Drawing Generalized Causal Inferences Based on Meta-Analysis

Georg E. Matt

INTRODUCTION

Research syntheses are more and more used to inform decisionmakers about the effects of a particular policy or of different policy options. For instance, do substance abuse prevention programs in junior high schools reduce drug use in high school? Are random drug tests more effective than drug education programs in reducing drug use? Are social influence prevention programs more effective with boys than with girls?

In the language of Cook and Campbell (1979) and others, these questions involve causal relationships of two kinds: bivariate causal relationships and causal moderator relationships. In bivariate causal relationships, one is examining whether deliberately manipulating one entity (e.g., introducing a prevention program) will lead to variability in another entity (e.g., onset of drug use). In causal moderator relationships, one is interested in identifying variables that modify the magnitude or sign of a causal relationship (e.g., in the presence of peer counselors, prevention programs are more effective than in their absence).

Meta-analyses seek to draw conclusions about populations, classes, or universes of variables. This is different from primary studies in which, for instance, researchers examine the causal effects of a particular drug education curriculum in a particular school with students in a particular grade. Instead, meta-analyses seek to draw conclusions regarding a universe of persons (e.g., students in grades 4 to 12), a universe of interventions (e.g., substance abuse prevention programs), a universe of outcomes (e.g., drug use), a universe of settings (e.g., schools), and a universe of times (e.g., 1980's). Thus, meta-analyses are concerned with generalized causal relationships. This chapter deals with specific threats to the validity of meta-analyses, examining generalized bivariate causal and causal moderator relationships.

As Campbell originally coined the term, "validity threats" refer to situations and issues in research practice that may lead to erroneous conclusions about a causal relationship. However, unlike the validity threats identified by Campbell and Stanley (1963) and Cook and Campbell (1979), this chapter is not concerned with validity threats in

primary studies. Because research synthesis relys on the evidence generated from many different studies, the issue is the total bias across studies rather than bias in a single primary study. Thus, the validity threats discussed in this chapter refer to issues in conducting a research synthesis that may lead to erroneous conclusions about a generalized causal relationship.

Drawing generalized causal inferences in meta-analysis involves three major steps. First, research synthesists need to establish that there is an association between the class of interventions and the class of outcomes. In other words, there has to be evidence that the intervention effect across studies is reliably different from zero. Second, research synthesists have to defend the argument that the relationship examined across studies is causal. Phrased differently, they have to rule out that factors other than the treatments as implemented were responsible for the observed change in the outcomes. Third, given the specific instances of interventions, outcomes, persons, settings, and times included in a review, research synthesists have to clarify the universes of interventions, outcomes, populations, settings, and times about which one can draw inferences. The following paragraphs discuss validity threats that research synthesists may encounter at each of these three steps of generalized causal inference. The research reviews by Bangert-Drowns (1988), Hansen (1992), and Tobler (1986, 1992) are used to provide examples of validity threats and to indicate ways for coping with them.

THREATS TO INFERENCES ABOUT THE EXISTENCE OF A RELATIONSHIP: IS THERE AN ASSOCIATION BETWEEN TREATMENT AND OUTCOME CLASSES?

The first group of validity threats deals with issues that may lead a research synthesist to draw erroneous conclusions about the existence of a relationship between a class of independent variables (i.e., interventions) and a class of dependent variables (i.e., outcomes). In the language of statistical hypothesis testing, these threats may lead to type 1 or type 2 errors because of deficiencies in either the primary studies or the meta-analytic review process. Because research syntheses are concerned with generalized relationships, a single threat in a single study is not likely to jeopardize meta-analytic conclusions in any meaningful way. More critical is whether the same source of bias operates across all or most of the studies being reviewed and whether different sources of bias fail to cancel each other out across studies. This may then lead to a predominant direction of bias, inflating or deflating estimates of a relationship. See

table 1 for a list of threats to valid inferences about the existence of a relationship in a meta-analysis.

Unreliability in Primary Studies

Unreliability in implementing or measuring variables contributes random error to the within-group variability of a primary study, thereby attenuating effect size estimates not only within such a study but also when studies are aggregated meta-analytically.

