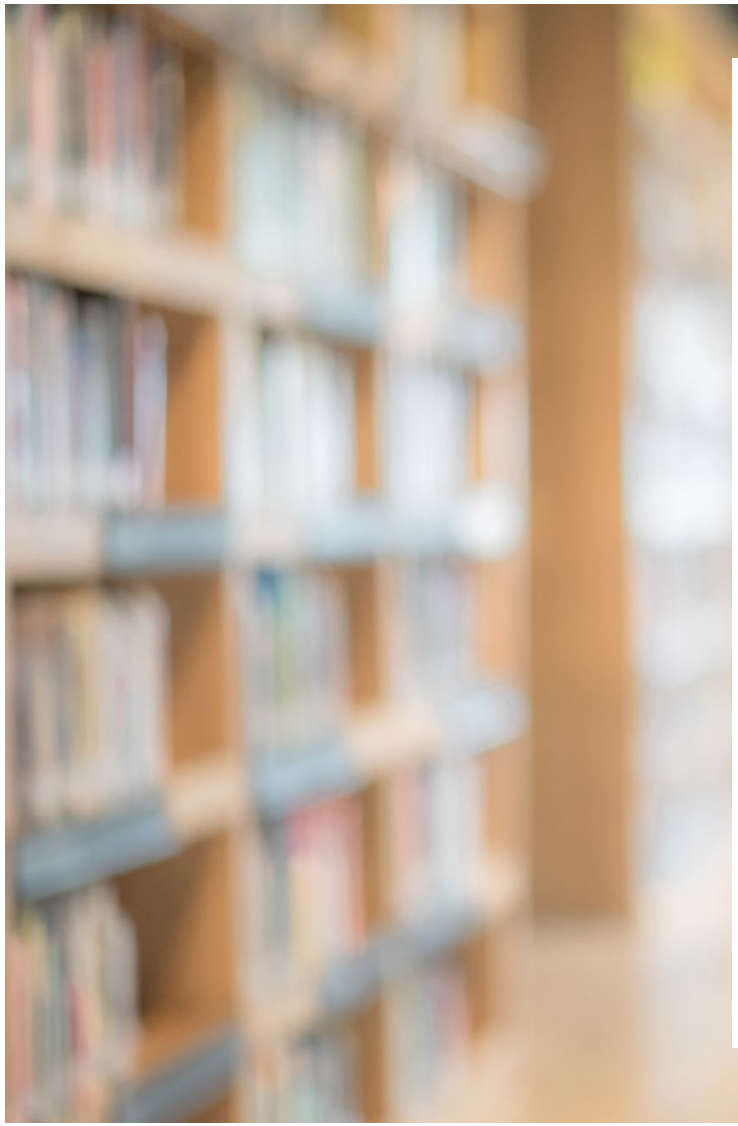
The logic of generalization from systematic reviews to policy and practice

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

Seminar on External Validity in Program Evaluation
Sciences Po, Paris, France
6 June 2023



Outline of the presentation

- Introduction
  - Systematic reviews (SRs)
  - Focus of the presentation
  - Working example: a case study
- Logic of generalization from SRs
  1. Consensus-based rubrics
  2. Probability theory and inferential statistics
  3. Proximal similarity and other principles for generalizing from nonprobability samples
- Conclusions and directions for further work

Systematic reviews: definition, types

- **Ideally**, systematic reviews (SRs)
  - follow established guidelines (e.g., PRISMA, STROBE, MOOSE, MECIR, MECCIR),
  - use *a prior* protocols and transparent methods to
  - locate, critically appraise, and synthesize results of multiple studies,
  - address a well-defined research question/hypothesis,
  - use meta-analysis to synthesize quantitative data when applicable, and
  - attempt to minimize bias and error at each step in the review process.
- SRs can address **different types of questions** about…
  - Rates and trends, associations, risk and protective factors, diagnostic test accuracy, interventions, methodological issues, etc.
- I will focus on systematic reviews of **interventions effects** because
  - Most of the work on SR methods and applications has been done in this area.

Systematic reviews of research on interventions

Conducted for various **purposes**, including

1. **Exploratory, descriptive, or theory-building** aims

   How does an intervention operate in various settings? Who does it serve? What experiences and views do clients and staffs have?

2. **Confirmatory** (hypothesis testing) aims

   E.g., one treatment is more effective than another treatment

I will focus on the generalizability of reviews aimed at **testing/confirming** claims about intervention **effects**.

Because there is considerable interest in generalizing from these kinds of reviews.

Working example (Littell et al., 2004, 2005, 2021)

DOI: 10.1002/cl2.1158

UPDATED SYSTEMATIC REVIEWS

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]

[1]Graduate School of Social Work and Social Research, Bryn Mawr College, Bryn Mawr, Pennsylvania, USA

[2]School of Public Health, Georgia State University, Atlanta, Georgia, USA

[3]Regional Centre for Child and Adolescent Mental Health, Eastern and Southern Norway (RBUP), Oslo, Norway

[4]Counseling and Psychological Services, Swarthmore College, Swarthmore, Pennsylvania, USA

[5]Richmond, Virginia, USA

Correspondence
Julia H. Littell, Graduate School of Social Work and Social Research, Bryn Mawr College, Bryn Mawr, PA 19010, USA.
Email: jhlittell@gmail.com and jlittell@brynmawr.edu

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.

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® (MST) compared with treatment as usual or other treatments for youth with social, emotional, and behavioural problems. The quality of this evidence is uneven. It shows that effects of MST vary across studies, settings, outcomes, and endpoints.

Although most MST trials produce a mixture of positive, negative, and null findings, many reports focus selectively on positive, statistically significant results instead of all results.

What is the aim of this review?

This Campbell updated systematic review and meta-analysis synthesises data from all eligible trials to test the claim that Multisystemic Therapy® is effective across clinical problems and populations.

What is this review about?

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.

We synthesise data from all eligible trials to test the claim that MST is effective across clinical problems and populations.

What studies are included?

Included studies examine outcomes of MST for juvenile offenders, sex offenders, offenders with substance abuse problems, youth with conduct or behaviour problems, those with serious mental health problems, autism spectrum disorder, and cases of child maltreatment.

This review summarises findings from 23 randomised controlled trials of the effects of MST. These trials were conducted in the USA, the UK, Canada, The Netherlands, Norway and Sweden.

