Chapter 9
Generalizations from Meta-analysis

Abstract This chapter discusses the kinds of inferences and generalizations we can make from a meta-analysis. The chapter reviews the framework outlined by Shadish et al. (2002) for meta-analysis, and provides examples from two recent syntheses that had an influence on policy.

9.1 Background

What kinds of decisions can we make from a meta-analysis? Are we justified in making policy decisions from the results of a meta-analysis about implementation of an intervention, such as the use of systematic phonics instruction? What about decisions of a more personal nature – such as should I have a mammography annually in my forties? These are questions asked about meta-analyses from policy makers, practitioners, and consumers of this information. This chapter reviews the basis whereby we can make inferences from a meta-analysis, and the kinds of inferences that can be supported. It also argues for the transparency that reviewers of evidence should provide so that the results of systematic reviews can be used appropriately.

As many researchers have pointed out, the results of meta-analyses are observational. Reviewers cannot manipulate the kinds of methods used, or the participants in the sample, and thus cannot fulfill the requirements of an experimental study that aims to identify causes. In a meta-analysis, we do not have the ability to assign the conditions of the study. The studies already exist, and use a variety of procedures, methods, participants, in a variety of settings. Despite this fact, we still want to make decisions about the types of interventions that are most effective, or about what personal choice I should make about my own health. Given the nature of meta-analysis, we cannot use statistical reasoning as we do in a randomized controlled trial to reach a causal inference. Instead, we need a different basis for arguing about cause.

T.D. Pigott, *Advances in Meta-Analysis*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4614-2278-5_9, © Springer Science+Business Media, LLC 2012

As Matt and Cook (2009) and Shadish et al. (2002) point out, causal inferences can be supported in a meta-analysis using a logic or basis different from classic arguments about cause. These researchers argue that the warrant for making causal claims from a meta-analysis depends on ruling out threats to the validity of that inference. In other words, we argue a causal claim from a meta-analysis by systematically addressing all other plausible explanations for the causal relationship we propose.

Shadish et al. outline how we can proceed by arguing from the five principles of generalized causal inference: (1) surface similarity, (2) ruling out irrelevancies, (3) making discriminations, (4) interpolation and extrapolation, and (5) causal explanation. Causal inferences from a meta-analysis require examining the conditions and methods used across studies. As Shadish et al. argue, the number of methods and conditions represented across studies in a meta-analysis allows us to have more evidence about how an intervention or a relationship varies than in a single study that cannot include all of these conditions.

This chapter will use the Preventive Health Services (Nelson et al. 2009) report on breast cancer screening and Ehri et al. (2001) meta-analysis on the effects of systematic phonics instruction to illustrate the five principles of generalized causal inference described by Shadish et al. In the discussion that follows, I use Cronbach (1982) acronym, UTOS, as a shorthand for the generalizations we want to make from the meta-analysis. UTOS stands for the Units (persons) who receive the treatment and to whom we wish to generalize, the Treatments in the study and those treatments we want to generalize about, the Observations (measures) used in the study and those we wish to generalize to, and the Settings where the study takes place, and settings where we want to generalize these findings. Below is an introduction to the report on breast cancer screening as well as to the work of the National Reading Panel (2000) followed by a discussion of each of the five principles of generalized causal inference in the context of these reviews.

9.1.1 The Preventive Health Services (2009) Report on Breast Cancer Screening

In 2009, the United States Preventive Health Services (USPHS) released an update of their previous synthesis on studies focusing on the outcomes of breast cancer screening. The update included two new trials, the Age trial from the United Kingdom (Moss et al. 2006), and an update of the data from the Gothenburg trial conducted in Sweden (Bjurstam et al. 2003). The Age trial specifically targeted outcomes in women aged 40–49, and resulted in the USPHS revising their original (US Preventive Services Task Force 2002) recommendations for this specific group of women. Essentially, the recommendations were that the evidence no longer supported routine, annual mammography for women aged 40–49, given the risk of false positive results and over diagnosis. The release of the results coincided with the healthcare reform debates that were occupying the US Congress and the media.