In the context of drug prevention programs, reliability issues include the measurement of outcome variables such as drug knowledge, attitudes toward drugs, actual drug use, and the fidelity with which prevention programs were implemented. To deal with this issue, correction formulas have been suggested to adjust effect estimates and their standard errors

TABLE 1. *Threats to inferences about the existence of a relationship between treatment and outcome classes.*

(1) Unreliability in primary studies
(2) Restriction of range in primary studies
(3) Missing effect sizes in primary studies
(4) Unreliability of codings in meta-analyses
(5) Capitalizing on chance in meta-analyses
(6) Bias in transforming effect sizes
(7) Lack of statistical independence among effect sizes
(8) Failure to weight study level effect sizes proportional to their precision
(9) Underjustified use of fixed- or random-effects models
(10) Lack of statistical power

(Hunter and Schmidt 1990; Rosenthal 1984). However, Tobler (1986) found that program implementation and the reliability of outcome measures are often poorly documented in primary studies, making comprehensive attempts to correct for attenuation unfeasible. Nevertheless, attenuation corrections are sometimes useful to make the degree of attenuation constant across studies and to better understand the magnitude of effects if interventions were consistently implemented and outcomes measured without error.

Restriction of Range in Primary Studies

When the range of an outcome measure is restricted in a primary study, all correlation coefficients involving this measure are attenuated. Range restrictions may influence other effect size measures differently. For instance, the selection of homogeneous subgroups, blocking, and matching reduce both within-group variability and range. Everything else being equal, this decreases the denominator of the effect size estimated, thereby increasing the magnitude of effect sizes. When such design characteristics operate, Kulik and Kulik (1986) refer to the resulting effect sizes as operative rather than interpretable. Aggregating such operative effect sizes may yield a predominant bias across studies.

In research syntheses of prevention programs, restricted ranges can occur if primary studies involve extreme groups or homogeneous subgroups from a larger population. Effect estimates based on these studies may overestimate program effects in populations with larger variances. Correction formulas can be applied to adjust effect size estimates (Hunter and Schmidt 1990) if valid estimates of population variances are available.

Missing Effect Sizes in Primary Studies

Researchers sometimes provide an incomplete report of findings because of page limitations in journals, the particular emphasis of a research paper, unexpected results, or poor measurement. This reporting practice may bias effect estimates in meta-analyses if researchers in primary studies fail to report, for instance, statistically nonsignificant findings or statistically significant findings in an unexpected direction.

Selective reporting in primary studies is a pervasive issue in many meta-analyses. To prevent possible biases, it is always desirable to code the most complete documents and to contact study authors to obtain information not available in research reports (Premack and Hunter 1988; Shadish 1992). If this strategy is not feasible, there is a need to consider imputation strategies (Little and Rubin 1987; Rubin 1987) and to explore

how missing effect sizes may have influenced effect estimates in a meta-analysis.

Unreliability of Codings in Meta-Analyses

All the data synthesized in a meta-analysis are collected through a coding process susceptible to human error. Thus, meta-analyses contribute sources of unreliability in addition to those in primary studies. Unreliability in the coding process adds error variation to the observations, increasing estimates of standard error and attenuating correlations among effect size estimates and study characteristics. Strategies for controlling and reducing error in codings include comprehensive coder training, pilot testing, and reliability assessments (Cooper 1989).

Capitalizing on Chance in Meta-Analyses

There are three major ways in which meta-analyses may capitalize on chance. First, a publication bias may exist such that studies with statistically significant findings in support of a study's hypotheses are more likely to be submitted for publication. If this is the case, the studies published in the behavioral and social sciences are likely to be a biased sample of all the studies actually carried out (Greenwald 1975; Rosenthal 1979). A second way meta-analysts may capitalize on chance is in extracting effect sizes within studies. Research reports frequently present more than one estimate, especially when there are multiple outcome measures, multiple treatment and control groups, and multiple delayed time points for assessment. Not all of these effect estimates may be relevant for a particular topic, and some relevant estimates may be more important than others. Meta-analysts must then decide which effect estimates should be included in the meta-analysis. Bias may occur when selected effect estimates are just as substantively relevant as those not selected, but differ in average effect size (Matt 1989). A third way that meta-analysts may capitalize on chance is by conducting a large number of statistical tests without adequately controlling for type 1 error.