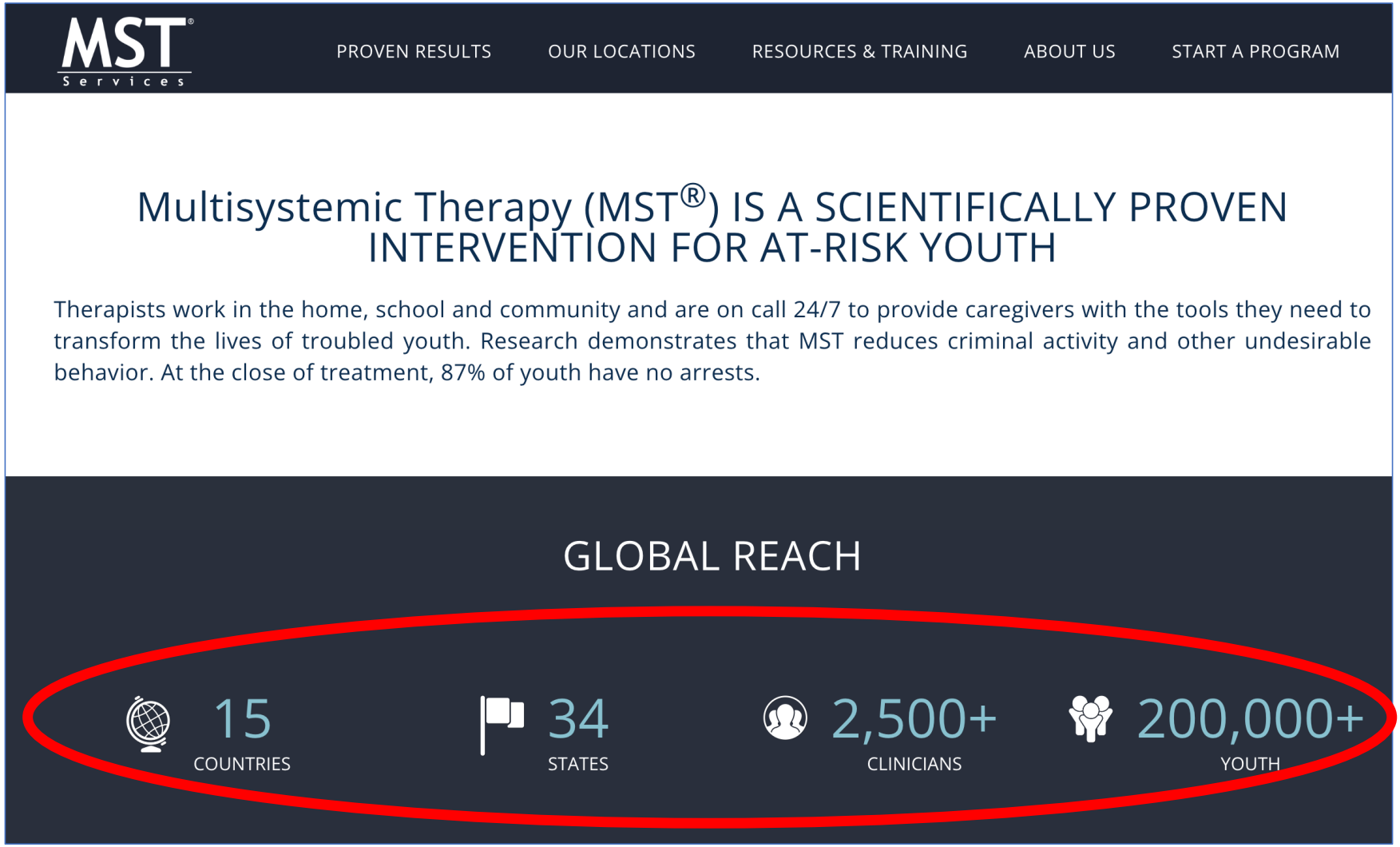
Most trials compare MST to treatment as usual (TAU). In the USA, TAU consists of relatively little contact and few services for youth and families, compared with more robust public health and social services available to youth in other high-income countries. One US study provided 'enhanced TAU' to families in the control group, and two US studies compared MST to individual therapy for youth.

What are the main findings of this review?

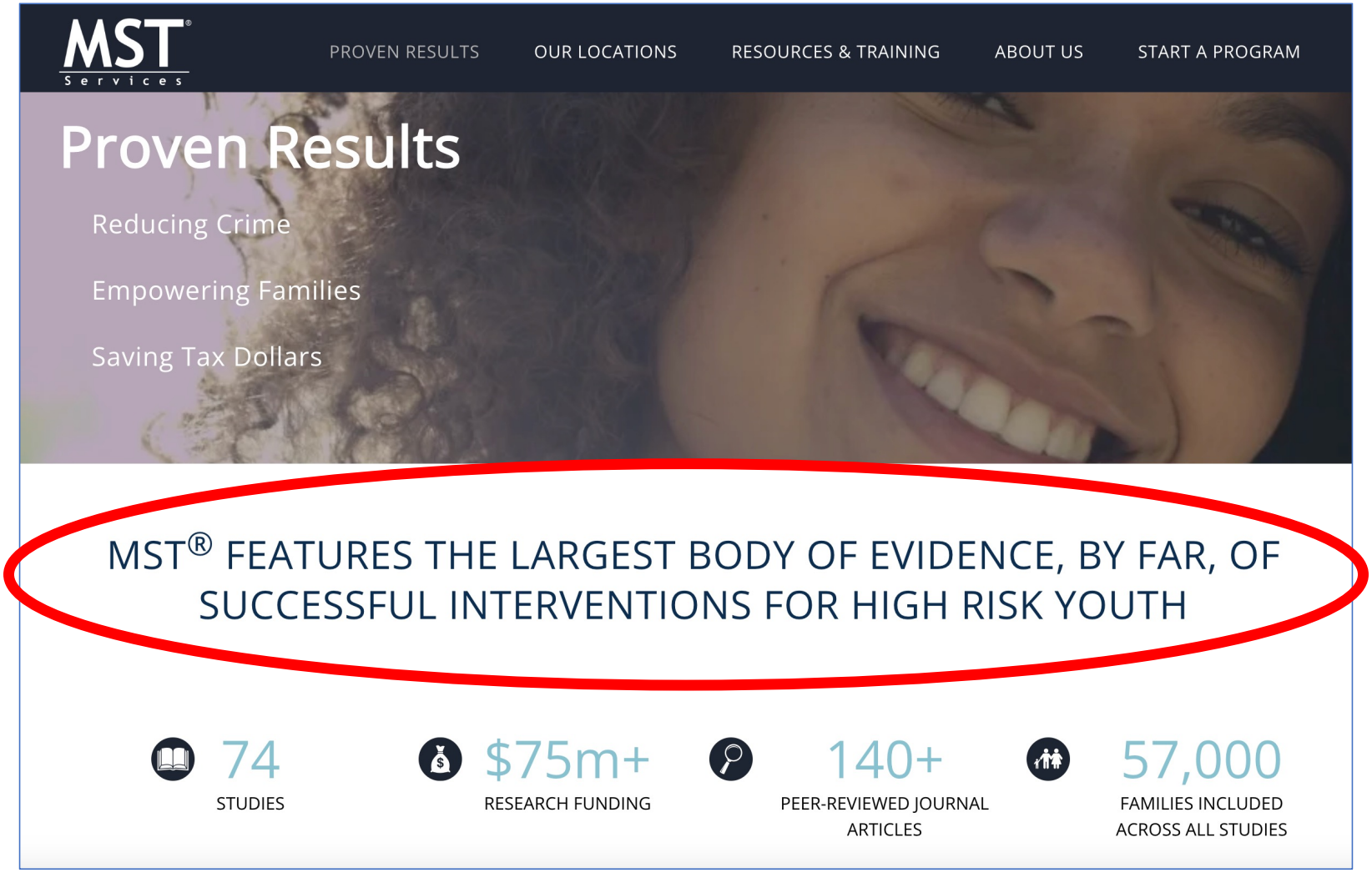
Available evidence shows that MST reduces rates of out-of-home placement and arrest or conviction

