The Logic of Generalization from Systematic Reviews to Policy and Practice

Julia H. Littell
Bryn Mawr College

Viewed as a powerful tool for generalized causal inference, systematic reviews can provide more robust empirical evidence than any single study or nonsystematic review. For this reason, proponents have argued that systematic reviews should be the preferred source of empirical evidence for policy and practice decisions, or "the basic unit of knowledge translation" (Grimshaw et al., 2012, p. 3). Unfortunately, the logic of generalization from systematic reviews to policy formulation is woefully underdeveloped. In this paper, I aim to address this gap by identifying problems and principles of generalizing from systematic reviews to policy and practice problems, and by applying these ideas in a worked example.

In the realms of knowledge transfer and implementation science, we hear repeated laments about the "know-do" gap; the refrain is: we know what to do, we just need to do more of it. Much attention has been paid to how to disseminate information about "what works" and how to implement effective interventions. Systematic reviews are seen as ideal starting points for this process. But knowledge transfer and implementation efforts tend to oversimplify results of systematic reviews (which do not produce dichotomous "this works, that doesn't" answers) and bypass issues of external validity altogether.

Probability samples are the "gold standard" for generalization (Tipton et al., 2017). Drawn at random from a well-defined study population, the logic of generalization from large-enough probability samples is relatively simple: results should generalize to the study population and not beyond that population.

Systematic reviews of intervention effects prioritize studies that support causal inferences, namely randomized controlled trials (RCTs) and strong quasi-experimental designs, which are rarely based on probability samples. Systematic reviews and meta-analyses build statistical power and may increase generalizability by including studies conducted in multiple settings with different samples. This supports broader analyses of trends and potential moderators. But systematic reviews of intervention effects tend to be based on convenience samples of nonrepresentative studies and, strictly speaking, a non-probability (grab-bag) sample of studies is not representative of any larger population.

As the logic of probability sampling is generally not available for use with systematic reviews, we turn to other principles for generalization. Shadish, Cook, and Campbell (2002) reminded us that validity is a property of knowledge claims, it is not a property of methods. External validity is not a property of probability samples. Shadish (1995) articulated five principles for generalization common to highly localized, nonrepresentative studies (experiments and ethnographies). In contrast to the logic of probability sampling, which is used to make inferences to a larger, well-defined population, these principles are used to assess how likely it is that results will apply to various groups or situations, based on their similarities and differences to study samples and circumstances (see Table 1).

[Insert Table 1 about here]

In this paper, I explore potential uses of these principles in formulating credible generalizations from systematic reviews. I use as a working example our recent systematic review and meta-analysis of effects of Multisystemic Therapy (MST) for families of youth with social, emotional, and behavioral problems (Littell et al., 2021). MST is an intensive, short-term, family-based intervention aimed at reducing out-of-home placements of youth, crime and delinquency, and other social and emotional problems among youth. MST has been implemented in 34 US states and 15 countries.

Our systematic review of effects of MST included 23 RCTs; of these, 13 were conducted by program developers in the USA and 10 RCTs were conducted by independent investigators (three in the USA, three in the UK, and one each in Canada, Sweden, the Netherlands, and Norway). None of these studies used probability samples. Sample characteristics, comparison conditions, and risks of bias varied across these studies. Effects were not consistent across studies, outcomes, or endpoints. There was some evidence that MST reduced the likelihood of out-of-home placement and re-arrest in the USA but not in other countries. Program developer involvement, USA location, and high risks of bias are all associated with “better” outcomes, and these moderators are highly confounded.

Proximal similarity. At first glance, the principle of proximal similarity presents us with difficulties: It is hard to imagine a program, population, and setting that mirrors the assortment of participants, comparison conditions, and settings found across 23 MST trials. Estimates of the overall mean effects are not generalizable to any country or countries. Results of this systematic review are not generalizable to MST programs, because the review does not include a representative sample of licensed MST programs; it includes programs that MST developers and others decided to study with RCTs. We find greater clarity in the subgroup and moderator analyses, which suggest that MST is more likely to reduce out-of-home placements and re-arrests in the USA than in other countries. But, as we will see, the reasons for this are unclear.