Bias in Transforming Effect Sizes

Meta-analyses require that findings from primary studies be transformed into a common metric such as a correlation coefficient, a standardized mean difference, or standard normal deviate. Because studies differ in the type of quantitative information they provide about intervention effects, transformation rules were developed to derive common effect size estimates from many different metrics. Bias results if some types of transformation lead to systematically different estimates of average effect

size or standard error when compared to others. For instance, this is likely to be the case when primary studies fail to report exact probability levels and truncated levels (e.g., $p < 0.05$) have to be used to estimate an effect size.

Lack of Statistical Independence Among Effect Sizes

Hedges (1990) states that there are at least four reasons why effect size estimates entering into a meta-analysis may lack statistical independence: (a) Different effect size estimates may be calculated on the same respondents using different measures; (b) effect sizes may be calculated by comparing different interventions to a single control group, or different control groups to a single intervention group; (c) different samples may be used in the same study to calculate an effect estimate for each sample; and (d) a series of studies may be conducted by the same research team, resulting in nonindependent results. A predominant bias may occur if stochastic dependencies among effect sizes influence average effect estimates and their precision (Hedges and Olkin 1985).

The simplest approaches for dealing with dependencies involve analyzing only one of the possible correlated effects or an average effect for each study. However, these approaches fail to take into account information concerning the differences between nonindependent effect sizes, and multivariate analyses or hierarchical linear models may be called for (Bryk and Raudenbush 1992; Raudenbush et al. 1988; Rosenthal and Rubin 1986).

Failure To Weight Study Level Effect Sizes Proportional to Their Precision

Even if one obtains unbiased effect estimates within a study, simply averaging them may yield biased average effect estimates and incorrect sampling errors if the effect sizes from different studies vary in precision (i.e., have different standard errors) (Shadish 1992). Similarly, t tests, analyses of variance (ANOVAs), and regression analyses may provide incorrect results unless weighted estimation procedures are used (e.g., weighted least squares).

Underjustified Use of Fixed- or Random-Effects Models

For the statistical analysis of effect sizes, Hedges and Olkin (1985) distinguish between postulating a model with fixed or random effects. In its simplest form, the fixed-effects model assumes that all studies (e.g., social influence programs) have a common but unknown effect size and that estimates of this population value differ only as a result of sampling variability. In the fixed-effect model, analysts are interested in estimating the unknown population effect size and its standard error. In the random-effects model, each treatment is assumed to have its own unique underlying effect and to be sampled from a universe of related but distinct treatments. Under the random-effects model, the effects of a sample of treatments are best represented as a distribution of true effects rather than as a point estimate.

There is no simple indicator for which model is correct. However, two factors should be considered in the decision whether to assume a fixed- or a random-effects model. The first concerns assumptions about the processes generating an effect. For instance, in the context of drug prevention programs, are all the prevention programs labeled "social influence" identical and are they standardized and administered consistently in all studies? Are the processes by which social influence programs affect drug use the same across all studies? If the answer to these questions is "no" or "probably no," a random-effects model is indicated. The second factor to consider is the heterogeneity of the observed effect sizes. A homogeneity test can be conducted to determine whether the observed variance exceeds what is expected based on sampling error alone. If the homogeneity hypothesis is rejected, the analyst may want to consider the possibility of a random-effects model. Alternatively, if one has reason to insist on a fixed-effects model, the search would begin for the variables responsible for the increased variability.

Lack of Statistical Power

When compared to statistical analyses in primary studies, statistical power will typically be much higher in meta-analyses, particularly when meta-analysts are only interested in estimating the average effect of a broad class of interventions. However, as the meta-analyses on drug prevention programs show (Bangert-Drowns 1988; Tobler 1986, 1992), research synthesists are frequently interested in examining effect sizes for subclasses of treatments and outcomes, different types of settings, and different subpopulations. These subanalyses often rely on a much smaller number of studies than the overall analyses and result in a large number

of statistical tests. The meta-analyst then has to decide which tradeoff to make between type 1 and type 2 error, or, in other words, between the number of statistical tests and the statistical power of these tests.

