Citation:

Matt, G. E. (2003). Will it work in Münster? Meta-analysis and the empirical generalization of causal relationships. In R. Schulze, H. Holling & D. Böhning (Eds.), *Meta-analysis: New developments and applications in medical and social sciences* (pp. 113-139). Hogrefe & Huber.

8

Will it Work in Münster? Meta-Analysis and the Empirical Generalization of Causal Relationships

Georg E. Matt

Department of Psychology
San Diego State University

Summary

If vigorous physical exercise increases cognitive skills among elderly residents of San Diego, California, will it do the same for the elderly in Münster, Westphalia? This chapter examines the role of meta-analysis in justifying generalized causal inferences. In the *experimental tradition* of the social, behavioral, and natural sciences, such causal generalizations can be justified through a complete understanding of the causal conditions and mechanisms that bring about a phenomenon. Thus, rigorous experimentation and causal modeling of micro-mediating processes should provide the keys to valid causal generalization. In the *observational and correlational tradition* of the social, behavioral, and natural sciences, these generalizations are often justified through the correspondence between samples (or cases, instances, exemplars) and the populations (or universes, constructs, categories, classes) they are meant to represent. This emphasis on correspondence between samples and populations about which inferences are sought suggests that causal generalization may be best accomplished through rigorous random sampling. This chapter argues that, causal explanation and random sampling are of limited use for justifying

generalized causal inferences because the causal moderating and mediating mechanisms are often poorly or incompletely understood and because random sampling – if at all possible – is infrequently practiced. Following a review of different validity models, Cook's (1990, 1993) five principles for strengthening empirical generalizations are presented in detail and illustrated in the context of meta-analysis. Finally, some conditions are outlined that promote generalizable inferences. The chapter concludes that valid empirical generalizations are best achieved through the synthesis of multiple studies, conducted by many research teams, with different populations, in different settings, with multiple operationalizations of interventions and outcomes. One form of research synthesis, meta-analysis, has particularly great promise to facilitate generalized inferences. Even though the best meta-analysis presents no shortcuts or guarantees for valid generalizations, it does provide research design and analytical tools to conduct principled investigations of generalizability claims, thus yielding stronger generalized inferences than are possible based on a single study alone.

8.1 INTRODUCTION

Imagine you just submitted a study for publication of the effects of physical exercise on cognitive skills in the elderly. The journal editor replies, wondering whether your findings apply only to the 49 volunteers from the retirement home across the street, your specific exercise regimen (i.e., ballroom dancing), the poorly-ventilated activity room at the retirement home, the specific operationalization of "cognitive skills" (e.g., standardized test involving verbal and numeric problems), and the foggy and cold November of 1999 when data were collected. Few scientific journals are interested in publishing a study if the research findings only apply to the unique circumstances in which they were conducted. In fact, textbooks on scientific methods (e.g., Babbie, 1995; Kerlinger, 1986) identify the pursuit of *general* truths as a defining feature of science. Do the findings from this study apply to other circumstances, possibly volunteers from other local retirement homes, other forms of exercise, better ventilated rooms, and different seasons? How can we justify such conclusions in the absence of strong sampling designs and strong causal explanatory theories?

This chapter deals with the empirical generalizability of causal relationships, the types of relationship that are at stake when we study the effects of a new drug to delay the onset of Alzheimer's disease, the effects of a "tough love" program for teaching employment skills to the long-term unemployed, or the effects of physical exercise on cognitive skills in the elderly. Following the work of Campbell and Stanley (1966), Cronbach (1982), and Cook (1993; Cook & Campbell, 1979), I will distinguish three types of empirical (i.e., data-driven) generalizations. The first concerns inferences to target populations, classes, or universes (e.g., the population of elderly, the class of retirement homes, the universe of cognitive skills). The second involves generaliza-

tion across populations or sub-populations (e.g., public and private retirement homes, men and women, verbal and numeric skills). The third involves extrapolation and interpolations about novel universes. After presenting some of the traditional approaches to justifying generalized inferences, I will review five principles proposed by Cook (1990, 1993) to strengthen empirical generalizations, illustrate their application in the context of meta-analysis, and discuss the conditions under which valid generalizations are most likely to emerge. The chapter argues that strong generalizations are rarely – if ever – possible based on single studies. Instead, generalizations are best justified by programmatic reviews of findings from many studies, the type of reviews to which carefully conducted meta-analyses can make significant contributions.

8.2 THE DIFFERENT MEANINGS OF GENERALIZATION

The term "generalization" has many meanings and connotations in everyday life and in scientific discourse. Some of these meanings are briefly reviewed in the following.

