6		
USA [d]	275	82.3	15	75.0
Total	334		20	
USA states				
Florida	16	5.8	1	6.7
Indiana	2	0.7	3	20.0
New Jersey	2	0.7	1	6.7
New Mexico [d]	1	0.4	6	40.0
Pennsylvania	15	5.5	1	6.7
Utah	2	0.7	1	6.7
Washington	12	4.4	1	6.7
31 states plus DC	225	81.8	1	6.7
Remaining 12 states	0		NA	
Subtotal	275		15	

[a] Licensed FFT programs; accessed on August 27, 2023 at: https://www.fftllc.com/sites and https://functionalfamilytherapy.com/sites

[b] Randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) of FFT for behavior problems among youth ages 11-18 as of August 2020, not including 5 studies that provided no usable data.

[c] Includes 1 study that provided no usable data.

[d] Includes 4 studies conducted in New Mexico that provided no usable data.

Licensed FFT programs and completed trials

2 states contain

- ~ 1% of FFT programs in USA
- 60% of FFT evaluations in USA
- 45% of FFT evaluations in the world

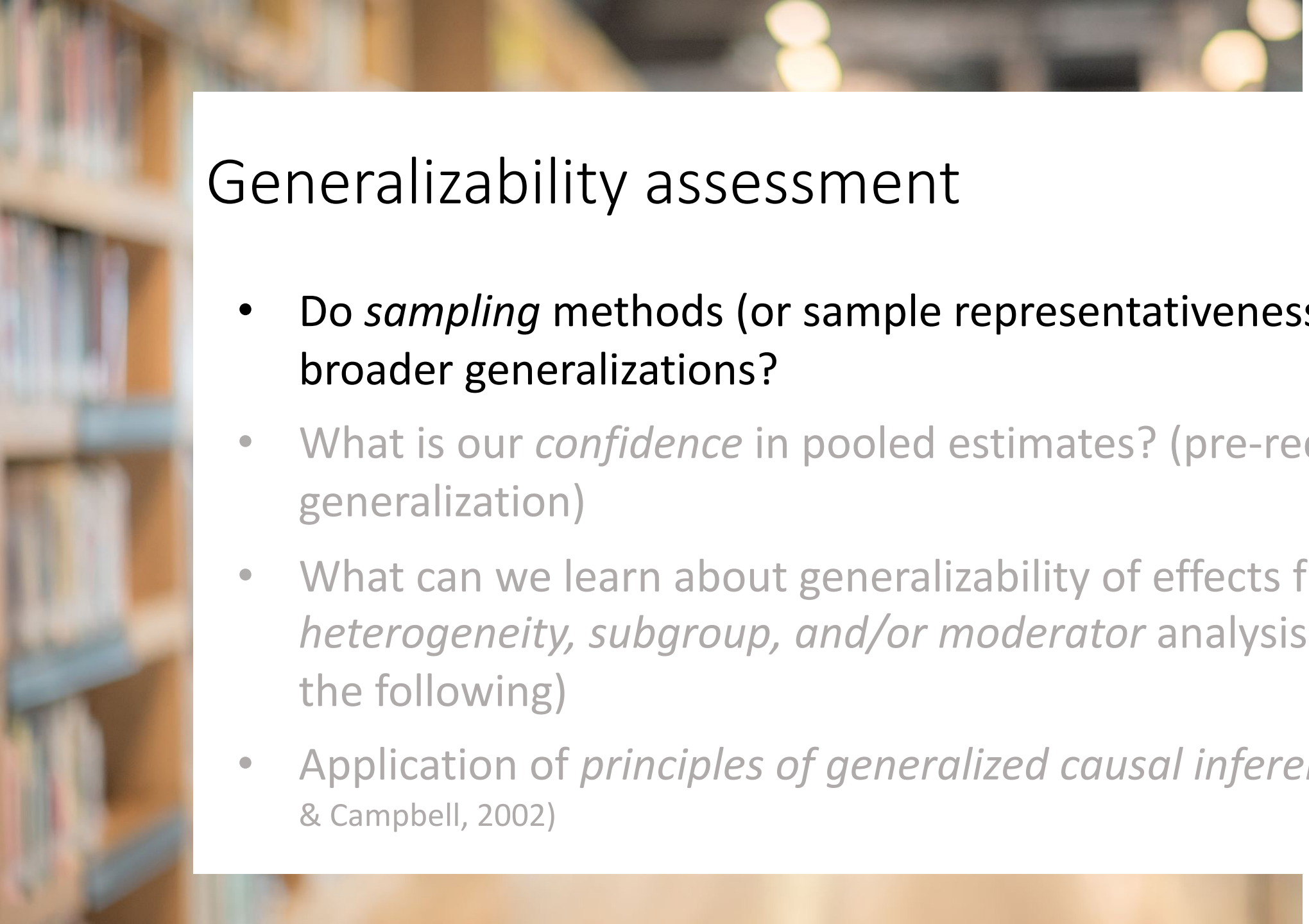
Country/State	Licensed FFT programs [a]		FFT RCTs/QEDs [b]	
	N	Column %	k	Column %
Australia	21	6.3		
Canada	2	0.6		
Denmark	8	2.4		
England	9	2.7	1	5.0
Ireland [c]	1	0.3	1	5.0
Netherlands	3	0.9	1	5.0
New Zealand	6	1.8		
Norway		0.0	1	5.0
Sweden		0.0	1	5.0
Scotland	7	2.1		
Singapore	2	0.6		
USA [d]	275	82.3	15	75.0
Total	334		20	
USA states				
Florida	16	5.8	1	6.7
Indiana	2	0.7	3	20.0
New Jersey	2		1	
New Mexico [d]	1	0.4	6	40.0
Pennsylvania	15	5.5	1	6.7
Utah	2	0.7	1	6.7
Washington	12	4.4	1	6.7
31 states plus DC	225	81.8	1	6.7
Remaining 12 states	0		NA	
Subtotal	275		15	

[a] Licensed FFT programs; accessed on August 27, 2023 at: https://www.fftllc.com/sites and https://functionalfamilytherapy.com/sites

[b] Randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) of FFT for behavior problems among youth ages 11-18 as of August 2020, not including 5 studies that provided no usable data.

[c] Includes 1 study that provided no usable data.

[d] Includes 4 studies conducted in New Mexico that provided no usable data.



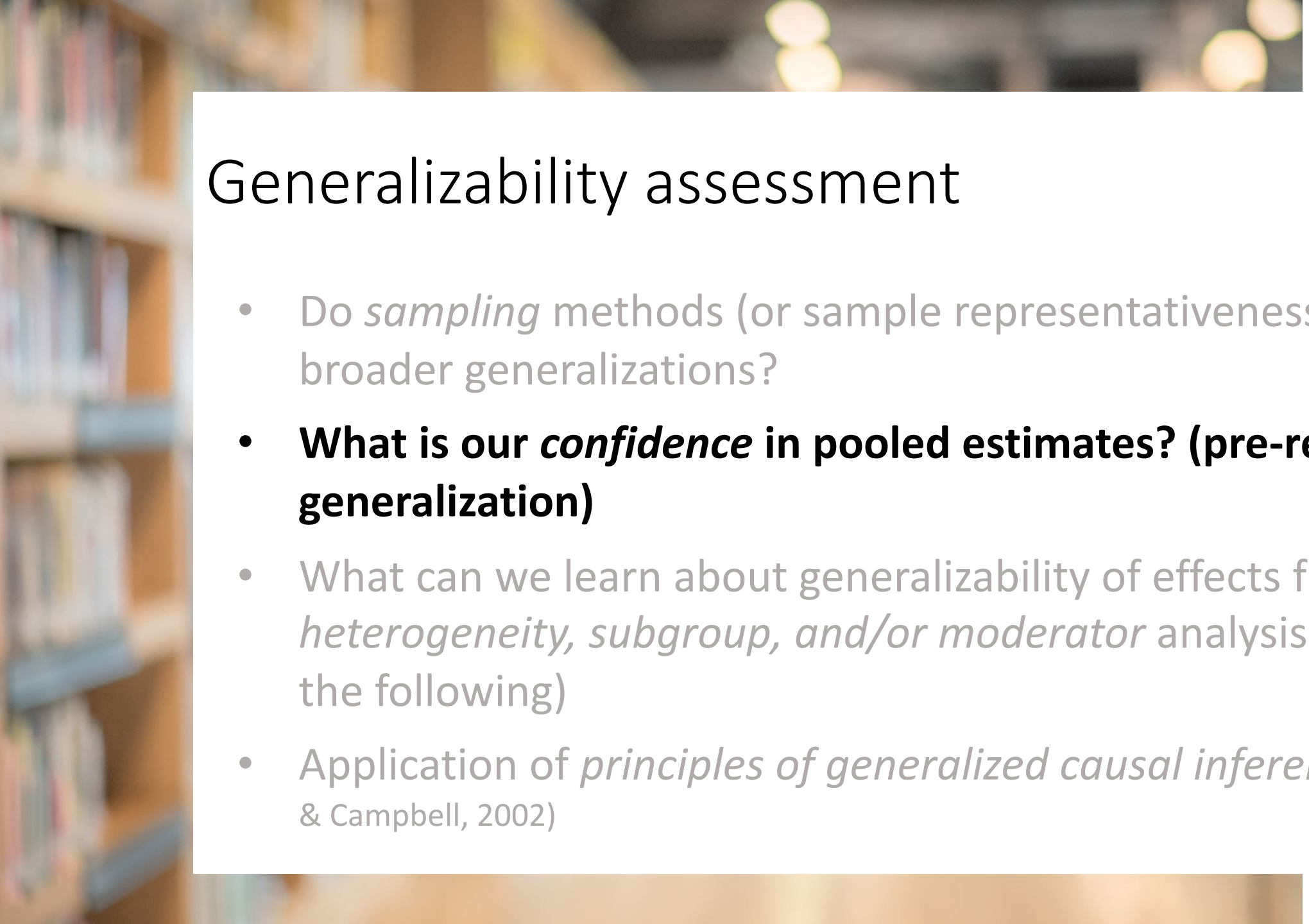
Generalizability assessment

- Do *sampling* methods (or sample representativeness) support broader generalizations?
- What is our *confidence* in pooled estimates? (pre-requisite for generalization)
- What can we learn about generalizability of effects from *heterogeneity, subgroup, and/or moderator* analysis? (to inform the following)
- Application of *principles of generalized causal inference* (Shadish, Cook, & Campbell, 2002)

Do *sampling* methods (or sample representativeness) support broader generalizations?