THREATS TO INFERENCES ABOUT CAUSATION: ARE THERE ANY NONCAUSAL REASONS FOR THE ASSOCIATION?

Whenever a reliable association between independent and dependent variables is presumed to be causal, some additional threats need to be considered. Note again that inferences about the possible causal nature of a treatment-outcome relationship are not necessarily jeopardized by deficiencies in primary studies. A plausible threat arises only if the deficiencies within each study combine across studies to create a predominant direction of bias. In the following, two aspects are considered: bivariate causal relationship and causal moderator relationship. Table 2 gives a brief summary of the threats. See Matt and Cook (1993) for a discussion of threats to causal mediating relationships.

TABLE 2. *Threats to inferences about causation.*

(1) Failure to assign at random
(2) Deficiencies in the implementation of treatment contrasts
(3) Confounding levels of the moderator with substantively irrelevant study characteristics

Failure To Assign at Random

If experimental units (e.g., students, classrooms, schools) are not assigned to treatment conditions at random, a variety of third-variable explanations can jeopardize causal inference in primary studies. The failure to assign at random jeopardizes meta-analytic conclusions if it results in a predominant bias across primary studies.

For research studies of school-based substance abuse prevention programs, Hansen (1992) argues that selection biases are potential threats in quasi-experimental designs comparing groups that inherently differ in expected drug use. In some studies, higher levels of initial risk for substance abuse may be a precondition for entry into a prevention program. Moreover, Hansen's (1992) research suggests that selection biases may be more likely in some program groups (e.g., alternatives)

than in others (affective education). However, despite the potential for selection biases, Tobler's meta-analysis (1986) found little evidence for a predominant bias when comparing randomized trials and quasi-experimental studies.

Deficiencies in the Implementation of Treatment Contrasts

Outside of controlled laboratories, random assignment is often difficult to implement; and even if successfully implemented, it does not ensure that comparability between groups is maintained beyond the initial assignment. Even the most carefully designed randomized experiments and quasi-experiments are not immune to implementation problems such as differential attrition and diffusion of treatments. If the reviewed studies share deficiencies of implementation, a predominant bias may result when studies are combined. However, in trying to examine the implementation of prevention programs more closely, Tobler (1986) found that primary reports often failed to report relevant information.

Hansen (1992) points out another type of implementation issue: studies of school-based prevention programs often involve small numbers of experimental units (i.e., schools), thus jeopardizing the equivalence of control and treatment groups even if experimental units are randomly assigned. While this may threaten the internal validity of a primary study, one would not expect that such nonequivalence necessarily yields a predominant bias when studies are combined in a meta-analysis.

Confounding Levels of a Moderator Variable With Substantively Irrelevant Study Characteristics

Moderator variables condition causal relationships by specifying how an outcome is related to different variants of an intervention, to different classes of outcomes, and to different types of settings and populations. All moderator variables imply a statistical interaction and identify those factors that lead to differently sized cause-effect relationships. Moderators can change the magnitude or the sign of a causal effect, as when Tobler (1986) concluded that peer programs are more effective in reducing drug use than other adolescent drug prevention programs. Threats to valid inference about the causal moderating role of a variable may arise if substantively irrelevant factors are differentially associated with each level or category of the moderator variable under analysis. If the moderator variable (e.g., information/knowledge versus social influence programs) is confounded with characteristics of the design, setting, or population (e.g., urban versus rural schools), differences in the size or direction of a treatment effect brought about by the moderator

cannot be distinguished from differential effects brought about by the potentially confounding variable.

Meta-analysts attempt to deal with confounding issues through statistical modeling (e.g., Tobler 1986, 1992) and through the use of within-study comparisons (e.g., Shapiro and Shapiro 1982). Within-study comparisons are particularly useful because they do not require making assumptions regarding the nature of the confounding. For instance, if the moderating role of prevention programs type A and B is at stake, a meta-analysis could be conducted of all the studies with internal comparisons of prevention programs A and B.