In some instances, the report's recommendations were linked to the healthcare reform debate (Woolf 2010), one clear example of how the results of this review were misinterpreted. There have been many commentaries in the media and in the medical literature discussing what conclusions can be drawn, and how women should respond to the report.

9.1.2 The National Reading Panel's Meta-analysis on Learning to Read

The U. S. Congress in 1997 asked that a panel be convened to review research on strategies to teach children to read. The panel, appointed by the National Institute of Child Health and Human Development (NICHD) and the Department of Education conducted a series of systematic reviews of the evidence. One of these subgroup analyses is a meta-analysis conducted by Ehri et al. (2001) on the effects of systematic phonics instruction on students' ability to read words. The National Panel's report had wide-spread influence, contributing to the language in the No Child Left Behind Act (Allington 2006) that calls for the use of research-based teaching strategies. The National Panel's report was widely criticized on various grounds (Camilli et al. 2006; Hammill and Swanson 2006; Pressley et al. 2004), with many researchers questioning how well the results could generalize to real classrooms. Below I outline the five principles of generalized causal inference, using both the Ehri et al. meta-analysis and the breast cancer screening study as examples.

9.2 Principles of Generalized Causal Inference

9.2.1 Surface Similarity

Surface similarity was first discussed by Campbell (1957) in terms of construct validity. We can more safely apply a generalization from one measure to another measure that is based on a similar construct. In the context of meta-analysis, we can apply a finding from a meta-analysis to those UTOS that are represented in the studies included in the synthesis. Conversely, we may caution about generalizing from a meta-analysis to a context that is not represented in the meta-analysis. One criticism of the breast cancer screening meta-analysis was that the studies in the meta-analysis did not include a sufficient number of African-American women, and thus the results should not be generalized to this group of women. For example, Murphy (2010) cautions that clinicians applying the findings of the report need to keep in mind that African-American women have a higher risk of mortality from breast cancer, and women of Ashkenazi Jewish descent are at higher risk of genetically mediated breast cancer. The synthesis did not include studies with a large sample of these groups of women, and thus the synthesis provides no evidence about how breast cancer screening is related to the mortality from breast cancer for these two groups. In terms of other groups of women who might be at higher risk, such as women exposed to high levels of radiation, there is not enough specific information included about the characteristics of women in the trial to make generalizations about particular sub-groups. Our ability to examine surface similarity from the report itself is limited. The report does update findings about one particular group of women, those aged 40–49, since new evidence from the Age Trial (Moss et al. 2006) provides more direct evidence about this group. There were, however, no new trials that could provide insight for the screening of women over the age of 70, and thus the report does not revise the guidelines for women in this age group.

In the debate over the National Reading Panel's meta-analysis on systematic phonics, Garan (2001) questioned the use of measures of different reading outcomes as equivalent in the meta-analysis. The Ehri et al. meta-analysis used the construct of general literacy to include decoding regular words, decoding pseudowords, spelling words and reading text orally to name a few. To Garan, these measures are not sufficiently similar to each other to constitute a single construct. In the Ehri et al. review, the effect sizes for the different measures are reported separately though they are treated as measuring the effectiveness of programs on systematic phonics.