No.

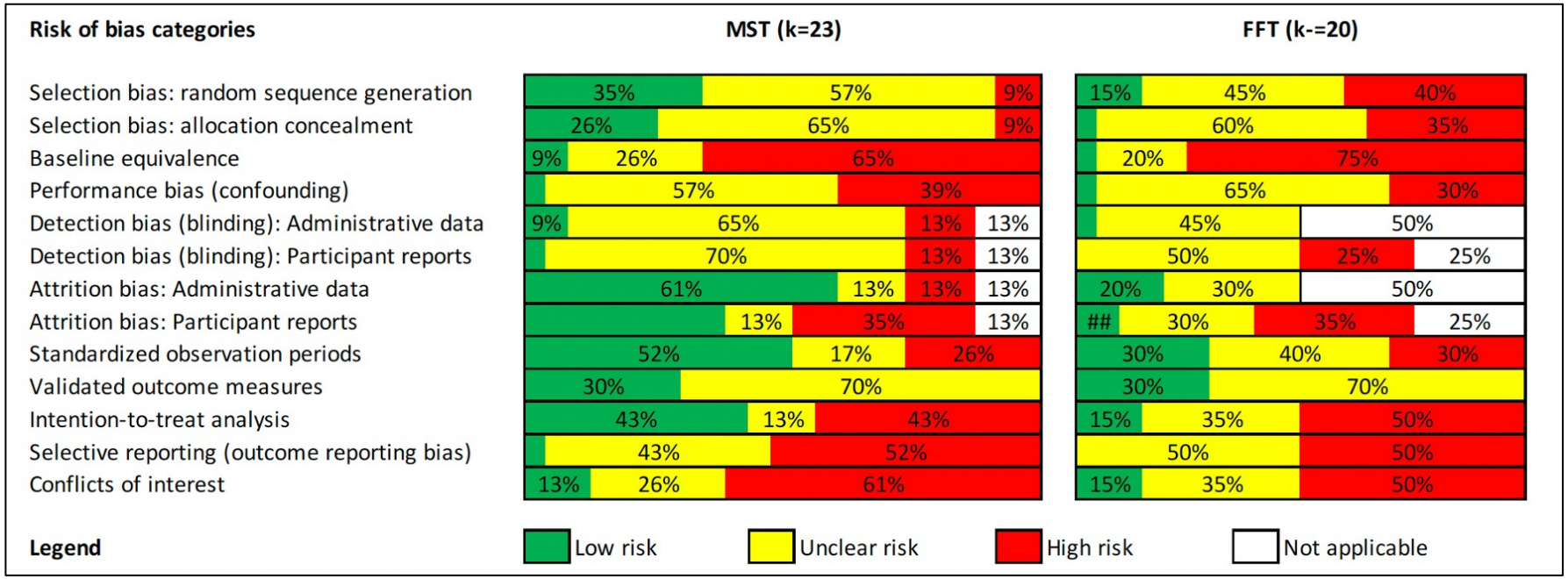
- Probability theory provides no basis for generalization from results of SRMAs based on
  - Nonprobability samples of studies
  - Studies that relied on nonprobability samples of participants
- Analysis suggests that available studies are not representative of MST/FFT programs in the countries or states in which they have been implemented.



Generalizability assessment

- Do *sampling* methods (or sample representativeness) support broader generalizations?
- **What is our *confidence* in pooled estimates? (pre-requisite for generalization)**
- What can we learn about generalizability of effects from *heterogeneity, subgroup, and/or moderator* analysis? (to inform the following)
- Application of *principles of generalized causal inference* (Shadish, Cook, & Campbell, 2002)

Confidence in results: Risk of bias ratings



- High risks of bias in > 50% of studies on: baseline equivalence, selective reporting of outcomes, conflicts of interest.
- 96% of MST trials and 100% of FFT impact evaluations have high risks of bias on at least one indicator.

Sources: Littell et al., (2021, 2023)

Consistency (PIs) and coverage (sparce data)

Results of correlated effects meta-analysis (Pustejovsky & Tipton, 2022)

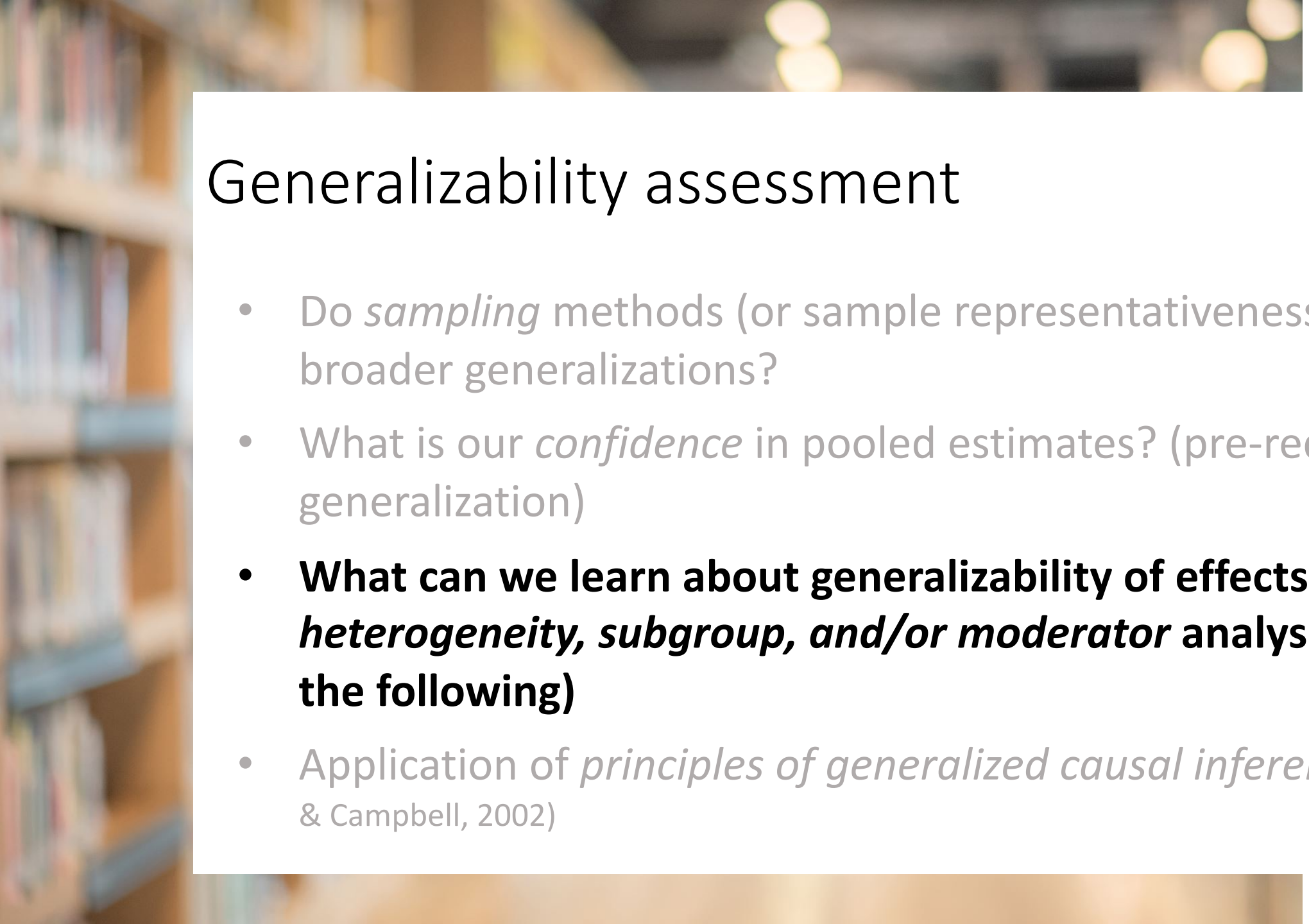
	MST (k=23)				FFT (k=20)			
Relative effects on outcomes (SMD)	95% PI		Valid k		95% PI		Valid k	
	LB	UB	k	%	LB	UB	k	%
Out of home placement	-0.72	0.17	17	74%			4	20%
Arrest or conviction	-0.55	0.26	18	78%	-0.39	0.76	8	40%
Delinquency	-1.31	0.77	14	61%			5	25%
Substance abuse	-1.35	1.20	9	39%			4	20%
Peer relations	-1.53	1.91	13	57%			3	15%
Youth behavior/symptom	-1.52	1.26	20	87%	-0.24	0.18	7	35%
Parent behavior/symptom	-0.77	0.45	16	70%			5	25%
Family functioning	-1.08	1.27	15	65%			5	25%
School	-1.92	2.55	8	35%			1	5%
All outcomes combined					-0.37	0.75	15	75%

Prediction intervals (PI) suggest that future studies are likely to find a wide range of positive and negative results

Sources: Littell et al., (2021, 2023)

Confidence in results: GRADE ratings

- **MST**: GRADE ratings of the certainty of evidence for the primary outcomes were ***moderate to low***,
  - meaning that further research is likely to affect confidence in estimates of effects and may change those estimates (Littell et al., 2021).
- **FFT**: GRADE ratings of the certainty of evidence were ***very low*** for all six primary outcomes,
  - meaning that any estimate of effects based on available data is very uncertain (PI for overall -0.37 to 0.75; Littell et al., 2023).
  - **Lacking confidence in evidence for FFT, we conclude that results are not generalizable** beyond the studies in the review.
- Generalizability assessment proceeds, based on the MST case study alone.