Heterogeneity of irrelevancies. A major aim of our review was to test the repeated assertion that the effectiveness of MST has been demonstrated “across problems, therapists, and settings. This shows that the treatment and methods of decision making can be extended and that treatment effects are reliable” (Kazdin 1998, pp. 27–28). This assertion about the heterogeneity of irrelevancies does not hold up in our review. That is, results do not “hold over variations in persons, settings, treatments, outcome measures, and times that are presumed to be conceptually irrelevant” (Shadish, 1995, p. 424).

Discriminant validity. Proponents have claimed that it is adherence to MST principles and methods that is responsible for better outcomes, not other factors. But adherence measures tap other constructs that have been shown to relate to outcomes (e.g., client engagement, client satisfaction, and alliance formation). There are no studies that show that the MST Therapist Adherence Measure (TAM) discriminates between MST and other treatments.

Empirical interpolation and extrapolation. Here again, we can capitalize on the wide variation in effects across studies to try to identify situations in which MST is more or less likely to be effective. But the review gives us little to go on, beyond the finding that MST appears to reduce out-of-home placements and arrests in the USA and not in other countries.

Explanation. If we could explain the differences in results in the USA and other countries, we could generalize more confidently. But there are competing explanations for these findings that cannot be unraveled because of confounded moderators. Greater program developer involvement and higher risks of bias in the USA studies could mean that results were affected by: allegiance bias, the quality of program implementation, and/or the integrity of the research methods. Other explanations are possible as well: Youth in the USA are more likely to be removed from their families and more likely to be arrested that youth in other countries. In fact, placement and arrest rates were much lower in the comparison groups in studies conducted outside of the USA than from those in the USA. Thus, there may be a ceiling effect on these outcomes outside of the USA.

In conclusion, there is very little to say about the generalizability of effects of MST based on our systematic review. Our prediction intervals indicated that future studies are equally likely to find positive or negative effects of MST on all outcomes; in other words, future results are not predictable. This is in stark contrast to the ways in which these results have been characterized by others. For example, the Youth Endowment Fund, a prominent knowledge broker in the UK, estimates that "based on international studies, the review estimates that MST reduces violence by 16% and offending by 17%" (https://youthendowmentfund.org.uk/toolkit/multi-systemic-therapy-2/); presented without confidence intervals, these estimates suggest that effects are consistent, when they are not.

The principles outlined by Shadish (1995) can help us think more critically about generalizing from systematic reviews. To do this well, we will require higher quality primary studies, better systematic reviews, and more careful knowledge brokers. Implications for future research and policy are considered, including the need to shift attention from "know-do" gaps, and focus on the very real gaps in current knowledge, and how to build legitimate bridges between research and policy. Examples from Canada and Sweden illustrate how different policy makers have made different (but equally defensible) decisions about MST, based on results of our systematic review.

References

Grimshaw, J. M., Eccles, M. P., Lavis, J. N., Hill, S. J., & Squires, J. E. (2012). Knowledge translation of research findings. *Implement Science*, *7*(1), 50.

Kazdin, A. E., & Weisz, J. R. (1998). Identifying and developing empirically supported child and adolescent treatments. *Journal of Consulting and Clinical Psychology, 66*(1), 19–36. https://doi.org/10.1037/0022-006X.66.1.19

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

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.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, *41*(5), 472–505. https://doi.org/10.1177/0193841X16655665

Table 1: Five principles of generalization (Shadish, 1995)

Principle	Explanation (direct quotes from Shadish, 1995, pp. 424-426)
Proximal similarity	We generalize most confidently to applications where treatments, settings, populations, outcomes, and times are most similar to those in the original research.
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.
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.
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.
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.