9.2.2 Ruling Out Irrelevancies

Related to surface similarity is the principle of ruling out irrelevancies. In order to generalize a finding to a set of UTOS that were not represented in the meta-analysis, we need to understand whether a given situation is similar to the ones represented in the meta-analysis, and what differences between our given situation and those in the meta-analysis are irrelevant to the findings. In the breast cancer screening review, one issue deemed irrelevant to mortality of breast screening is whether the mammography used film or digital technology. The research question guiding the review includes both of these mammography procedures, but does not provide a comparison of their effectiveness on mortality outcomes in the review. Thus, the reviewers conducted the review on the assumption that film and digital mammography lead to the same mortality rates. However, some researchers do raise issues about whether the method of screening is really irrelevant. For example, Berg (2010) presents evidence that magnetic resonance imaging (MRI) for women at high risk improves detection by 40% over mammography and ultrasound combined. The report does not compare outcomes using MRI versus film or digital technology. Murphy (2010) also suggests that to avoid higher rates of false positives, younger women should consider having their screening at facilities with radiologists that focus on breast imaging and that use digital technology. Here Murphy questions whether film versus digital technology is actually an irrelevant factor. It may not be possible to test empirically whether outcomes are different between film and digital mammography with the current evidence, so this may be an area that needs more research.

Camilli et al. (2006), in their review of the findings of the National Reading Panel on systematic phonics, note that the meta-analysis compares treatments that received various levels of systematic phonics with a no-treatment control. Camilli et al. argues that while the report's findings (as indicated in the title the Ehri et al. 2001) states that systematic phonics increases student reading achievement, the meta-analysis itself did not and could not examine the differences among the different types of systematic phonics programs represented in the sample of studies. Thus, we are not able to determine from this meta-analysis whether the difference among systematic phonics programs in the amount of phonics instruction is an irrelevant factor.

9.2.3 Making Discriminations

As Shadish et al. (2002) describe, we make discriminations in a meta-analysis about the conditions where the cause and effect relationship does not hold, or, in other words, for those persons, treatments, measures and settings where the findings are found not to apply. This principle is different from surface similarities in that it refers to the examination of moderators of a given cause and effect relationship. For example, Littell et al. (2005) has found that the reported effectiveness of multisystemic therapy for at-risk children varies as a function of the involvement of the researcher in the development of the intervention. Studies that were conducted by researchers other than the original developers have smaller effect sizes. We can think of this finding as discriminating about the conditions where the treatment is most effective. The breast cancer screening study is limited in its ability to make discriminations partly due to the lack of information about the backgrounds of the women involved in the studies, and partly due to the small number of trials (seven). Moderator analyses examining how the results might vary systematically among persons, treatments, measures and settings are not possible since there are only seven trials that meet the inclusion criterion. We do not have enough statistical power to make discriminations about the relative effectiveness of screening across different UTOS.

One finding from the Ehri et al. (2001) meta-analysis that was not subject to debate was that systematic phonics instruction did not appear as effective for older elementary school children as for those in kindergarten and first grade. This finding was based on a number of studies that included older children. In fact, Ehri et al. used simple moderator analyses to examine both grade and reading ability, finding that kindergartners and first graders at risk had the largest benefit from systematic phonics instruction. Children in 2nd through 6th grade had little benefit.

9.2.4 Interpolation and Extrapolation

Another interrelated principle is interpolation and extrapolation. In examining a causal claim from a meta-analysis, we need to specify the range of characteristics of UTOS where the cause and effect relationship applies. In a single, primary study, we are careful not to extrapolate to contexts outside of the ones represented in the study itself – a single study cannot provide much evidence about whether the findings hold outside of the UTOS used in the study. In some meta-analyses, we could have a wide range of persons, treatments, measures and settings represented across studies, and we can systematically examine whether the cause and effect relationship applies across the studies. One method for interpolating and extrapolating studies is to use modeling strategies with effect sizes, using meta-regression, for example, to see what combinations of characteristics of studies may find larger or smaller effect sizes. As described above, the breast cancer screening review does not include enough studies to model the range of possible study characteristics where the results do or do not apply. The breast cancer screening review does not make recommendations on the effects of screening on women older than 70 – the sample of studies simply does not provide evidence about this group, and the authors of the review do not extrapolate the results. One classic example of the use of modeling in this way is illustrated in Raudenbush and Bryk (1985). Using a random effects meta-regression model, Raudenbush shows that the effect size in the teacher expectancy studies drops off considerably when the induction of expectancy is performed after the teachers have known their students for 3 weeks or more.