Generalizability assessment

- Do *sampling* methods (or sample representativeness) support broader generalizations?
- What is our *confidence* in pooled estimates? (pre-requisite for generalization)
- **What can we learn about generalizability of effects from *heterogeneity, subgroup, and/or moderator* analysis? (to inform the following)**
- Application of *principles of generalized causal inference* (Shadish, Cook, & Campbell, 2002)

Potential sources of heterogeneity (MST)

Effects sizes tend to be larger in studies…

- conducted in the USA vs other countries,
- by MST program developers vs independent teams,
- with higher risks of bias.

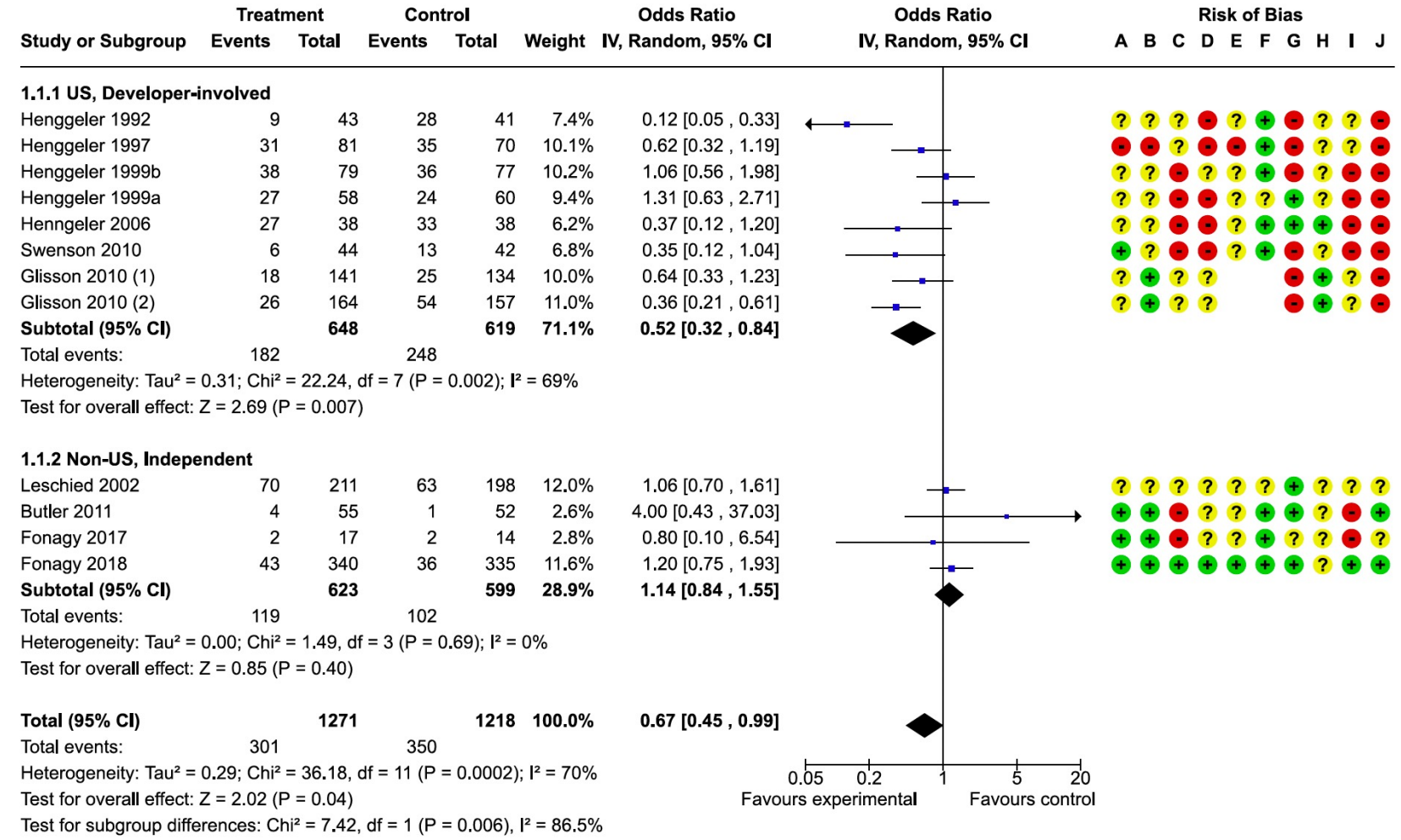
These three moderators are highly confounded

Studies conducted by MST developers are largely in the USA and have relatively high risks of bias.

Not possible to *explain* observed differences in effects between subgroups formed by these moderator variables.

MST effects on out-of-home placements at one year: US developers vs Non-US independents

Comparison 1: Out-of-home placement, Outcome 1: Out-of-home placement, 1 year



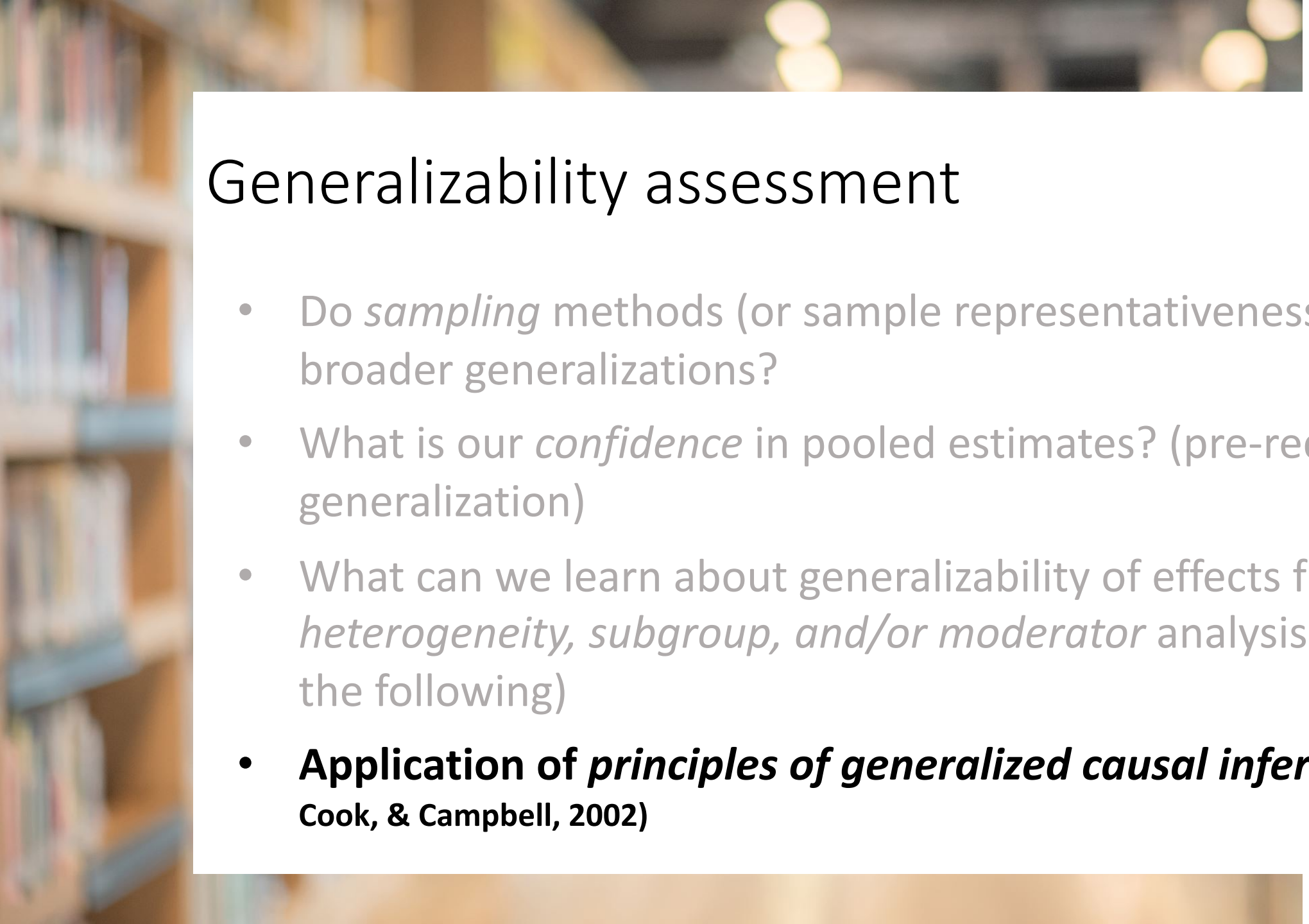
Study or Subgroup	Treatment Events	Treatment Total	Control Events	Control Total	Weight	Odds Ratio IV, Random, 95% CI
1.1.1 US, Developer-involved						
Henggeler 1992	9	43	28	41	7.4%	0.12 [0.05 , 0.33]
Henggeler 1997	31	81	35	70	10.1%	0.62 [0.32 , 1.19]
Henggeler 1999b	38	79	36	77	10.2%	1.06 [0.56 , 1.98]
Henggeler 1999a	27	58	24	60	9.4%	1.31 [0.63 , 2.71]
Henngeler 2006	27	38	33	38	6.2%	0.37 [0.12 , 1.20]
Swenson 2010	6	44	13	42	6.8%	0.35 [0.12 , 1.04]
Glisson 2010 (1)	18	141	25	134	10.0%	0.64 [0.33 , 1.23]
Glisson 2010 (2)	26	164	54	157	11.0%	0.36 [0.21 , 0.61]
Subtotal (95% CI)		**648**		**619**	**71.1%**	**0.52 [0.32 , 0.84]**
Total events:	182		248			
Heterogeneity: Tau² = 0.31; Chi² = 22.24, df = 7 (P = 0.002); I² = 69%						
Test for overall effect: Z = 2.69 (P = 0.007)						
1.1.2 Non-US, Independent						
Leschied 2002	70	211	63	198	12.0%	1.06 [0.70 , 1.61]
Butler 2011	4	55	1	52	2.6%	4.00 [0.43 , 37.03]
Fonagy 2017	2	17	2	14	2.8%	0.80 [0.10 , 6.54]
Fonagy 2018	43	340	36	335	11.6%	1.20 [0.75 , 1.93]
Subtotal (95% CI)		**623**		**599**	**28.9%**	**1.14 [0.84 , 1.55]**
Total events:	119		102			
Heterogeneity: Tau² = 0.00; Chi² = 1.49, df = 3 (P = 0.69); I² = 0%						
Test for overall effect: Z = 0.85 (P = 0.40)						
Total (95% CI)		**1271**		**1218**	**100.0%**	**0.67 [0.45 , 0.99]**
Total events:	301		350			
Heterogeneity: Tau² = 0.29; Chi² = 36.18, df = 11 (P = 0.0002); I² = 70%						
Test for overall effect: Z = 2.02 (P = 0.04)						
Test for subgroup differences: Chi² = 7.42, df = 1 (P = 0.006), I² = 86.5%						