Multisystemic Therapy® (MST) is…

- A prominent “evidence-based” program
- Intensive, short-term, family- and community-based treatment
- For families of youth involved in juvenile justice, mental health, and/or child welfare service systems
- Primary goals:
  - Reduce crime and delinquency
  - Reduce out-of-home placements of youth (detention, hospital, foster care)
  - Improve youth and family functioning
- Proponents claim that MST has “consistent, positive effects” on primary outcomes across populations, problems, settings, and over time (Kazdin, 1998).
  - Aim of our review was to test this hypothesis.
- MST is licensed and supported by a for-profit consulting firm, MST Services LLC



https://www.mstservices.com/



https://www.mstservices.com/proven-results

Systematic review of effects of MST (Littell et al., 2021)

Included 23 randomized controlled trials (RCTs) conducted 1983 to 2020

- Compared licensed MST programs to treatment as usual (TAU) and/or other active treatments in 6 high-income countries (total N ~ 4,000 families)

- 13 RCTs conducted by MST program developers in the USA
- 10 RCTs conducted by independent investigators (3 in the USA, 3 in the UK, 1 each in Canada, Sweden, the Netherlands, and Norway).

Like most RCTs, none of these studies used probability samples.

Sample characteristics, intervention characteristics, settings, and risks of bias varied across studies. Most RCTs did not describe these characteristics well.

Results of our systematic review

Effects of MST were not consistent across studies, outcomes, or endpoints. Some different results for studies in USA vs other countries.

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

The logic of generalization

1. **Consensus-based rubrics**
2. Probability theory and inferential statistics
3. Proximal similarity and other principles

1. Consensus-based rubrics for generalizations from reviews

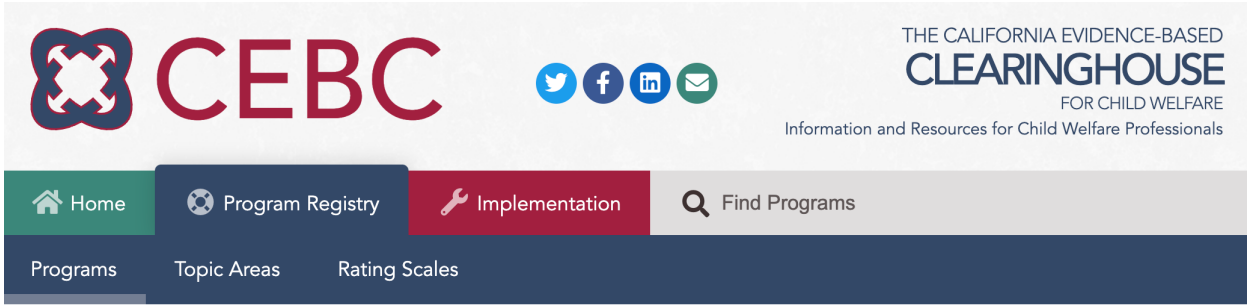
Common short-cuts used to characterize a body of evidence

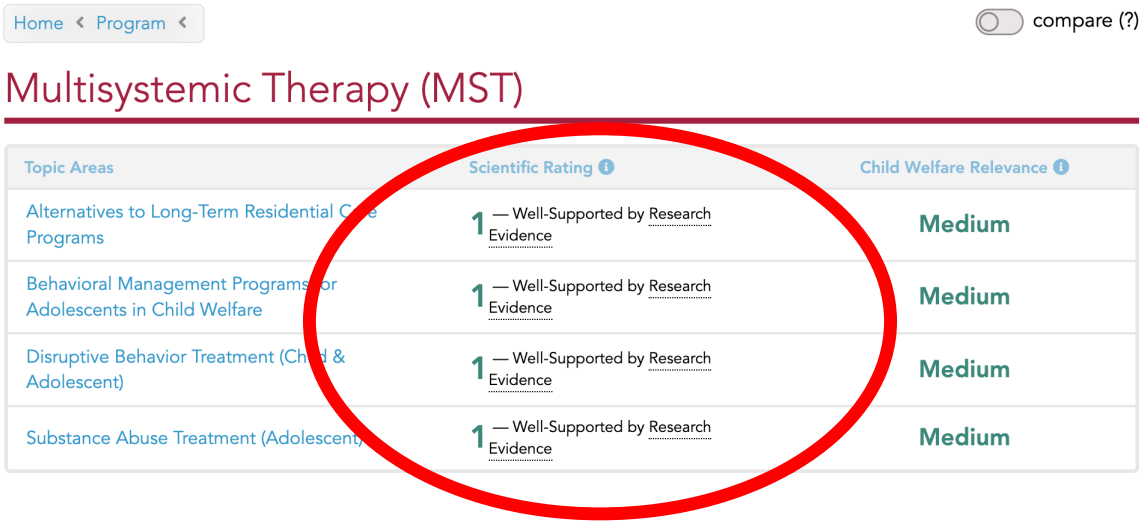
a. Thresholds set by USA government agencies and clearinghouses to rate interventions