THREATS TO GENERALIZED INFERENCES

Research syntheses promise to generate findings that are more generalizable than those of single studies. Following Cronbach (1982) and Campbell and Stanley (1963), generalizations may involve universes of persons, treatments, outcomes, settings, and times. With respect to research syntheses, Cook (1990) distinguishes three separate though interrelated types of generalized inferences. The first concerns general-ized inferences about classes of persons, treatments, outcomes, settings, and times from which the reviewed studies were sampled. These are the generalizations that meta-analysts like to make; for instance, the effects of goal-setting programs (the treatment class) on drug use (the outcome class) among 8- to 12-year-olds (the target population) in public schools (the target setting class) during the 1980s (the target time).

The second type of generalized inferences concern generalizations across universes. Here, the issue is probing the robustness of a relationship across different populations of persons, different classes of interventions, different categories of settings, different outcome classes, and different time periods. When a relationship is not robust, the analyst seeks to specify the contingencies on which its appearance depends. At issue here are moderator variables, and of particular importance are moderator variables that specify the conditions under which a program has no effect or negative effects.

The third type of generalized inferences concern the generalizability of findings beyond the universes of persons, treatments, outcomes, and settings for which data are available. For example, can the effects of comprehensive prevention programs on the onset of drug use observed in school settings be generalized to church, YMCA, and prison settings? Are the effects of social influence programs observed during the 1970s and

1980s generalizable to programs to be implemented during the 1990s? In each of these examples, the issue is how one can justify inferences to novel universes of persons, treatments, outcomes, settings, and times on the basis of findings in other universes.

Generalizing on the basis of samples is most warranted when formal statistical sampling procedures have been used to draw the particular instances studied. That is, a sampling frame has been designed and instances have been selected with known probability. However, in meta-analyses the instances of person, samples, treatments, outcomes, settings, and times rarely if ever constitute probability samples from whatever universes were specified in the guiding research question. Nevertheless, Cook (1990) argues that generalized inferences about persons, treatments, outcomes, and settings can be tentatively justified even in the absence of random sampling. Cook discusses several principles for justifying generalized inferences in meta-analyses; two of these are further elaborated below. The first requires making a case for the proximal similarity of the sample and population (Campbell 1986). This requires identifying the prototypical, identity-inferring elements (Rosch 1978) of the target classes of persons, settings, causes, and effects and then examining whether they are adequately represented in the sample of studies entering a meta-analysis. In addition to the prototypical elements making a study relevant to a target universe, each individual study's setting, population, measure, and treatments are likely to have unique components that are not part of the target classes. It is crucial that these irrelevancies are made heterogeneous in the sample of studies entering a meta-analysis to avoid confounding prototypical and irrelevant characteristics (Campbell and Fiske 1957).

The second principle for generalizing when random selection cannot be assumed is empirical interpolation and extrapolation. Simply put, the more regularly intervention effects occur across different levels of an independent variable (e.g., length of intervention, type of counselor, type of school), the more tenable is the assumption that a causal effect can be extrapolated to not yet studied but related levels (e.g., shorter or longer interventions, different types of schools and counselors). The more dissimilar the yet unstudied levels are from the levels for which intervention effects have been examined, the more difficult interpolations and extrapolations are to justify. The wider and more diverse the conditions under which the intervention effects follow a predictable pattern, the more justified are generalizations to yet unstudied levels. Table 3 lists threats related to the different types of generalized inference desired in meta-analyses.

TABLE 3. *Threats to generalized inferences.*

(1) Unknown sampling probabilities associated with the set of persons, settings, treatments, outcomes, and times entering a meta-analysis
(2) Underrepresentation of prototypical attributes
(3) Failure to test for heterogeneity in effect sizes
(4) Lack of statistical power for studying disaggregated groups
(5) Restricted heterogeneity of substantively irrelevant aspects
(6) Confounding of subclasses with substantively irrelevant study characteristics
(7) Restricted heterogeneity of classes of populations, treatments, outcomes, settings, and times

Unknown Sampling Probabilities Associated With the Set of Persons, Settings, Treatments, Outcomes, and Times Entering a Meta-Analysis