Contextual differences in effects: Base rates in control groups

Outcome @ one year	Overall RD	USA			Non-USA		
		MST	Control	RD	MST	Control	RD
Arrest or conviction	-3%	40%	49%	-9%	25%	27%	-2%
Out-of-home placement of youth	-5% *	28%	40%	-12% **	19%	17%	+2%

RD = risk difference, * p <0.05, ** p<0.01

Source: Littell, Pigott, et al. (2021)



Generalizability assessment

- Do *sampling* methods (or sample representativeness) support broader generalizations?
- What is our *confidence* in pooled estimates? (pre-requisite for generalization)
- What can we learn about generalizability of effects from *heterogeneity, subgroup, and/or moderator* analysis? (to inform the following)
- **Application of *principles of generalized causal inference* (Shadish, Cook, & Campbell, 2002)**

Principles for generalized causal inference
(Shadish, Cook, & Campbell, 2002)

- Validity is a property of knowledge claims (inferences based on data),
  - not a property of research methods.
- External validity is a property of certain inferences (extrapolation),
  - not a property of probability sampling methods.
- Logic of generalization: a conceptual problem, with empirical referents.

The logic of generalization:
A conceptual problem with empirical referents

How can we transfer knowledge developed in

- one set of multi-attribute contexts (studies) to
- other sets of multi-attribute contexts (targets for generalization)?

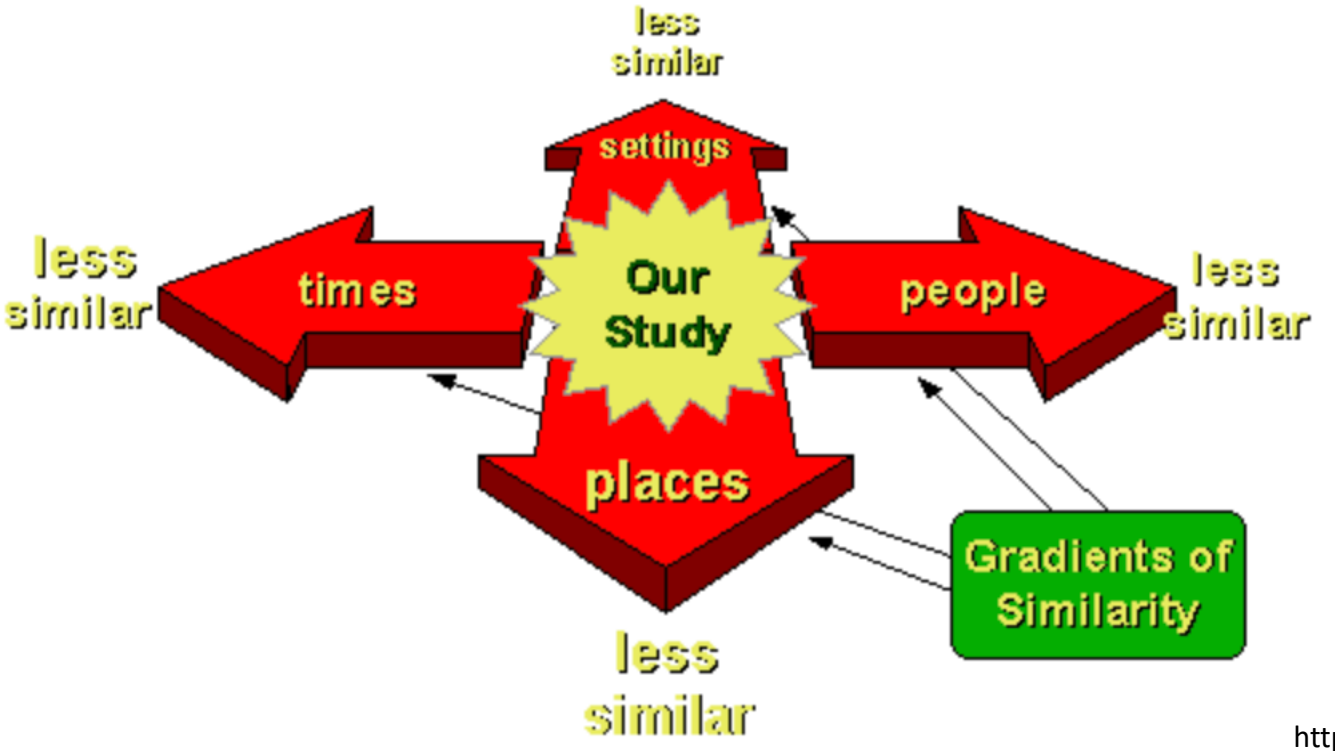
  - “[W]e need something more appropriate than the generalization rhetoric and the solution of it by representative sampling from a universe designated in advance… In this shift, the validity of theoretical interpretation replaces atheoretical generalization…” (Campbell, 1986, p. 73).

Principles for generalized causal inference (Shadish, Cook, & Campbell, 2002)

1) Proximal similarity (or surface similarity)
2) Ruling out irrelevancies
3) Making discriminations
4) Interpolation and extrapolation
5) Causal explanation

1) Proximal (or surface) similarity

"We generalize most confidently **to applications** where treatments, settings, populations, outcomes, and times are **most similar** to those in the original research" (Shadish, 1995; emphasis added).



https://conjointly.com/kb/external-validity

Which gradients of similarity are most salient?

<u>PICOTS gradients of similarity</u>:

Populations

Interventions

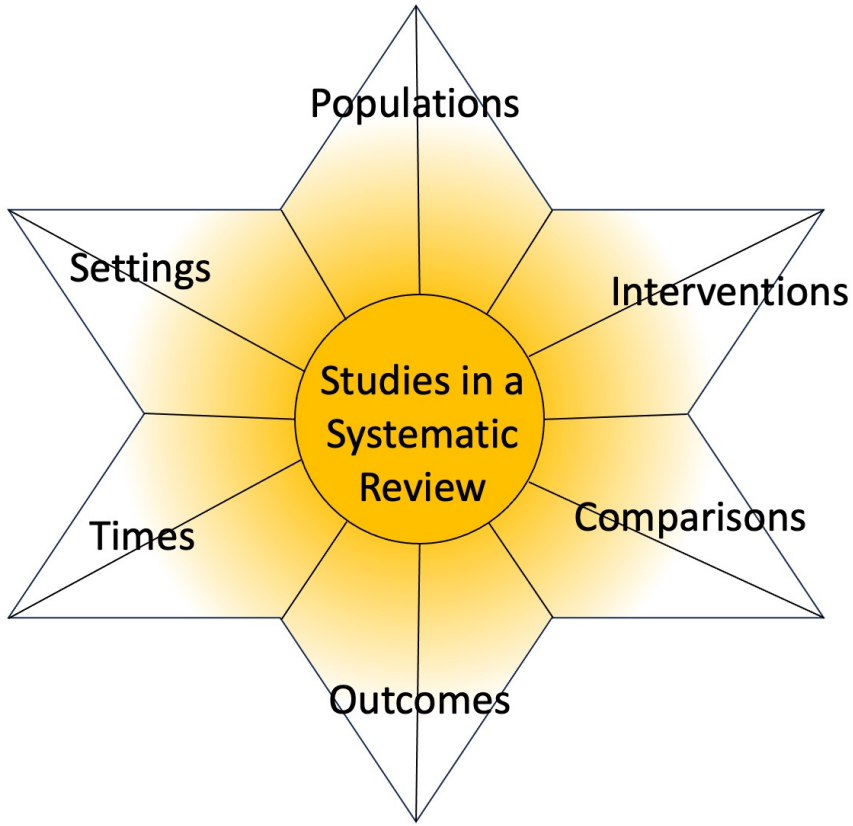
Comparisons

Outcomes

Times

Settings

Other gradients?