One issue of extrapolation and interpolation raised in the systematic phonics meta-analysis relates to the nature of the phonics programs. As Camilli et al. (2006) explains, the reading treatment described in the literature can rarely be classified as including systematic phonics instruction versus less systematic phonics instruction as might occur in a classroom where phonics is only taught when needed. Underlying this criticism of the phonics meta-analysis is the question of whether the phonics treatment as described in these studies is implemented in a similar way in classrooms. Pearson (2004) raises this question in a history of the whole language movement prior to the National Reading Panel report; teachers were less involved and invested in the critiques around the National Reading Panel and No Child Left Behind than academics. The realities of systematic instruction of phonics in a classroom may not resemble the studies in the meta-analysis, and may also be difficult to classify.

9.2.5 Causal Explanation

The fifth principle is causal explanation. Though a meta-analysis may not include information about how an intervention works, Shadish et al. (2002) argue that with good theory, meta-analyses can contribute to our understanding about causes.

Causal explanation can be facilitated in a meta-analysis by breaking down the intervention reviewed into its component parts, and positing a theory about both the critical ingredients of an intervention and how those ingredients relate to one another. A meta-analysis can then focus on the parts of this theory of action, using effect size modeling to examine what components of the intervention are most strongly associated with the magnitude of the effect size. In addition, a logic model or theory of action can provide a map of what evidence exists in the literature about particular aspects of a mediating process, and where more studies are needed to provide insight into aspects of the model. In the breast cancer screening study, there are not enough studies to map out an elaborated logic model. However, there are areas in the social sciences that may have the potential of supporting this type of analysis.

Pressley et al. (2004) raise the issue of theory of action or model of reading that is implied by the National Reading Panel work. For Pressley et al., the Reading Panel focused on a set of skills that are related to reading but may be much more narrow than intended. Pressley et al. argue that the theory of reading underlying the National Panel work suggests that "beginning reading only requires instruction in phonemic awareness, phonics, fluency, vocabulary, and comprehension strategies" (p. 41). The criticism of the report may have been tied to this difference in theory of how reading develops in children. The report may not have emphasized enough that the meta-analyses examined component parts of an effective reading program, and were not intended to define a comprehensive reading curriculum.

9.3 Suggestions for Generalizing from a Meta-analysis

Both the breast cancer screening study and the meta-analysis on systematic phonics instruction captured much attention due to the characterization of their findings by various groups. In the breast cancer screening case, the findings appeared to contradict current practice (yearly mammography) particularly in women aged 39–50. The meta-analysis on systematic phonics stirred controversy since its findings were influential on subsequent education policy. The question is what can reviewers do to decrease potential for misinterpreting meta-analysis findings and misapplying them to policy and practice? One suggestion is based on the Cochrane Handbook's (Higgins and Green 2011) risk of bias tables. With the assistance of experts in the field of study, reviewers might attempt a summary of what aspects of UTOS in a given field appear to have enough evidence to make a recommendation, and where we have equivocal or no evidence. Table 9.1 below is an attempt at a table for the Ehri et al. (2001) work.

Table 9.1 is not complete, but may serve as a way to summarize where we do have evidence to take an action. For each element of UTOS, I indicate the level of evidence for particular generalizations from Ehri et al. Both Pearson (2004) and Pressley et al. (2004) mention the role of policymakers in using the results of the National Panel report in ways that went beyond the data gathered. We do need ways

Table 9.1 Outline of generalizations supported in Ehri et al. (2001)

Area	Evidence	Equivocal evidence	No evidence
Units	Adequate for K-1 graders at risk		Second language learners
Treatments		Differences in effectiveness among types of programs, and how much systematic phonics instruction is necessary	
Observations/ measures	Word reading and pseudo –word reading	General reading ability not well defined so that not clear that tests of comprehension are equivalent to tests of work reading	
Settings		No differences found among instructional delivery units of tutoring, small group or whole class	

to communicate complex findings to those who may use our reviews. Those who are interested in this book are by nature interested in meta-analysis and summarizing the evidence in an area, and thus we also must be as careful in how we describe what actually can be done with our results.