“Effective” if there is more than one study

https://crimesolutions.ojp.gov/ratedprograms/192

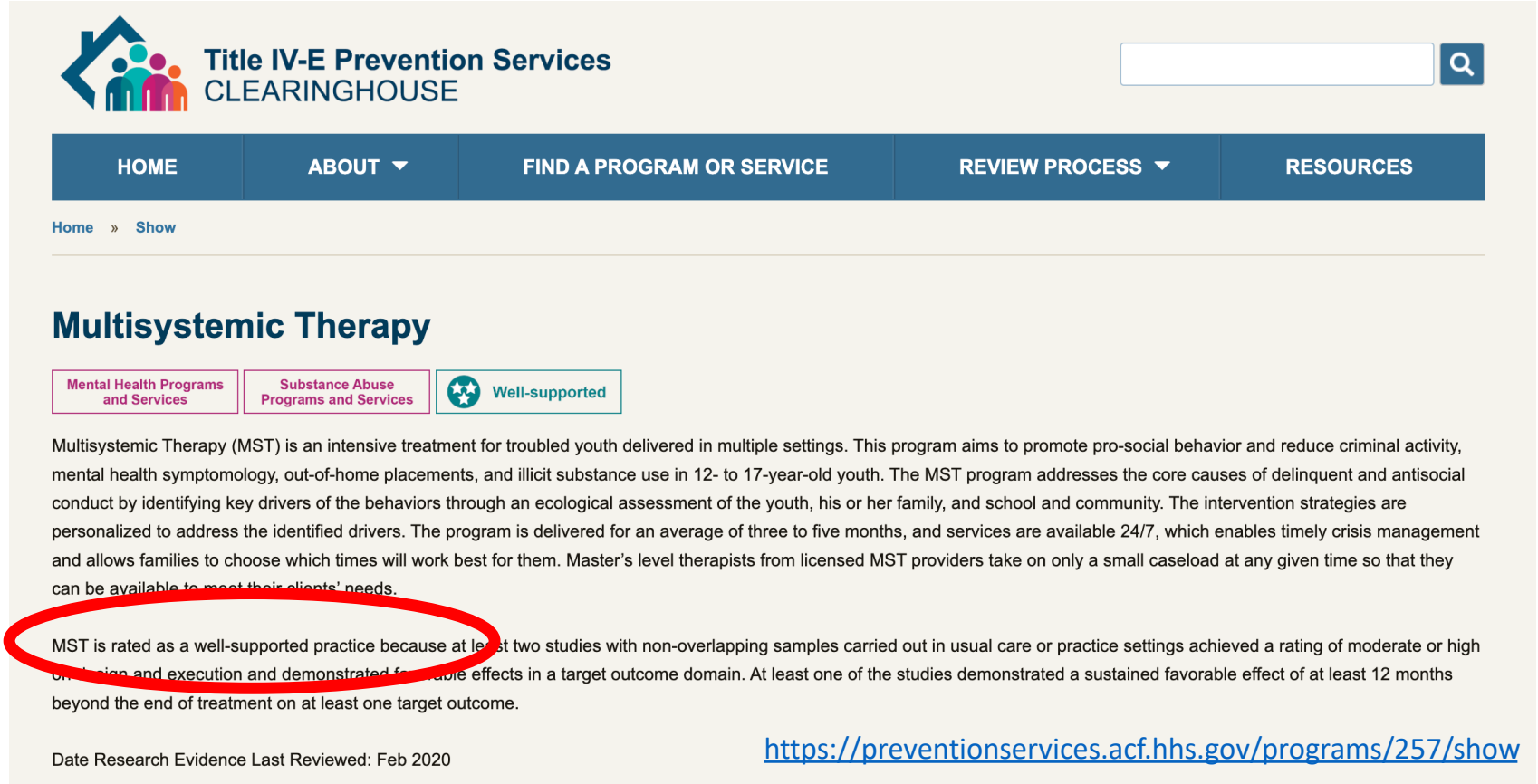
“Well supported” if effects last > 1 year





https://www.cebc4cw.org/program/multisystemic-therapy/detailed

“Well supported” if more than one study…



Title IV-E Prevention Services CLEARINGHOUSE

HOME | ABOUT | FIND A PROGRAM OR SERVICE | REVIEW PROCESS | RESOURCES

Home » Show

Multisystemic Therapy

Mental Health Programs and Services | Substance Abuse Programs and Services | Well-supported

Multisystemic Therapy (MST) is an intensive treatment for troubled youth delivered in multiple settings. This program aims to promote pro-social behavior and reduce criminal activity, mental health symptomology, out-of-home placements, and illicit substance use in 12- to 17-year-old youth. The MST program addresses the core causes of delinquent and antisocial conduct by identifying key drivers of the behaviors through an ecological assessment of the youth, his or her family, and school and community. The intervention strategies are personalized to address the identified drivers. The program is delivered for an average of three to five months, and services are available 24/7, which enables timely crisis management and allows families to choose which times will work best for them. Master’s level therapists from licensed MST providers take on only a small caseload at any given time so that they can be available to meet their clients’ needs.

MST is rated as a well-supported practice because at least two studies with non-overlapping samples carried out in usual care or practice settings achieved a rating of moderate or high on design and execution and demonstrated favorable effects in a target outcome domain. At least one of the studies demonstrated a sustained favorable effect of at least 12 months beyond the end of treatment on at least one target outcome.

Date Research Evidence Last Reviewed: Feb 2020

https://preventionservices.acf.hhs.gov/programs/257/show

1. Consensus-based rubrics for generalizations from reviews

Common short-cuts used to characterize a body of evidence

a. Thresholds set by USA government agencies and clearinghouses to rate interventions:

- "Effective" if 2+ RCTs showed some positive results (US NIJ Crime Solutions)
- "Well supported" if some positive results last > 1 year (CEBC4CW)
- "Well established" with 2+ independent RCTs (JCCAP, McCart & Sheidow, 2016)
- "Ready for broad dissemination" if previous conditions are met and there is a treatment manual, training and technical assistance, fidelity monitoring tool (Gottfredson et al., 2015).

MST meets all of these criteria.

1. Consensus-based rubrics for generalizations from reviews

Common short-cuts used to characterize a body of evidence

a. Thresholds set by USA government agencies and clearinghouses to rate interventions

b. The **pooled effect size** (weighted average across studies) used as the best estimate of likely effects.

- Example: UK Youth Endowment Foundation (YEF) online Toolkit

YOUTH ENDOWMENT FUND

WHAT WORKS TO PREVENT VIOLENCE?