What variables/factors should define each gradient?

N-dimensional comparison space, where N=number of variables/factors on salient gradients

Gradients of similarity for MST?

PICOS gradients of similarity for MST

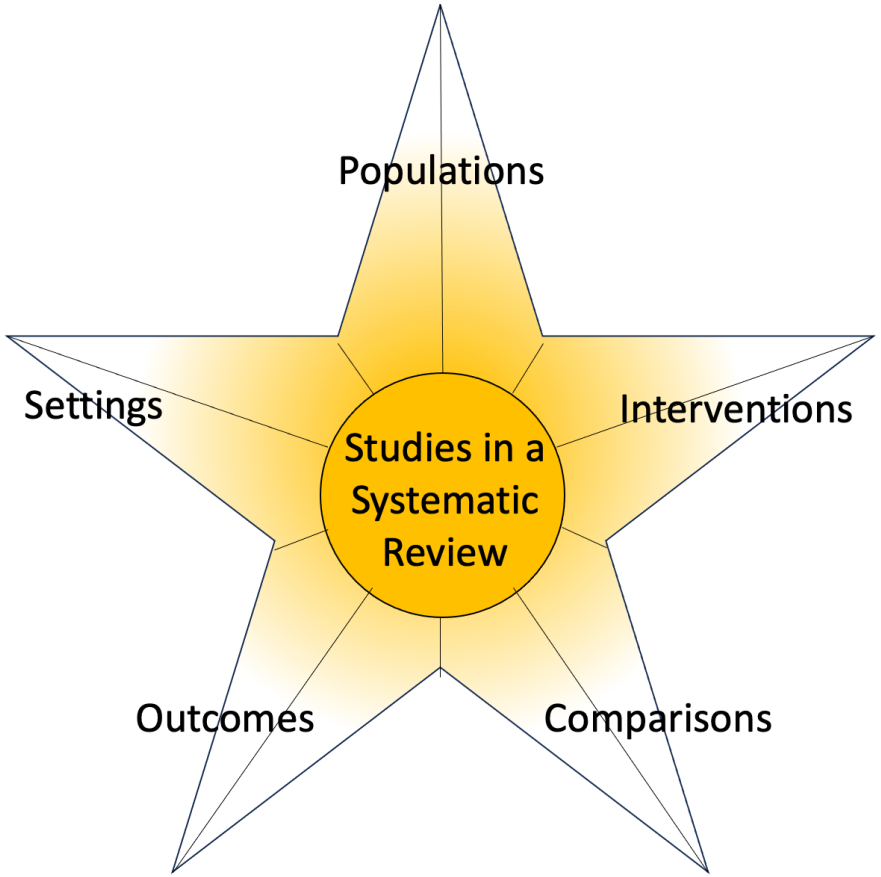
Populations: families of youth with SEB problems

Interventions: short-term, home & community services

Comparisons: varied amounts & types of services

Outcomes: depends on stakeholder goals

Settings: WEIRD countries



WEIRD = Western, Educated, Industrialized, Rich, Democratic

What do we know about these gradients? Study data

- Descriptive data are often not comparable across studies
  - e.g., diverse measures of SES/household income/poverty status; household composition
- Comparable data are superficial or sparce
  - treatment duration, amount
  - participants age, gender, race
- Uneven measurement of outcomes across studies (valid k ranges from 8 to 20 /23)

Descriptive data	MST (k=23)			
	Min	Max	Valid k (%)	
Treatments				
Mean duration (days)	94	231	17	74%
Mean hours of direct contact	21	92	8	35%
Comparisons				
Mean duration (days)	83	380	6	26%
Mean hours of direct contact	23	76	3	13%
Treatment as usual			20	87%
Other active treatments			3	13%
Settings				
USA			16	70%
Other countries			7	30%
Developers' studies			13	57%
Mix of urban, suburban, rural locations			12	52%
Participants				
Mean age (focal youth)	13.4	16.0	22	96%
% male	44%	100%	23	100%
% White	10%	95%	21	91%
% Black	7%	81%	19	83%
Times				
Year enrollment began	1983	2014	17	74%
Relative effects on outcomes (SMD)	95% PI		Valid k	
	LB	UB	k	%
Out of home placement	-0.72	0.17	17	74%
Arrest or conviction	-0.55	0.26	18	78%
Delinquency	-1.31	0.77	14	61%
Substance abuse	-1.35	1.20	9	39%
Peer relations	-1.53	1.91	13	57%
Youth behavior/symptoms	-1.52	1.26	20	87%
Parent behavior/symptoms	-0.77	0.45	16	70%
Family functioning	-1.08	1.27	15	65%
School	-1.92	2.55	8	35%

Proximal similarity is necessary but insufficient for generalized causal inferences (Shadish et al., 2002)

"Perhaps the principle of proximal similarity merely describes the route to theory-based generalization…" (Campbell, 1986, p. 73).

Need other principles to flesh out generalizability assessment…

2) Ruling out irrelevancies

"We generalize most confidently when a research finding **continues to hold over variations** in persons, settings, treatments, outcome measures, and times that are **presumed to be conceptually irrelevant**" (Shadish, 1995; emphasis added).

Which variations are thought to be conceptually irrelevant?

MST claims that no PICOTS are irrelevant, effects are robust across all PICOTS variations.

- SRMA refutes these claims, by showing that results are inconsistent within and across studies and
- Inconsistent across PICOTS.

3) Making discriminations (Discriminant validity)

"We generalize most confidently when we can show that it is **the target construct**, and not something else, that is necessary to producing a research finding" (Shadish, 1995; emphasis added).

Obstacles:

- MST treatment is confounded with other variables that might account for effects
  - MST cases received more time and attention than control cases
  - MST therapists received more training and supervision than workers who provided services to control cases (Littell, Pigott, et al., 2021).
- MST fidelity measures are confounded with other variables known to predict positive outcomes (therapeutic alliance, client satisfaction, client engagement, early outcomes) and have not been shown to discriminate between MST and other treatments.

4) Interpolation and extrapolation

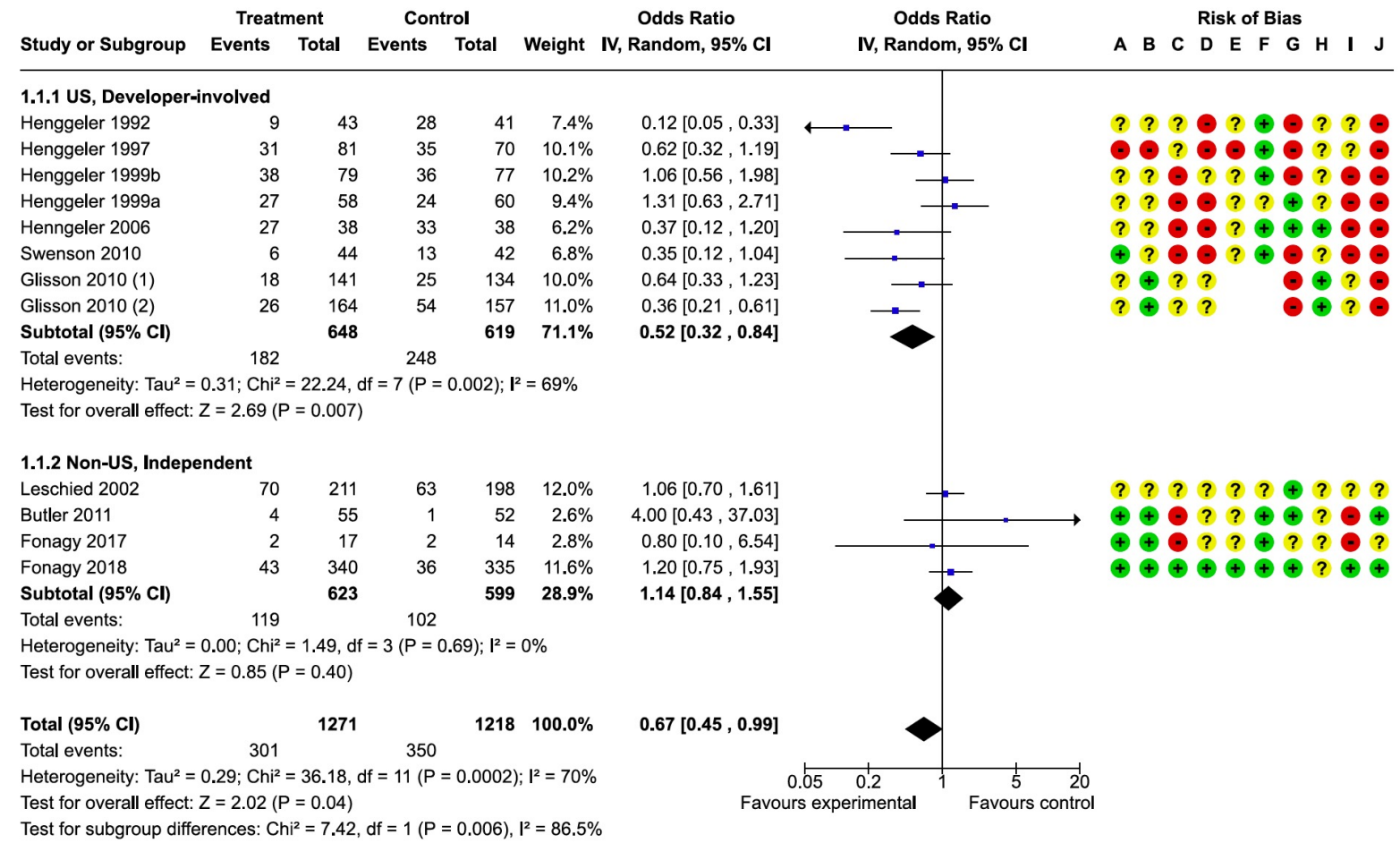
"We generalize most confidently when we can **specify the range** of persons, settings, treatments, outcomes, and times over which the **finding holds more strongly, less strongly, or not at all**" (Shadish, 1995).