8.2.1 Crisp and Fuzzy

In everyday discourse, "generalization" refers to a proposition that applies to a large number of instances of a class or group (Webster, 1986). For instances, "It never rains in Southern California." could be interpreted in a crisp manner to mean that there are zero days with precipitation south of Santa Barbara, CA. In addition, "generalization" has the connotation of "vague" or "fuzzy" in that the proposition may not always apply in exactly the same way to each and every member of a group or class (Zadeh & Yager, 1987). That is, that fact that San Diego is in Southern California, receives on average 10 inches of rain per year, and received about $^1/_2$ inch of rain last night does not necessarily disprove the "fuzzy" generalization that one should not expect rain on prototypical days in southern California.

8.2.2 Inductive and Deductive

Generalizations are often the result of inductive inferences, in which general statements are made based on specific observations. Watching a few episodes of Baywatch may lead many TV viewers to the conclusion that, in general, residents of Southern California are very attractive, athletic, and adventurous.

Generalizations may also be the result of deductive inferences in which a general proposition leads to more specific conclusions. Given the general wisdom that Westphalians are stubborn but produce excellent ham, one would expect to find headstrong persons and excellent hams throughout Westphalia's large industrial area around Essen, university towns like Münster, its hamlets along the Dutch border, and its expatriates in Wisconsin.

8.2.3 Logical, Empirical, and Theoretical

In formal logic the validity of an inference depends entirely on its form or structure and not on the subject matter (Groeben & Westmeyer, 1975). A valid inference is one in which a proposition (e.g., a general conclusion) follows with strict necessity from a set of premises (e.g., syllogism). The deduction of the conclusion from the premises must follow the formal rules of logic. For instance, given the premises that "All psychotherapies are effective" and that "Interpersonal therapy is a form of psychotherapy", it follows that "Interpersonal therapy is effective". As soon as the truth of the premises has been established, the validity of the argument is ensured based on the structure of the argument alone. Formal logic is concerned with inference forms rather than with the particular instances.

In contrast, empirical generalizations – the topic of this chapter – are informed primarily by patterns of observations made in particular instances, with inference forms playing a secondary role (Groeben & Westmeyer, 1975). The logic underlying empirical generalizations is closely related to Popper's falsificationist (Popper, 1959, 1972) approach as applied in quasi-experimental designs where plausible alternative hypotheses are identified and ruled out (Cook & Campbell, 1979). The goal of empirical generalizations is not to establish truth but to explore the dependability of generalization claims by subjecting them to falsification tests. That is, empirical generalizations are always tentative and approximate. The more a generalization has been subjected to credible empirical falsification tests, the stronger the belief in its validity.

Empirical generalizations also have to be distinguished from the class of general propositions that make up a theory. Theoretical generalizations are law-like statements that do not directly apply to the empirical world. Instead they rely on theoretical constructs, abstractions and simplifications of complex empirical phenomena, and idealized conditions. Theoretical generalizations present an ideal model of the real world; empirical generalizations present an empirical model of the real world.

8.2.4 Universal and Specific

Universals are empirical generalizations that are of near complete generality (Abelson, 1995) such that they apply to all humans or all humans of a certain type (e.g., retirees taking physical exercise classes). While there are more universals in the natural sciences, there are many near universals in the social and behavioral sciences as well. Examples of nearly universal findings in psychology include: (a) Limitations of short-term memory cause humans to chunk information into groups of no more than about seven items (i.e., Magic Number 7 ± 2); (b) Language acquisition only occurs in humans if they have had exposure to a language community during a critical period in infancy; (c) When deciding among courses of action with equal objective payoffs humans are risk-averse and select the least risky option; and (d) The emotional interpretation of prototypical facial expressions. While exceptions exist to all these

near universals, these exceptions are rare, highlighting that they are "exceptions to the rule", thereby corroborating the existence of the rule.

Generalizations do not have to be universals. Instead, propositions can be phrased at different levels of generality, ranging from near universals to statements identifying specific circumstances under which the proposition holds. The higher the level of generality, the broader the range of circumstances across which the proposition presumably holds.

8.2.5 Transfer, Extrapolation, and Analogs

Transfer refers to a form of generalization where observations made in one condition are extrapolated to another. In learning theory (Estes, 1978; Hommel & Prinz, 1997), transfer is said to have occurred when a subject who learned to respond to a particular stimulus (e.g., a 440 Hz sound) responds as well to similar stimuli beyond the original conditions of training (e.g., 597 Hz, 293 Hz). As differences between two conditions increase, the effects of generalization decrease until there may be no transfer from one situation to another. Alternatively, the more the two situations have in common, the greater is the amount of predictable transfer.