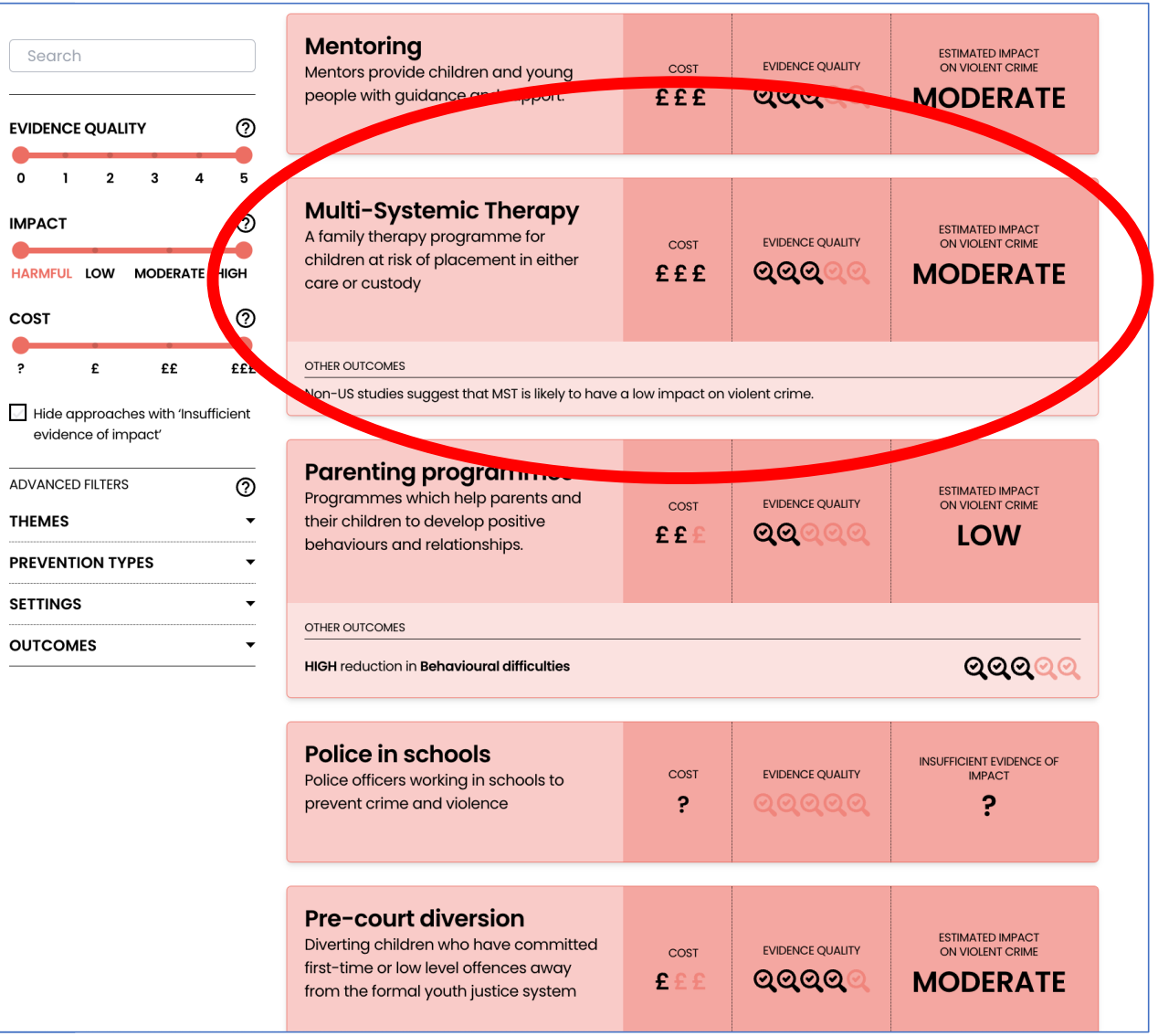
YEF Toolkit

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 does “moderate impact”mean? (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
- Incorrect, the overall risk difference is **3%** (p>.05)
- The overall effect is not statistically different from zero (no effect)

We found that a 9% reduction in offending in the USA and 2% reduction elsewhere (p>.05).

- An overall average may have no real meaning/relevance anywhere in the world.
- Tyranny of the mean effect size

Consensus-based rubrics: Limitations

- Conflate internal validity and external validity
- Thresholds are low, encouraging over-generalization from weak evidence.
- Mean effect size are relatively uninformative
  - Ignore heterogeneity of results (confidence intervals, prediction intervals, systematic differences between studies)

The logic of generalization

1. Consensus-based rubrics
2. **Probability theory and inferential statistics**
3. Proximal similarity and other principles

2. Probability theory and inferential statistics

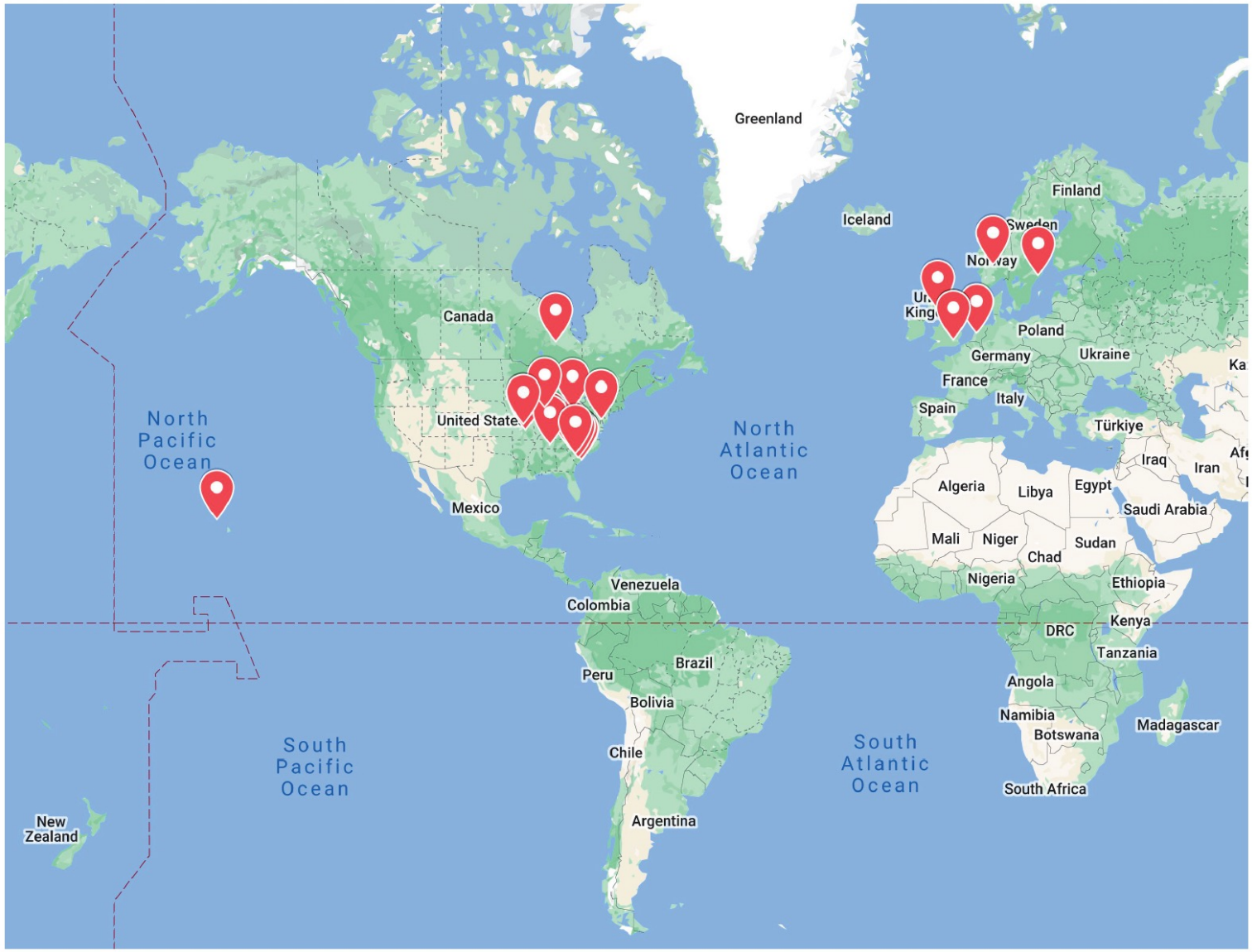
- Probability samples are the “gold standard” for generalization (Tipton et al., 2017)
  - Support use of inferential statistics to make generalizations to a larger, target population

- Types of studies included in systematic reviews of intervention effects (RCTs and credible QEDs) **rarely use probability samples**