Obstacles:

- Moderators of effects of MST (USA/other, control conditions, risks of bias, developers/others) are confounded and
- There are unexplained variations within subgroups (Littell, Pigott, et al., 2021).

MST effects on out-of-home placements at one year: US developers vs Non-US independents

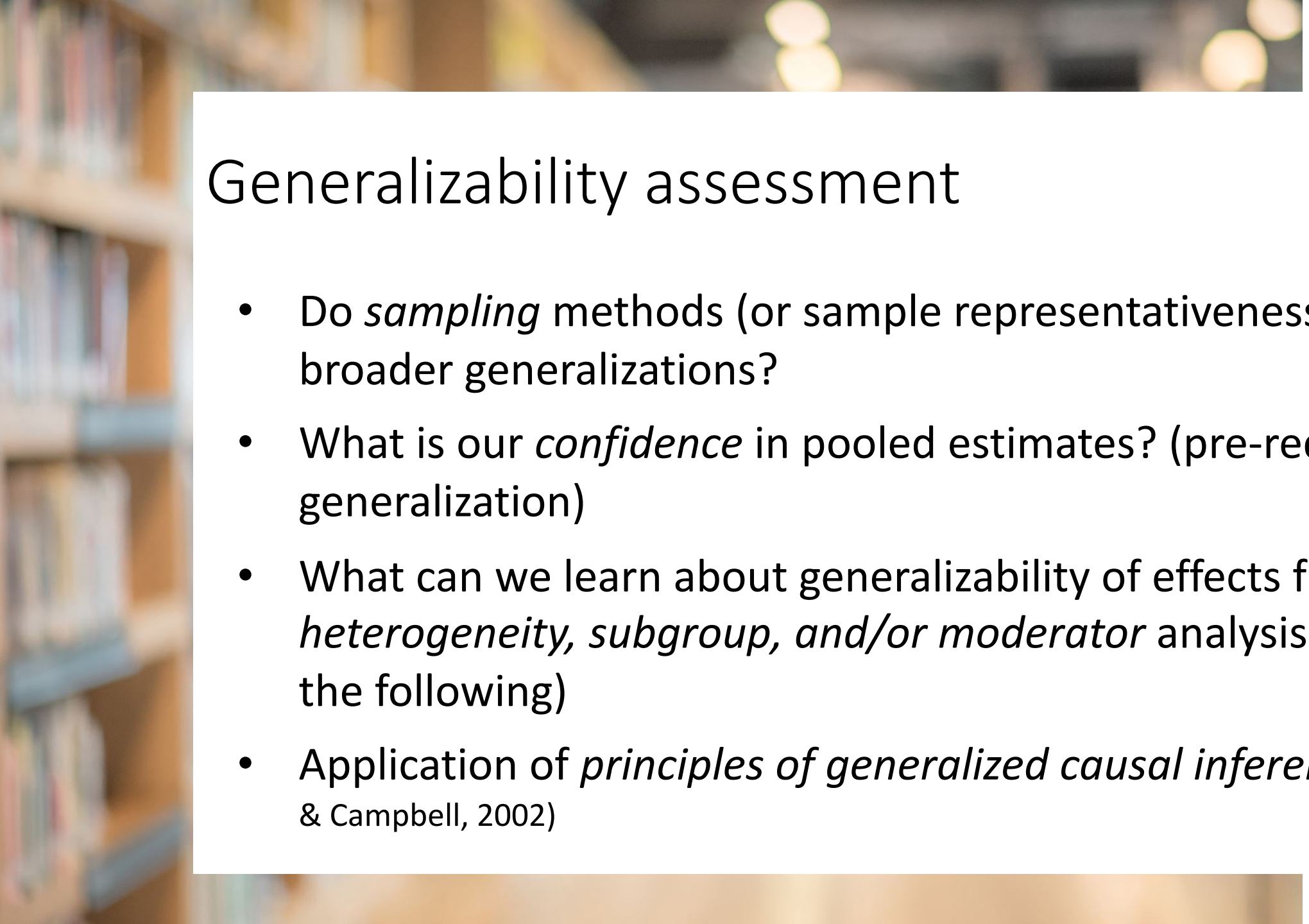
Comparison 1: Out-of-home placement, Outcome 1: Out-of-home placement, 1 year



5) Causal explanation

"We generalize most confidently when we can **specify completely and exactly** (a) which parts of one variable (b) are related to which parts of another variable (c) through which mediating processes (d) with which salient interactions, for then we can transfer only those essential components to the new application to which we wish to generalize" (Shadish, 1995).

Obstacle: MST theory of change is under-developed, does not fully explain hypothesized effects of treatment, or account for actual (inconsistent) results of studies and SRMAs.



Generalizability assessment

- Do *sampling* methods (or sample representativeness) support broader generalizations?
- What is our *confidence* in pooled estimates? (pre-requisite for generalization)
- What can we learn about generalizability of effects from *heterogeneity, subgroup, and/or moderator* analysis? (to inform the following)
- Application of *principles of generalized causal inference* (Shadish, Cook, & Campbell, 2002)

Summary: MST generalizability assessment

Criteria	Support for generalized causal inferences	
	Rating	Reasons/Support
Probability/representative sampling	None	Samples are not representative of countries or states with MST programs; MST developer-led studies are over-represented.
Certainty of evidence	Moderate/Low	Risk of bias and GRADE ratings.
Proximal similarity	Unclear	Insufficient descriptive data.
Ruling out irrelevancies	None	Results are inconsistent within and across: studies, USA vs other countries, developers/other investigators, outcome measures, endpoints.
Discriminant validity	None	MST is confounded with amount of service provided (time and attention), worker training and supervision; MST fidelity measures lack face validity, content validity, and discriminant validity.
Interpolation, extrapolation	Unclear	Confounded moderators and unexplained variations within subgroups.
Causal explanation	None	MST theory of change is under-developed, does not fully explain hypothesized effects of treatment, or empirical results of studies and SRMAs.

Summary: FFT generalizability assessment

Criteria	Support for generalized causal inferences	
	Rating	Reasons
Probability/representative sampling	None	Samples are not representative of countries and states with MST programs; developer-led studies are over-represented.
Certainty of evidence	Very Low	Risk of bias and GRADE ratings.
Proximal similarity		Insufficient data
Ruling out irrelevancies		Insufficient data
Discriminant validity		Insufficient data
Interpolation, extrapolation		Insufficient data
Causal explanation		Insufficient data

Generalizability assessment suggests

- Results of MST and FFT are **not widely generalizable**.
- Need better primary studies (with lower risks of bias) that control for factors confounded with treatment (time, attention, training, supervision).
- **Application** of MST results to specific contexts *might* be possible
  - Begin with knowledge of relevant PICOS in local context
  - Identify MST trials most similar to target context(s)
    - Assess credibility of estimates produced by these trials (risk of bias, GRADE)
  - Re-analysis of SRMA data if necessary/possible to estimate likely effects based on selected subgroup of studies (see Shackleford et al., 2021, on dynamic meta-analysis)
    - Obstacle: little statistical power for subgroups analysis in MST review

Common rubrics and rhetoric re: generalization

Use of the **mean effect size--**or a **rating** based on mean ES--as the best available estimate of likely effects.

Often presented without confidence intervals or prediction intervals.

YOUTH ENDOWMENT FUND

WHAT WORKS TO PREVENT VIOLENCE?