One can rarely assume that the instances of persons, treatments, outcomes, settings, and times represented in a meta-analysis were randomly selected from the population of persons, settings, treatments, and outcomes to which generalization is desired. Even if there are random samples at the individual study level, it is rare that the studies entering into a meta-analysis constitute a formally representative sample of all such possible study-specific populations. The samples entering primary studies are chosen for proximal similarity and convenience rather than for reasons of formal sampling theory, and the studies containing these samples have an unknown relationship to all the studies that have been completed and that might be done on a particular topic. To tentatively justify generalized inferences in the absence of random sampling, the meta-analyst may follow the principles suggested by Cook (1990).

Underrepresentation of Prototypical Attributes

To demonstrate proximal similarity between a sample and its referent universe requires matching theoretically derived prototypical elements of the universe with the elements of the studies at hand. For substance abuse prevention programs, the question is whether the samples of students, prevention programs, settings, outcomes, and times examined in the reviewed studies represent the core attributes of the populations to which one is interested in generalizing. For instance, Hansen (1992) identified a group of school-based programs and labeled them "social influence programs." Hansen explicates that their "… primary purpose is to teach

students about peer pressure and other social pressures and develop skills to resist these pressures" (p. 415). Thus, a meta-analysis of all the interventions that teach students about peer pressures but fail to include the development of skills to resist peer pressures might not constitute a social influence program. Consequently, such a meta-analysis would not allow generalized inferences to the target population of social influence programs. In a similar vein, program success could be explicated in terms of long-term abstinence from using illegal substances. A meta-analysis in which the majority of studies examine short-term effects, alcohol and tobacco use, the onset of drug use, and attitudes towards drugs would make questionable generalized inferences to the target population of outcomes (i.e., long-term abstinence from illegal substances).

Failure To Test for Heterogeneity in Effect Sizes

A statistical test for homogeneity has been developed (Hedges 1982; Rosenthal and Rubin 1982) that assesses whether the variability in effect estimates exceeds that expected from sampling error alone. Homogeneity tests play an important role in examining the robustness of a relationship and in initiating the search for factors that might moderate the relationship. If the homogeneity hypothesis is rejected, the implication is that subclasses of studies exist that differ in effect size. The failure to test for heterogeneity may result in lumping manifestly different subclasses of persons, treatments, outcomes, settings, or times into one category (i.e., apples-and-oranges problem). The heterogeneity test indicates when studies yield such different results that average effect sizes need to be disaggregated through blocking study characteristics that might explain the mean differences in effect size. Homogeneity tests also protect against searching for moderator variables when effects are robust.

Lack of Statistical Power for Studying Disaggregated Groups

If there is evidence that effect sizes are moderated by substantive variables of interest, then aggregated classes of treatments, outcomes, persons, or settings can be disaggregated to examine the conditions under which an effect changes in sign or magnitude. Such subgroup analyses rely on a smaller number of studies than main effect analyses and may involve additional statistical tests, thus lowering the statistical power for the subanalyses in question. Large samples mitigate against this problem, as do statistical tests adjusted to take into account the number of tests made. Even more useful are analyses based on aggregating within-study estimates of consequences of particular moderator variables.

Restricted Heterogeneity of Substantively Irrelevant Characteristics

Even if prototypical attributes of a universe are represented in the reviewed studies, a threat arises if a meta-analysis cannot demonstrate that the generalized inference holds across substantively irrelevant characteristics. For instance, if the reviewed studies on social influence programs were conducted by just one research team, relied on voluntary participation by students, depended on teachers and principals being highly motivated, or were all conducted in metropolitan areas of California, the threat would then arise that all conclusions about the general effectiveness of homework are confounded with substantively irrelevant aspects of the research context. To give an even more concrete example, if school-based programs were explicated to involve programs administered and implemented in school during grades 4 to 12, it is irrelevant whether the schools are in urban or rural settings, parochial or nonparochial schools, military schools, or elite academic schools. To generalize to school-based programs in the abstract requires being able to show that relationships are not limited to one or a few of these contexts—say, urban or Catholic schools.