The transfer view of generalizability underlies training approaches in areas where learning on the job can be prohibitively expensive, outright dangerous, or inappropriate for practical and ethical reasons (Cormier & Hagman, 1987). For instance, training pediatric surgeons on critically ill infants to perform a new form of heart catherization, training navy F-14 fighter pilots in actual war situations for combat missions, or training operator personnel of nuclear power plants in actual catastrophic accidents are ethically indefensible. In all of these examples, an important training component takes place on analog counterparts of the actual situations, such as animal models, flight simulators, or analog control rooms. These analogs are designed to maximize the amount of positive transfer (i.e., training that facilitates actual performance), and minimize negative transfer (i.e., training that hinders actual performance).

The transfer view of generalizability is also at the core of *technology transfer* models that aim at facilitating the transition of research findings obtained under laboratory conditions to commercial applications (National Academy of Sciences, 1997). A good example of the implementation of the technology transfer model is the approval process that guides the development of new pharmaceutical products in the U.S.A. (U.S. Food and Drug Administration, 1998). Experimental new drugs are first tested in preclinical studies for safety and efficacy, involving cell cultures, computer models, or animals. If the Food and Drug Administration (FDA) review panels come to the conclusion that findings from the lab are likely to extrapolate to humans, approval is granted for Phase I clinical studies in humans. These are short-term studies on small samples, focusing on safety, and often involve healthy subjects. If these initial studies demonstrate a drug's safety, Phase II studies follow, in which short-term efficacy and drug safety are investigated in larger samples. If Phase II studies continue to demonstrate a drug's safety and efficacy, large-scale Phase

III clinical trials are conducted with a focus on drug dosage, long-term effectiveness, and drug safety. Data collected in the preclinical and clinical trials are again reviewed by FDA expert panels to approve, to request additional studies, or to deny approval of a new drug. Following the approval of a new drug, monitoring systems are put in place to detect adverse reactions and to investigate quality control. The FDA estimates that only 5 of 5,000 compounds entering preclinical testing make it to human testing. Of those, only 1 in 5 are eventually found to be safe and effective and approved for marketing.

8.2.6 Replicability and Robustness

The better research findings replicate across different conditions, the more general the effect is said to be. Robust empirical findings suggest broad main effects of interventions, the type of effects that are particularly useful for policy decisions affecting large and diverse constituencies (Abelson, 1995).

If research findings are replicable within conditions but vary across conditions, interaction effects are present that moderate the direction or magnitude of an effect. Such interactions help identify the boundaries of generalizability. Of particular interest are conditions that reverse the direction of an effect (i.e., qualitative interaction) as is the case when physical exercise lowers blood pressure in some groups but increases blood pressure in others.

8.2.7 Fixed and Random

While robust main effects promise broad generalizability, Abelson (1995) points out that there is a catch. If main effects were investigated based on a limited number of fixed levels (e.g., 7 hours vs. 0 hours of vigorous exercise per week), a disclaimer is necessary stating that the generality is limited to the specific levels represented by the factor. To avoid the limitations of a fixed factor, contexts across which one intends to generalize should be considered as random factors with many different levels from which a sample is being investigated to draw inferences about the whole (e.g., many different durations, types, and intensities of physical exercise).

While a random effects model of contexts will be desirable in many generalizability situations, there are exceptions. Sometimes, researchers deliberately constrain their inferences to particular fixed levels on some factor and random levels on some other factors. For instance, a new psychotherapeutic intervention may rely on a highly standardized treatment manual, administered in controlled inpatient hospital settings for a specific eating disorder (e.g., binge eating). Similarly, the effects of a new drug may be of interest at limited and fixed dosages and for a carefully selected subset of persons suffering from a specific illness.

8.3 A FRAMEWORK FOR EMPIRICAL GENERALIZATIONS

The following framework relies on the work of Brunswik (1947), Campbell and Stanley (1966), Cronbach (1982), and Cook (1990, 1993) on the validity of causal inferences in field settings.

8.3.1 Representative Designs

Brunswik (1947) and Hammond (1948, 1951) were among the first psychologists to raise objections against studying macro-level behaviors with experimental methods under laboratory constraints. Brunswik (1952) argued that when studying behavior at a macro-level "... care must be exercised not to interfere with naturally established mediation patterns." (p. 26). Such an approach calls for research designs that are representative of the natural conditions in which the behavior takes place, that is, designs which Brunswik referred to as having situational representativeness, "naturalness, normalcy, 'closeness to life' " (Brunswik, 1952, p. 29). Clearly, at issue are designs with ecological or situational validity.