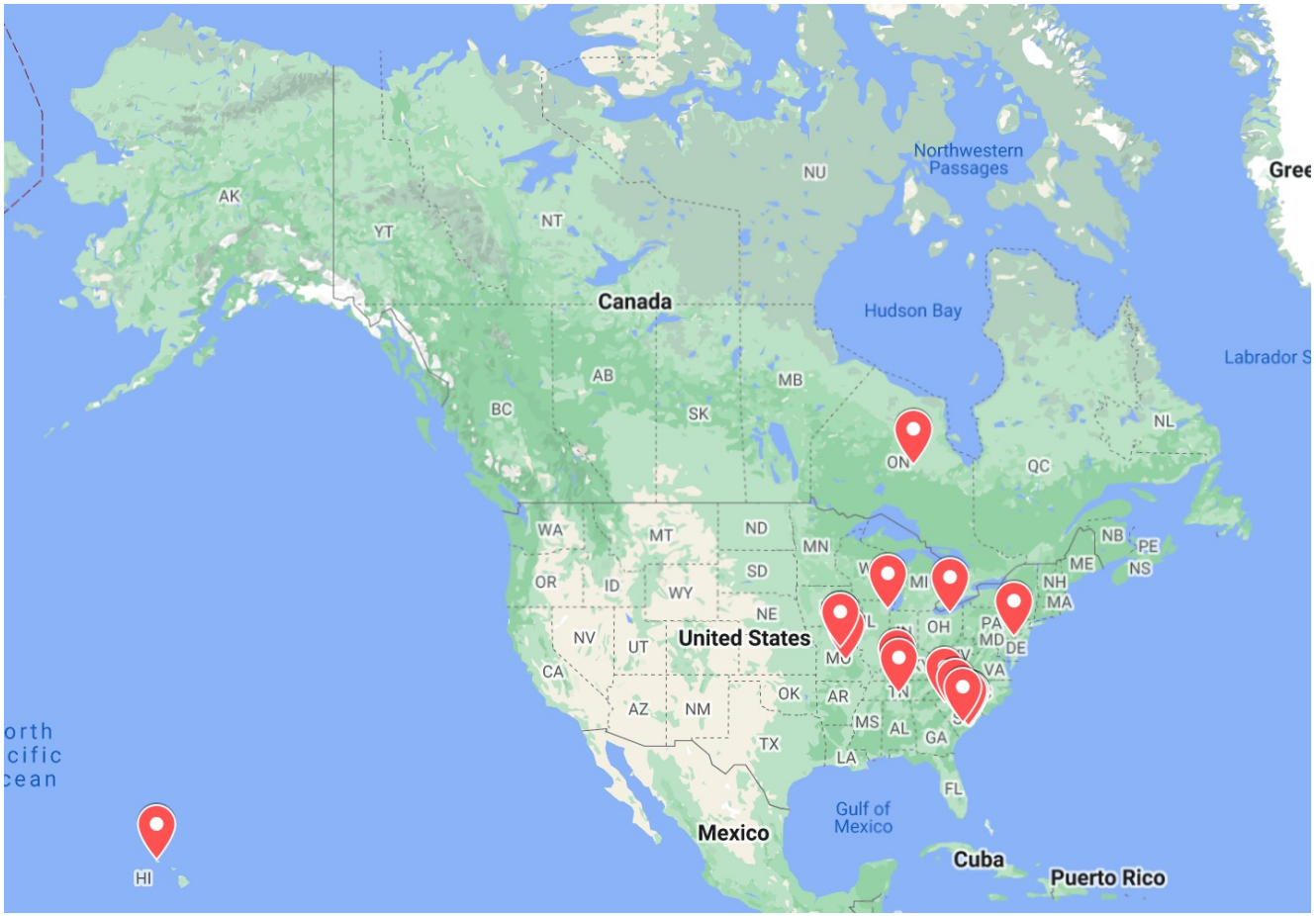
23 MST trials in six (6) high-income countries

Number of studies (k) per country:

16 USA (incl Hawaii)
3 UK
1 Canada
1 Sweden
1 Norway
1 Netherlands



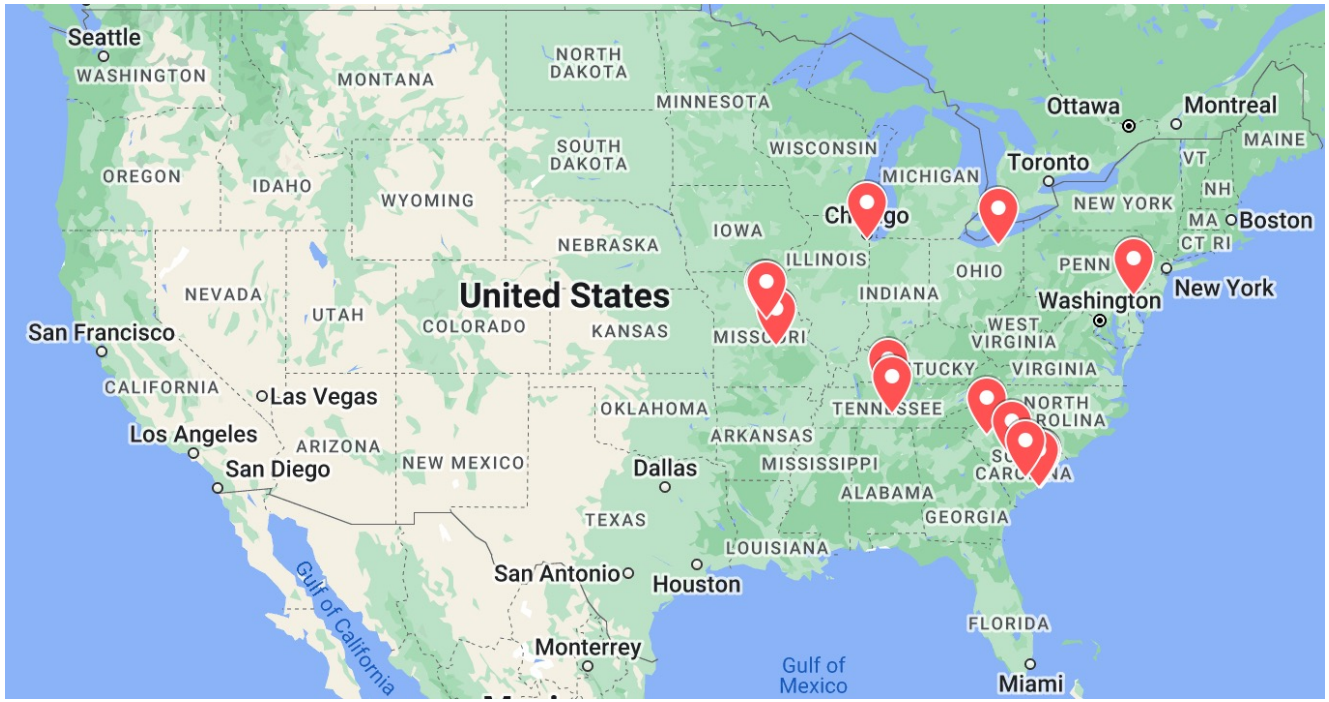
17 MST trials in USA and Canada



Number of studies (k) per state/province:

6 South Carolina
4 Missouri
2 Tennessee
1 Illinois
1 Ohio
1 Delaware
1 Hawaii
1 Ontario

15 MST trials in six (6) mainland USA states



Number of studies (k) per state:

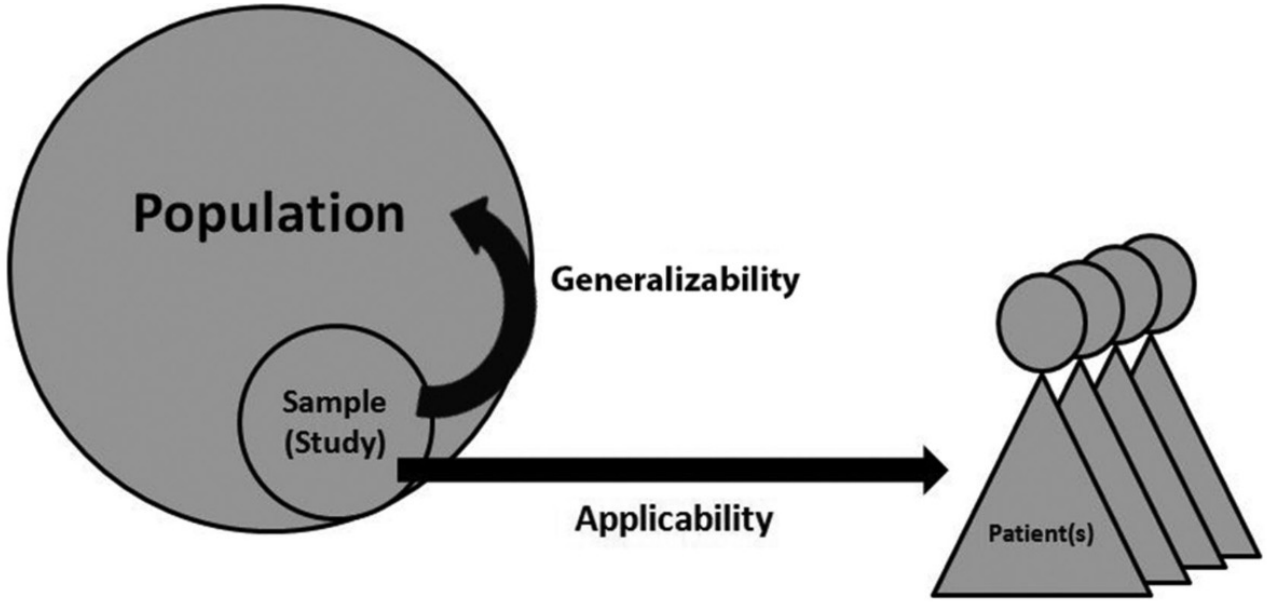
6 South Carolina
4 Missouri
2 Tennessee
1 Illinois
1 Ohio
1 Delaware

Systematic review of a nonprobability sample of studies: Limitations

- Programs that have been **studied** are not representative of any larger population of programs
  - Programs studied with RCTs are not representative of programs that have been studied with other methods

- Results are not **generalizable** to any larger target population

Probability theory provides little/no basis for generalizability or applicability of results of systematic reviews



Source: Murad et al., 2018

The logic of generalization

1. Consensus-based rubrics
2. Probability theory and inferential statistics
3. **Proximal similarity and other principles**

3. Proximal similarity and other principles for generalization from highly localized, nonrepresentative studies

Building on work of Cook (1990) and Shadish (1995)

Applying principles developed for use with experiments and ethnographies to SRs

American Journal of Community Psychology, Vol. 23, No. 3, 1995

Commentaries

The Logic of Generalization: Five Principles Common to Experiments and Ethnographies

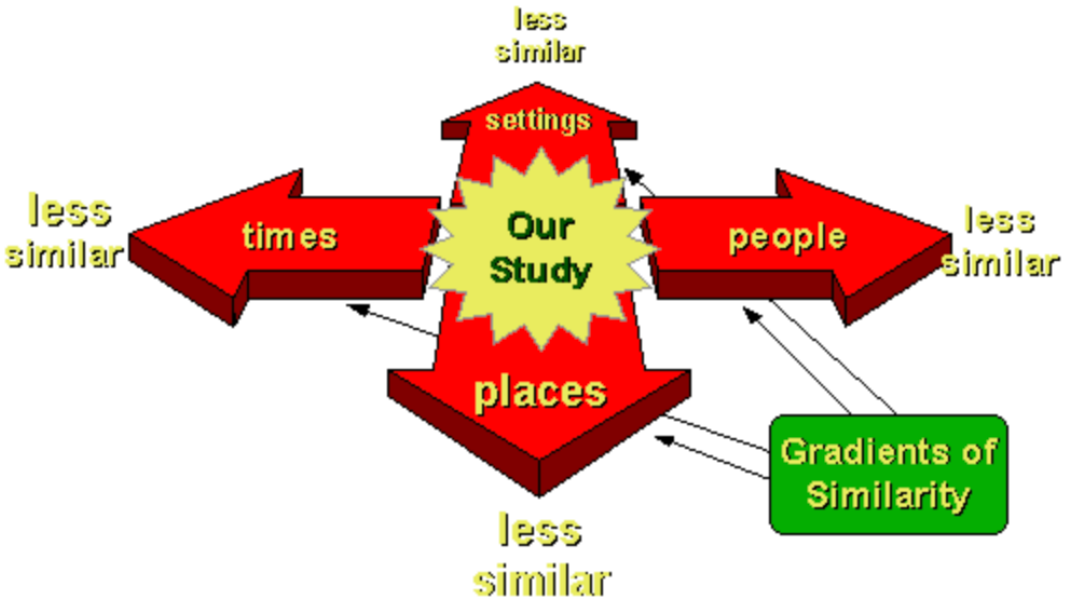
William R. Shadish[1]
University of Memphis

Both experiments and ethnographies are highly localized, so they are often criticized for lack of generalizability. The present article describes a logic of generalization that may help solve such problems. The logic consists of five principles outlined by Cook (1990): (a) proximal similarity, (b) heterogeneity of irrelevancies, (c) discriminant validity, (d) empirical interpolation and extrapolation, and (e) explanation. Because validity is a property of knowledge claims, not methods, these five principles apply to claims about generalization generated by any method, including both ethnographies and experiments. The principles are illustrated using Rizzo and Corsaro's interesting ethnographies as examples.

KEY WORDS: experiments; ethnographies; generalization; logic.

3-1. Proximal 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).



Could use concept mapping to identify different points on gradients of similarity (e.g., more/less relevant populations)

But may need to focus on **subgroups** of studies most similar to target application(s) (which results are most relevant for France? USA vs elsewhere?)

3-2. Heterogeneity of 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).

- Effects of MST are <u>not</u> consistent across studies, settings, outcomes, and endpoints (time) (Littell et al., 2021).
- This informs – reduces -- our confidence in the generalizability of results of MST.

3-3. 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).