References

Allington, R.L. 2006. Reading lessons and federal policymaking: An overview and introduction to the special issue. *The Elementary School Journal* 107: 3–15.

Berg, W.A. 2010. Benefits of screening mammography. *Journal of the American Medical Association* 303(2): 168–169.

Bjurstam, N., L. Bjorneld, J. Warwick, et al. 2003. The Gothenburg breast screening trial. *Cancer* 97(10): 2387–2396.

Camilli, G., P.M. Wolfe, and M.L. Smith. 2006. Meta-analysis and reading policy: Perspectives on teaching children to read. *The Elementary School Journal* 107: 27–36.

Campbell, D.T. 1957. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* 54(4): 297–312.

Cronbach, L.J. 1982. *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Ehri, L.C., S. Nunes, S. Stahl, and D. Willows. 2001. Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research* 71: 393–448.

Garan, E.M. 2001. Beyond the smoke and mirrors: A critique of the National Reading Panel report on phonics. *Phi Delta Kappan* 87(7): 500–506.

Hammill, D.D., and H.L. Swanson. 2006. The National Reading Panel's meta-analysis of phonics instruction: Another point of view. *The Elementary School Journal* 107: 17–26.

Higgins, J.P.T., and S. Green. 2011. *Cochrane handbook for systematic reviews of interventions*. Oxford, UK: The Cochrane Collaboration.

Littell J.H., M. Campbell, S. Green, and B. Toews. 2005. Multisystemic therapy for social, emotional and behavioral problems in youth aged 10–17. *Cochrane Database of Systematic Reviews* (4). doi:10.1002/14651858.CD004797.pub4.

Matt, G.E., and T.D. Cook. 2009. Threats to the validity of generalized inferences. In *The handbook of research synthesis and meta-analysis*, ed. H. Cooper, L.V. Hedges, and J.C. Valentine, 537–560. New York: Russell Sage.

Moss, S.M., H. Cuckle, A. Evans, et al. 2006. Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: A randomised controlled trial. *Lancet* 386(9552): 2053–2060.

Murphy, A.M. 2010. Mammography screening for breast cancer: A view from 2 worlds. *Journal of the American Medical Association* 303(2): 166–167.

National Reading Panel. 2000. *Report of the National Reading Panel: Teaching chidren to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Rockvill: NICHD Clearinghouse.

Nelson, H.D., K. Tyne, A. Naik, C. Bougatsos, B. Chan, P. Nygren, and L. Humphrey. 2009. Screening for breast cancer: Systematic evidence review update for the U. S. Preventive Services Task Force (trans: Agency for Healthcare Research and Quality). Rockville, MD: U. S. Department of Health and Human Services.

Pearson, P.D. 2004. The reading wars. *Educational Policy* 18: 216–252.

Pressley, M., N.K. Duke, and E.C. Boling. 2004. The educational science and scientifically based instruction we need: Lessons from reading research and policymaking. *Harvard Educational Review* 74: 30–61.

Raudenbush, S.W., and A.S. Bryk. 1985. Empirical Bayes meta-analysis. *Journal of Educational Statistics* 10: 75–98.

Shadish, W.R., T.D. Cook, and D.T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

US Preventive Services Task Force. 2002. Screening for breast cancer: Recommendations and rationale. *Annals of Internal Medicine* 137(5 Part 1): 344–346.

Woolf, S.H. 2010. The 2009 breast cancer screening recommendations of the US Preventive Services Task Force. *Journal of the American Medical Association* 303(2): 162–163.