For Hammond (1948, 1951) and Brunswik (1952), representative designs lead to generalized statements if a reference class or universe has been specified from which situations are sampled and about which inferences are sought. To achieve a representative design requires not only representative sampling of individuals but also sampling the situational circumstances under which a person functions outside of the research laboratory. This includes stimuli or interventions, responses or outcomes, and settings.

8.3.2 Domains About Which Generalizations May Be Desired

Campbell and Stanley (1966), Cronbach (1982), and Cook and Campbell (1979) have identified five entities about which generalizations may be desired. First are persons or, more generally, the units (U) to which treatments have been assigned. Units may consist of individual humans, animals, or cells as well as larger aggregates such as families, schools, neighborhoods, or states. A second entity is treatments (T), for which a specific operationalization was implemented in a study (i.e., cause constructs). A third entity is outcomes (O) of which specific operationalizations were used to measure effects of interest (i.e., effect constructs). A fourth entity is settings (S), referring to the social and physical environment in which the study takes place. A fifth entity is time (Tm), indicating the historical context in which the study takes place[1].

Each of these domains can involve universes at different levels of generality. For instance, the domain of U may consist of U.S. residents 60 years and older

[1]Note that Cronbach subsumes time in his definition of setting, a distinction with minor influence for the discussion that follows. I will keep with Cook and Campbell's (1979) distinction to better reflect distinct generalizability questions with regard to the social/physical and historical contexts.

or the subset (i.e., sub-U) of residents living in California with annual household income between $20,000 and $30,000 who are registered as independent voters. Similarly, the domain of T may consist of all types of vigorous physical exercises or the subset (i.e., sub-T) of exercises involving stationary bicycles.

A domain may consist of a few fixed levels or an infinite number of random levels. For instance, in some clinical trials of new drugs great efforts are made to control (i.e., fix) as many components of a study as possible. Research protocols are designed and their implementation is carefully monitored in different sites, prescribing in detail the specific characteristics of subjects to be recruited, specific levels of drug dosages to be administered, specific end points to be measured, and specific settings in which treatments are administered and patients are monitored. The goal of these studies is to collect evidence about a specific type of patients, specific dosage levels, specific outcomes, and in closely controlled settings.

In other studies, the aim is to draw inferences about universes consisting of large number of instances. In these situations, the researcher samples a subset of instances to represent the entire domain. For instance, a new algebra curriculum may be tested in public and private schools, with junior and senior instructors, in rural, suburban, and urban areas to draw inferences about the curriculum's effectiveness across a wide variety of school and students.

8.3.3 Generalizability Questions

In their classic work on quasi-experimentation and the validity of causal inferences in field settings, Cook and Campbell (1979) distinguish two major generalizability questions. The first question concerns *generalizations to* well-explicated target domains. This question is invoked when we ask whether a particular sample of retired persons allows valid inferences about the target population consisting of all retired persons. In Cronbach's notation, this question asks whether we can draw inferences from utos to UTOS where u, t, o, s designate the samples of U, T, O, S realized in a particular study. These concepts will be elaborated on later. The first question is most closely associated with inductive probabilistic inferences from samples to populations.

The second question concerns *generalizations across* well-explicated subdomains (i.e., inferences about sub-UTOS). This question is invoked when we ask which different populations or subpopulations (e.g., rural vs. urban vs. suburban; public vs. private; 8th grade vs. 9th grade vs. 10th grade) have been affected by an intervention. The second question is most closely associated with deductive inferences, in which the robustness of a general proposition is investigated across different context conditions. In Cronbach's notation, this question asks whether we can draw inferences from utos to different subsets of UTOS.

Cronbach argues that there is a third generalizability question, which is of particular concern in applied areas of research. This question involves *generalizing from the specific samples and the universes they represent to novel universes* not yet studied. For instances, having found that co-payments for doctor's visits

reduce the number of unnecessary visits in San Diego, CA, will co-payments have the same effect in Münster, Westphalia? In Cronbach's notation, this question concerns inferences from utos to *utos, inferences from the domain of observation to the domain of application. The third generalizability question is closely related to the transfer view of generalizability as it clearly invokes an extrapolation from conditions in which a research finding was investigated to related yet new and distinct conditions.

8.3.4 Justifying Empirical Generalizations

8.3.4.1 Complete Causal Explanation In the experimental tradition of the social and behavioral sciences, generalizations are justified through complete explanation, that is the complete understanding of the causal conditions and mechanisms that bring out a phenomenon. The assumption behind this belief is that when we understand how or why a phenomenon occurs, we can recreate that phenomenon wherever and however its causal ingredients can be brought together (Bhaskar, 1978). This is why causal explanation is often considered the "Holy Grail" of science and the scientific method the path leading to it.