The wider the range and the larger the number of substantively irrelevant aspects across which a finding is robust and the better moderating influences are understood, the stronger the belief that the finding will also hold under the influence of not yet examined contextual irrelevancies. Limited heterogeneity in substantively irrelevant variables will also impede the transfer of findings to new universes because it hinders the ability to demonstrate the robustness of a causal relationship across substantive irrelevancies of design, implementation, or measurement method. Tobler (1986) addresses the issue in examining whether program effects are robust regardless of substantively irrelevant characteristics of research design.

Confounding of Subclasses With Substantively Irrelevant Study Characteristics

Even if substantively irrelevant aspects are heterogeneous across studies, the possibility arises that subclasses of treatments, outcomes, settings, persons, or times are confounded with substantively irrelevant characteristics of studies. This situation arose in a meta-analysis of psychotherapy outcomes; differences in treatment effects were observed across different types of psychotherapy, but psychotherapy types were confounded with such substantively irrelevant research design features as the way psychotherapy outcomes were assessed (Wittmann and Matt 1986). This confounding impedes the ability to identify treatment type as a characteristic that moderates intervention effects.

Restricted Heterogeneity in Classes of Populations, Treatments, Outcomes, Settings, and Times

Generalizations across universes and generalizations to novel universes are facilitated if intervention effects can be studied for a large number and a wide range of persons, treatment, outcomes, settings, and times. This is the single most important potential strength of research syntheses over individual studies. For instance, a generalization to a novel universe of time is required if the question is whether school-based drug prevention programs developed and studied during the 1970s and 1980s can be expected to have similar effects in the 1990s. The confidence in such a generalization would be increased if one could demonstrate that the intervention effects were robust throughout the 1970s and 1980s, across different school settings, across different drugs, across different outcome measures, for students from different backgrounds, and so forth. The more robust the findings and the more heterogeneous the populations, settings, treatments, outcomes, and times in which they were observed, the greater the belief that similar findings will be observed beyond the populations studied.

SUMMARY AND CONCLUSIONS

Meta-analyses of drug prevention programs address questions regarding the causal relationship between prevention efforts and substance abuse. Different from primary studies of substance abuse prevention programs, meta-analyses involve generalized causal inferences. At issue are causal effects involving classes or universes of students, prevention programs, outcomes, settings, and times. This chapter presented threats to drawing such generalized inferences regarding bivariate causal and causal moderator relationships. The first group of threats concerns issues that could lead to erroneous conclusions regarding the existence of a relationship between a class of interventions and a class of outcomes. The second group concerns issues that may lead to erroneous conclusions regarding the causal nature of the relationship. Note that in all these instances, deficiencies in primary studies do not necessarily jeopardize the generalized inferences of a meta-analysis; in theory, such deficiencies may cancel each other out. A plausible threat only arises if deficiencies combine across studies to create a predominant bias. The third group of threats concerns issues that may lead to erroneous conclusions about the universes of persons, treatments, settings, outcomes, and times.

All validity threats are empirical products; they are the result of theories of method and the practice of research. Consequently, no list of validity threats is definite. Threats are expected to change as theories of method are improved and more is learned about the practice of research synthesis. All threats are potential; the existence of a threat by itself does not make it a plausible alternative explanation to a causal claim. Research synthesists have to use the empirical evidence, logic, common sense, and any background information available to determine whether a potential threat indeed provides a plausible alternative explanation.

REFERENCES

Bangert-Drowns, R.L. The effects of school-based substance abuse education—a meta-analysis. *J Drug Educ* 18:243-264, 1988.

Bryk, A.S., and Raudenbush, S.W. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage, 1992.

Campbell, D.T. Relabeling internal and external validity for applied social scientists. In: Trochim, W.M.K., ed. *Advances in Quasi-Experimental Design and Analysis*. San Francisco: Jossey-Bass, 1986.

Campbell, D.T., and Fiske, D.W. Covergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 59:81-105, 1957.

Campbell, D.T., and Stanley, J.C. Experimental and quasi-experimental designs for research on teaching. In: Gage, N.L., ed. *Handbook of Research on Teaching*. Chicago: Rand McNally, 1963.