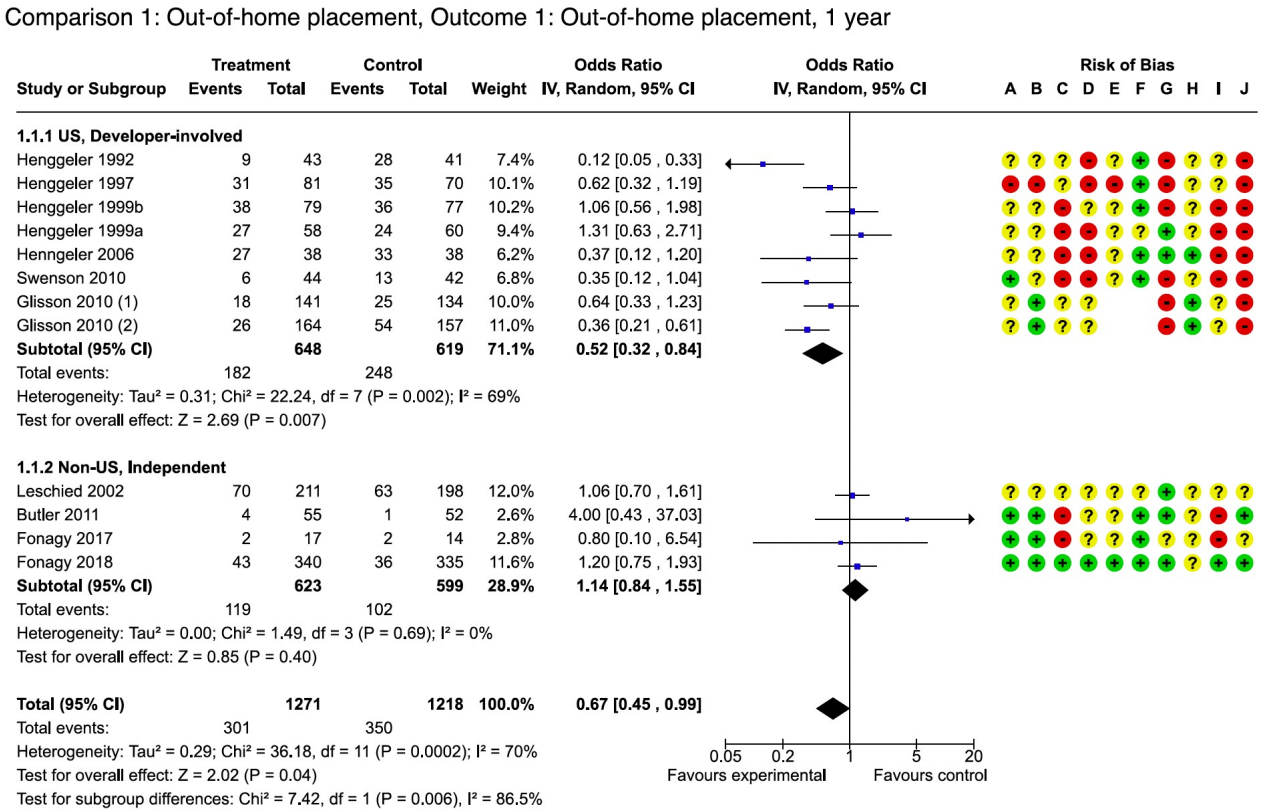
- Proponents claim that **adherence** to MST program principles is responsible for better outcomes.
  - But the MST Therapist Adherence Measure (TAM) lacks face validity and content validity, because it taps *other constructs* (client engagement, client satisfaction, relationship/alliance formation) that predict outcomes.
  - There are no studies that show that the TAM adherence measure successfully discriminates between MST and other treatments.
- Implementation of MST is **confounded** with other variables:
  - MST cases often receive more **time and attention** than control cases. MST therapists get more **training and supervision** than therapists who provide services to control groups (Littell et al., 2021).
- There is no convincing evidence for discriminant validity in MST studies.
- This informs – reduces – our confidence in the generalizability of results.

3-4. Empirical 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).

- MST review shows that positive effects are more likely in studies conducted in the USA, by program developers, using weaker research methods.
- But results are heterogeneous within subgroups of studies formed by these variables:
  - There are unexplained variations within the USA, and also among studies conducted outside of the USA.

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



3-5. 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 processed (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).

- Differences between the USA and other countries could be explained by:
  - Greater **program developer** involvement in USA studies: Better implementation? Implicit allegiance bias? Conflict of interest?
  - Higher **risks of bias** in USA studies: Weaker research methods (e.g., selective reporting of outcomes), inflated effect sizes?
  - **Contextual differences**: Higher "base rates" (likelihood of arrest/conviction, out-of-home placements) in the USA, ceiling/floor effects?

Context matters: Base rates

Harder to reduce events that are relatively rare

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

3-5. Explanation

- Possible explanations:
  - Quality of implementation
  - Variables confounded with treatment (additional time, attention, training, supervision)
  - Conflicts of interests
  - Implicit allegiance bias
  - Strength of research methods
  - Contextual differences
  - Ceiling/floor effects
- Competing explanations cannot be unraveled because these variables (potential moderators) are **confounded**.
- Inability to explain results informs – reduces – our confidence in ability to generalize.

Evaluation of three logical frameworks

Logical framework	Pros	Cons
1. Consensus-based rubrics	Easy to translate into policy-relevant metrics and conclusions	Illusions of precision and certainty, lead to over-generalization, potentially misleading
2. Probability theory, inferential statistics	Strong theoretical and statistical foundations for use with probability samples	Not relevant for syntheses of data from nonprobability samples (i.e., most reviews of intervention effects)
3. Proximal similarity and other principles for use with nonprobability samples	Potentially useful, especially for applications	Complex, inaccessible language? Few worked examples/illustrations

Summary: Logical frameworks for generalization from SRs

Logical framework	IRL (in real life)
1. Consensus-based rubrics	**What we're using now** (deeply flawed, common approaches)
2. Probability theory, inferential statistics	**What we think we're using now** (wishful thinking)
3. Proximal similarity and other principles for use with nonprobability samples	**What we're probably stuck with** (most realistic)

Conclusions

- Validity is a property of *inferences* that are based on data and methods, not a property of methods or data (Shadish, Cook, & Campbell, 2002)
- Current consensus-based rubrics (short cuts) and probability theory are of limited usefulness in generalizing from most systematic reviews.
- Proximal similarity and other principles may help us think through issues of generalizability from SRs.
  - Can inform the kinds of inferences we can make and confidence in our ability to generalize (or not) from SRs.
  - Regarding MST: We are not confident in generalizations from our SR. This conclusion is informed by principles articulated by Shadish (1995).

Directions for further work

- Primary studies
  - Improve reporting of characteristics of participants, interventions, and settings to support assessments of external validity.
- Systematic reviews and meta-analyses
  - Explore potential sources of heterogeneity (subgroup and moderator analyses).
  - Use proximal similarity and other principles to explore potential limitations on generalizability and applicability.
- Knowledge brokers
  - Stop using consensus-based rubrics that over-estimate precision, certainty, and generalizability.
  - Focus on proximal similarity and other principles to help decision makers apply results of SRs in specific policy/practice contexts.

Merci beaucoup!

jlittell@brynmawr.edu
jhlittell@gmail.com