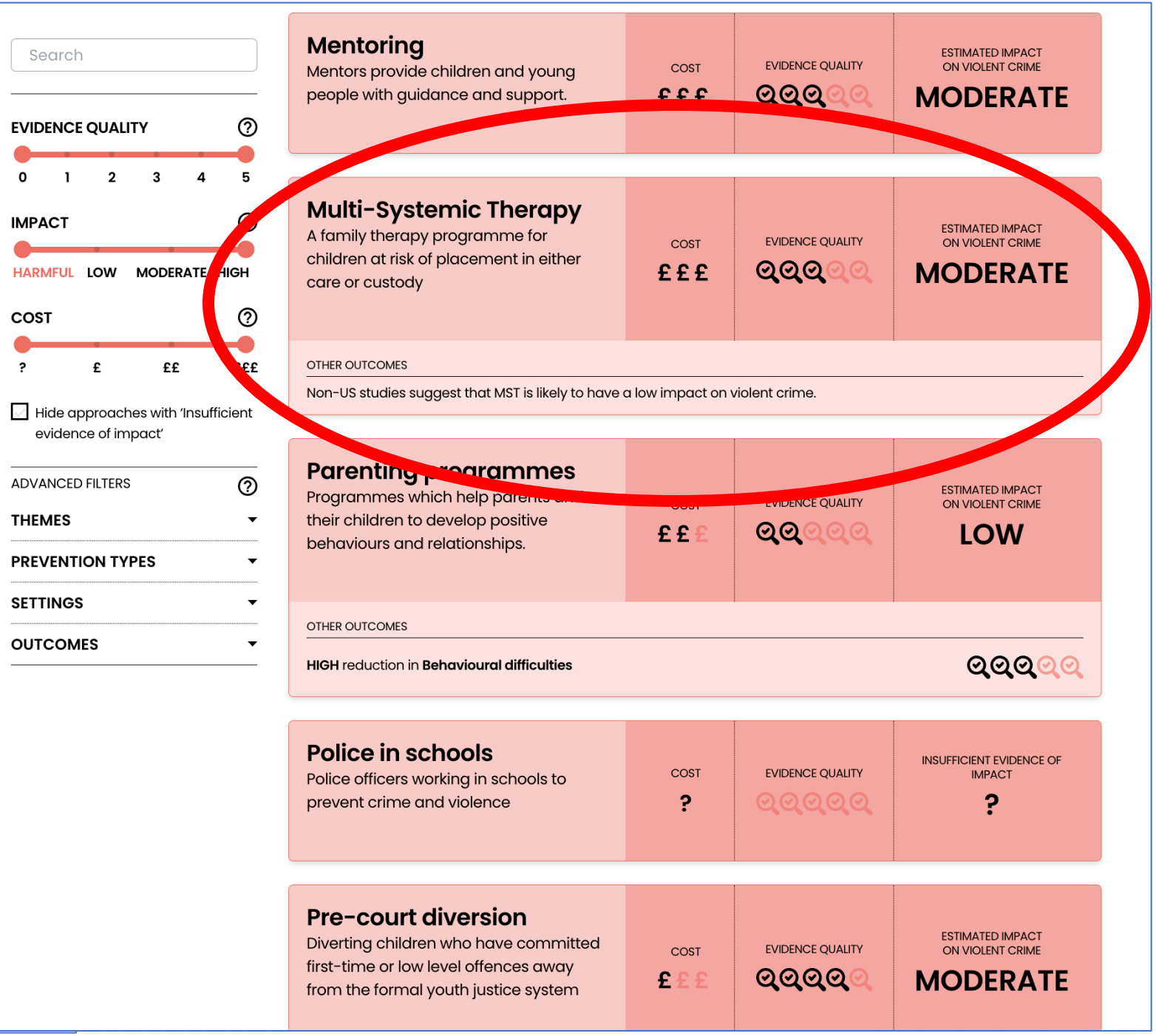
YEF Toolkit

A free online resource to help you put evidence of what works to prevent serious violence into action.

VISIT THE TOOLKIT →

About the Toolkit →

https://youthendowmentfund.org.uk/toolkit/



What is “moderate impact”? (in YEF Toolkit)

“The review estimates that MST reduces… offending by 17%.”
https://youthendowmentfund.org.uk/toolkit/multi-systemic-therapy-2/

- Estimate derived from our meta-analysis of data on arrests/convictions at one year (Littell, Pigott, et al., 2021).
- Incorrect. The overall risk difference is 3% and it is not statistically different from zero (p>.05), but this is not mentioned.
- No confidence interval (or prediction interval) is provided.

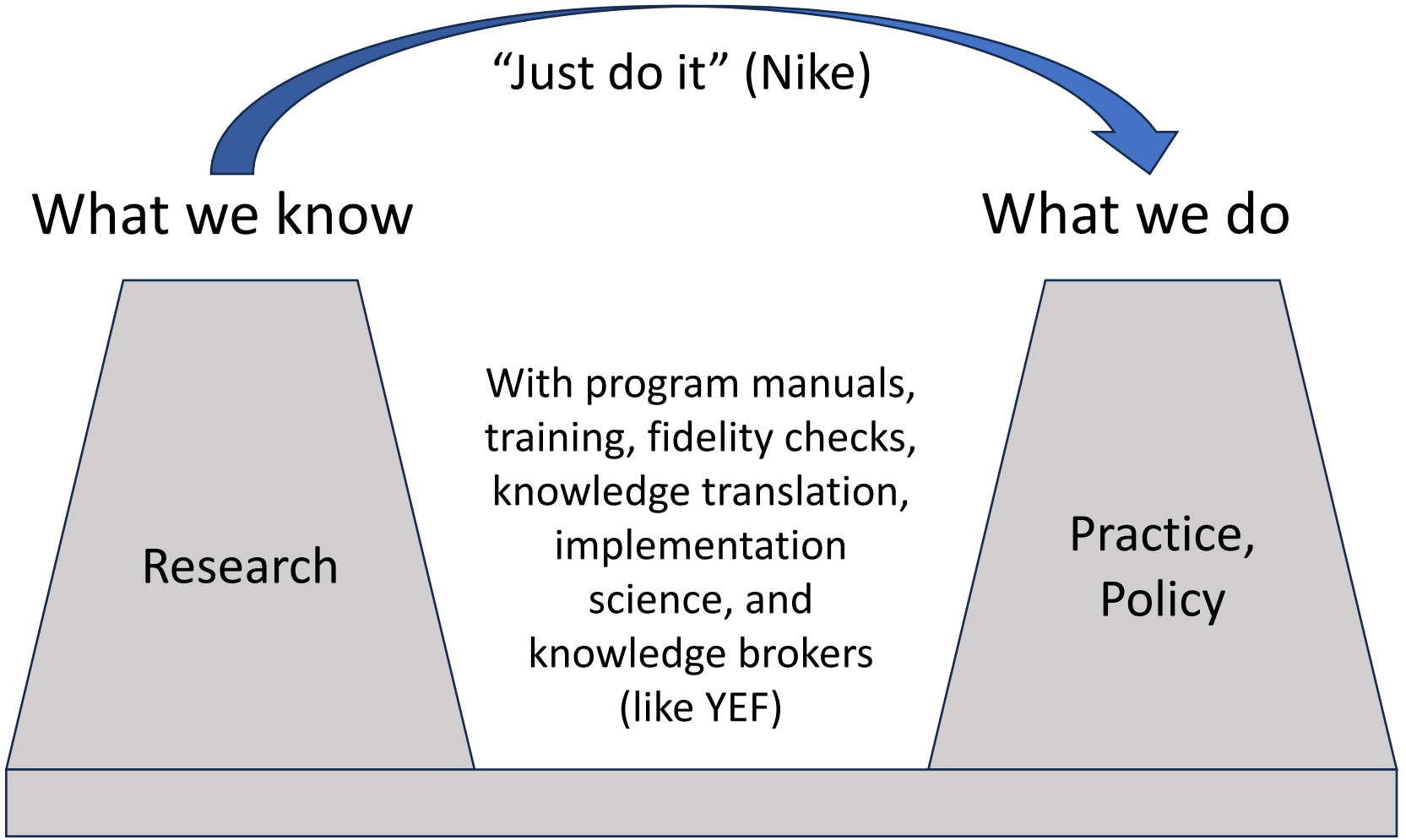
Consistent with APA and SPR guidelines, focus on “positive, pooled effect sizes” without specifying magnitude of ES or whether CIs or PIs can include null/negative effects (Tolin et al., 2015; Gottfredson et al., 2015).

Mean effect sizes are relatively uninformative for purposes of generalization

- Ignore heterogeneity, confidence/credibility of estimates (risks of bias), subgroup differences, moderators…
- Mean effects may have no real **meaning** anywhere in the world.
- When presented without CIs or PIs, point estimates convey "**incredible certitude**" (Manski, 2013).
  - Credible estimates are provided within a range (CI, PI)

Related rhetoric: Bridging the “know-do gap”