Cook, T.D. The generalization of causal connections: Multiple theories in search of clear practice. In: Sechrest, L.; Perrin, E; and Bunker, J., eds. *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*. DHHS Publication No. (PHS) 90-3454. Washington, DC: U.S. Department of Health and Human Services, 1990.

Cook, T.D., and Campbell, D.T. *Quasi-experimentation. Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company, 1979.

Cooper, H.M. *Integrating Research: A Guide for Literature Reviews*. Newbury Park, CA: Sage, 1989.

Cronbach, L.J. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass, 1982.

Greenwald, A.G. Consequences of prejudice against the null hypothesis. *Psychol Bull* 82:1-20, 1975.

Hansen, W.B. School-based substance abuse prevention: A review of the state of the art in curriculum, 1980-1990. *Health Educ Res Theory Pract* 7:403-430, 1992.

Hedges, L.V. Estimation of effect sizes from a series of independent experiments. *Psychol Bull* 92:490-499, 1982.

Hedges, L.V. Directions for future methodology. In: Wachter, K.W., and Straf, M.L., eds. *The Future of Meta-Analysis.* New York: Russell Sage Foundation, 1990.

Hedges, L.V., and Olkin, I. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press, 1985.

Hunter, J.E., and Schmidt, F.L. *Methods of Meta-Analysis. Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage, 1990.

Kulik, J.A., and Kulik, C.-L.C. "Operative and Interpretable Effect Sizes in Meta-Analysis." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 16-20, 1986.

Little, R.J.A., and Rubin, D.B. *Statistical Analysis with Missing Data*. New York: Wiley, 1987.

Matt, G.E. Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychol Bull* 105:106-115, 1989.

Matt, G.E., and Cook, T.D. Threats to the validity of research syntheses. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis.* New York: Russell Sage Foundation, 1993.

Premack, S.L., and Hunter, J.E. Individual unionization decisions. *Psychol Bull* 103:223-234, 1988.

Raudenbush, S.W.; Becker, B.J.; and Kalaian, H. Modeling multivariate effect sizes. *Psychol Bull* 103:111-120, 1988.

Rosch, E. Principles in categorization. In: Rosch, E., and Lloyd, B.B., eds. *Cognition and Categorization*. Hillsdale, NJ: Erlbaum, 1978.

Rosenthal, R. The 'file drawer problem' and tolerance for null results. *Psychol Bull* 86:638-641, 1979.

Rosenthal, R. *Meta-Analytic Procedures for Social Research*. Beverly Hills, CA: Sage, 1984.

Rosenthal, R., and Rubin, D.B. Comparing effect sizes of independent studies. *Psychol Bull* 22:500-504, 1982.

Rosenthal, R., and Rubin, D.B. Meta-analytic procedures for combining studies with multiple effect sizes. *Psychol Bull* 99:400-406, 1986.

Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

Shadish, W.R., Jr. Do family and marital therapies change what people do? A meta-analysis of behavioral outcomes. In: Cook, T.D.; Cooper, H.M.; Cordray, D.S.; Hartman, H.; Hedges, L.V.; Light, R.J.; Louis, T.A.; and Mosteller, F., eds. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation, 1992.

Shapiro, D.A., and Shapiro, D. Meta-analysis of comparative therapy outcome research: A replication and refinement. *Psychol Bull* 92:581-604, 1982.

Tobler, N. Meta-analysis of 143 adolescent drug prevention programs: Quantitative outcome results of program participants compared to a control or comparison group. *J Drug Issues* 16:537-567, 1986.

Tobler, N.S. Drug prevention programs can work: Research findings. *J Addict Dis* 11:1-27, 1992.

Wittmann, W.W., and Matt, G.E. Meta-Analyse als Integration von Forschungsergebnissen am Beispiel deutschsprachiger Arbeiten zur Effektivität von Psychotherapie [Integration of German-language psychotherapy outcome studies through meta-analysis.]. *Psychologische Rundschau* 37:20-40, 1986.

AUTHOR

Georg E. Matt, Ph.D.
Department of Psychology
San Diego State University
San Diego, CA 92182-4611

Click here to go to page 183