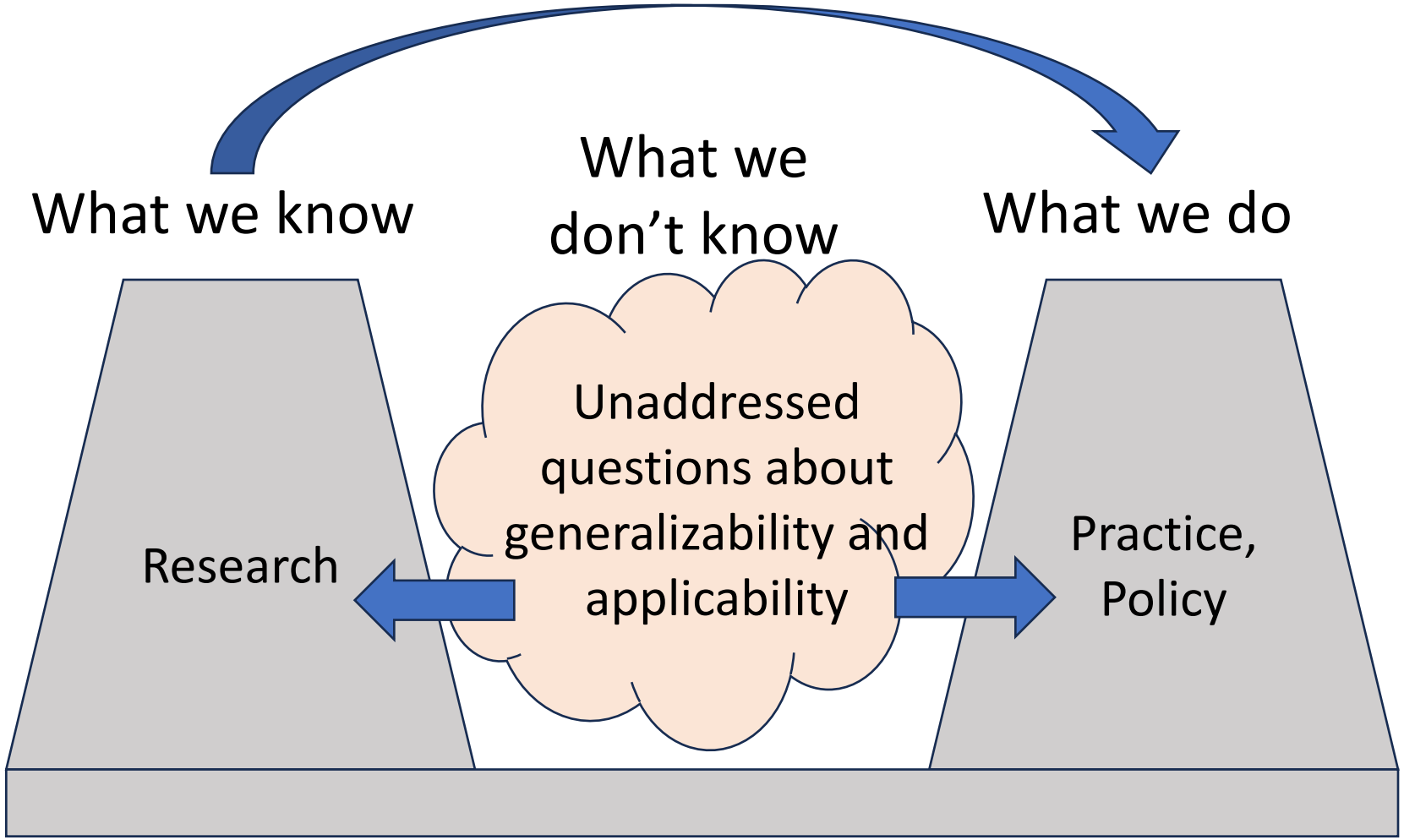
- Knowledge translation
- Implementation science



“Wishful extrapolation” (Manski, 2013)

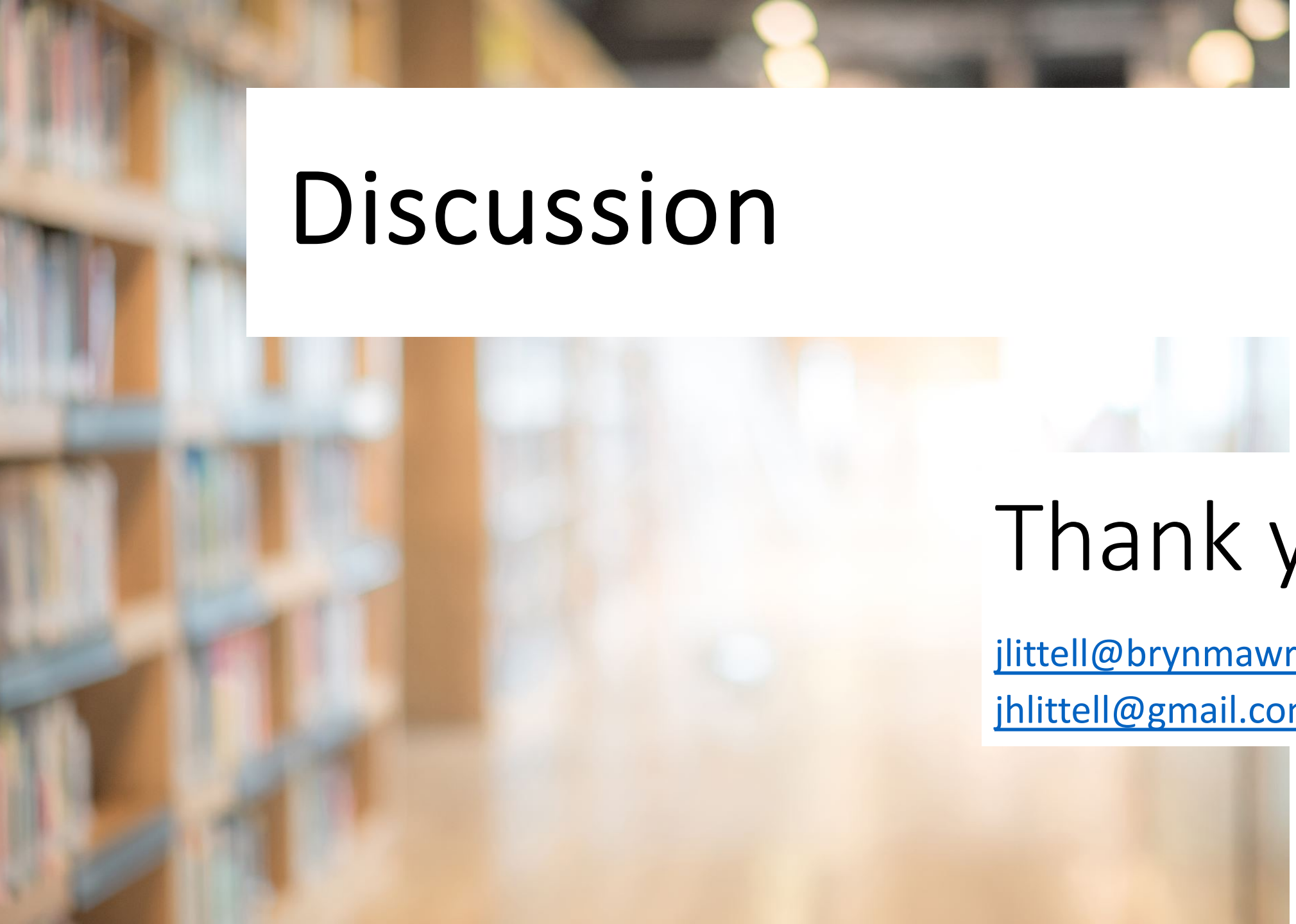
SRMAs can identify gaps in knowledge and directions for further work.

Generalizability assessment might help.



Conclusions

- Do pooled estimates from SRMAs have greater external validity than study-level effect sizes?
  - Short answer: not necessarily
  - Long answer: Logic of generalization from SRMAs is woefully underdeveloped.
- How can we use SRMAs to inform inferences about generalizability?
  1. Test generalizability claims
  2. Use subgroup/moderator analyses to identify limits on generalizability
  3. Apply principles of generalized causal inference to SRMA data
  4. Identify directions for further primary research to address unanswered questions about generalizability
- More attention to the logic of generalization is needed.



Discussion

Thank you!

jlittell@brynmawr.edu
jhlittell@gmail.com

References

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. *New Directions for Program Evaluation*, *1986*(31), 67–77. https://doi.org/10.1002/ev.1434

Cook, T. D. (1990). The generalization of causal connections: Multiple theories in search of clear practice. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), Research methodology: Strengthening causal interpretations of nonexperimental data (DHHS Publication No. PHS 90-3454, pp. 9-31). Rockville MD: Department of Health and Human Services.

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation. *Prevention Science*, *16*(7), 893–926. https://doi.org/10.1007/s11121-015-0555-x

Littell, J. H., Pigott, T. D., Nilsen, K. H., Green, S. J., & Montgomery, O. L. K. (2021). Multisystemic Therapy® for social, emotional, andbehavioural problems in youth aged 10-17: An updated systematic review and meta-analysis. *Campbell Systematic Reviews*, *17*(4), e1158. https://doi.org/10.1002/cl2.1158

Littell, J. H., Pigott, T. D., Nilsen, K. H., Roberts, J., & Labrum, T. K. (2023). Functional Family Therapy for families of youth (age 11–18) with behaviour problems: A systematic review and meta-analysis. *Campbell Systematic Reviews*, *19*(3), e1324. https://doi.org/10.1002/cl2.1324

Manski, C. F. (2013). *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press. https://doi.org/10.4159/harvard.9780674067547

McCart, M. R., Sheidow, A. J., & Jaramillo, J. (2023). Evidence Base Update of Psychosocial Treatments for Adolescents with Disruptive Behavior. *Journal of Clinical Child & Adolescent Psychology*, *52*(4), 447–474. https://doi.org/10.1080/15374416.2022.2145566

Shackelford, G. E., Martin, P. A., Hood, A. S. C., Christie, A. P., Kulinskaya, E., & Sutherland, W. J. (2021). Dynamic meta-analysis: A method of using global evidence for local decision making. *BMC Biology*, *19*(1), 33. https://doi.org/10.1186/s12915-021-00974-w

Shadish, W. R. (1995). The logic of generalization: Five principles common to experiments and ethnographies. *American Journal of Community Psychology*, *23*, 419–427. https://doi.org/10.1007/BF02506951

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

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically Supported Treatment: Recommendations for a New Model. *Clinical Psychology: Science and Practice*, *22*(4), 317–338. https://doi.org/10.1111/cpsp.12122

Youth Endowment Fund Toolkit. https://youthendowmentfund.org.uk/toolkit/