Take for instance the recent approval of thalidomide (alias Contergan) for the treatment of uncontrolled blood vessel growth and severe immuno-modulated diseases. In the 1950s, Chemie Grünenthal, a German pharmaceutical company, developed a sedative called thalidomide so harmless to rodents that an LD50 could not be established (i.e., lethal dose 50 is a measure of acute single exposure toxicity; it indicates the dosage at which 50% of the animals die). The causal mechanisms set in motion by thalidomide were not completely understood, a situation not uncommon even today in many popular drugs (e.g., aspirin, ibuprofen). In the late 1950s and early 1960s, at least 10,000 pregnant women in 46 countries took the sedative in their first trimester, eventually giving birth to infants with missing or stunted limbs. In the early 1960s McBride (1961), Lenz (1962), and Pfeiffer and Kosenow (1962) reported the association between maternal thalidomide usage and limb defects in babies, leading to a world-wide ban. It was not until 30 years later, that an explanation was found for how this potent human teratogen caused missing and stunted limbs.

D'Amato, Loughnan, Flynn, and Folkman (1994) discovered that thalidomide inhibits angiogenesis (i.e., blood vessel growth) in rabbit corneas, changes similar to those found in the deformed limb bud of thalidomide-exposed embryos. It is the ability to inhibit angiogenesis that most likely caused limb defects in babies after maternal thalidomide usage. The causal understanding of how thalidomide works is now being applied to treat conditions characterized by uncontrolled angiogenesis, including diabetic retinopathy and macular degeneration in populations not at risk of becoming pregnant (e.g., Verheul, Panigrahy, Yuan, & D'Amato, 1999).

8.3.4.2 Sampling Theory In the observational and correlational traditions of the social and behavioral sciences, causal generalizations are often justi-

fied through the correspondence between samples (or cases, instances, exemplars, etc.) and the populations (or universes, constructs, categories, classes, etc.) they are meant to represent. The assumption behind this belief is that causal relationships must be easiest to reproduce under the same or similar circumstances they were originally demonstrated. Carefully selecting the specific conditions under which causal effects are demonstrated (e.g., subject characteristics, outcome measures) may allow to approximate important larger classes in which the causal effects hold.

To justify inferences from samples to populations, statistical sampling theory has long played a crucial role in survey research and quality control in industry (Kish, 1965). The crucial element in sampling theory involves selecting units (e.g., persons, hospitals, observers, therapists) with known probability from some clearly designated universe so as to match the sample and population distributions on all (measured and unmeasured) attributes within known limits of sampling error. That is, if one can demonstrate that co-payments for doctor's visits causally reduce unnecessary visits in a random sample of doctors' offices belonging to a particular HMO, the effect in the entire population of HMO subscribers can be estimated.

Note that valid causal explanations are neither a necessary nor a sufficient condition for causal generalizations based on sampling theory. Instead, the causal generalizations based on sampling theory may be a particularly useful tool when complete causal explanations are not available. Similarly, causal explanations may be achieved based on careful experimentation on a few specimens and under highly controlled conditions with little consideration given to sampling theory. For instance, probability sampling did not play a significant role in the original development of thalidomide as a sedative, in the discovery of its devastating side effects, or in the recent approval for new medical indications. However, a case could be made that the thalidomide tragedy could have been reduced had more careful attention been given to sampling principles, including the definition of the universes of persons and outcomes about which inferences are desired and the selection of exemplars from these universes.

8.3.4.3 Campbell's Models for Increasing External Validity Campbell and Stanley (1966) distinguished internal validity from external validity to highlight two distinct inferences about the validity of experiments in field settings. Internal validity refers to the approximate validity with which we infer that the relationship between the manipulated cause and the measured effect is causal. External validity refers to inferences about the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of cause and effect, and across different persons, settings, and times.

Campbell and Stanley (1966) and later Cook and Campbell (1979) acknowledge that internal validity by itself is concerned only with the specific circumstances of how a presumed cause was manipulated and how a presumed effect was measured. Clearly, external validity is necessary before we can expect to replicate a causal relationship in a different sample of persons, with different

manipulations of the presumed cause, different operationalizations of the outcome, or in different settings.

To strengthen external validity in an individual study requires implementing strategies to better represent classes of persons, treatments, outcomes, or settings. If feasible, such strategies include random sampling (e.g., drawing a sample of affectively valenced words from a list of all such words in the English language), impressionistic samples of modal instances (e.g., selecting prototypical public and private schools from rural, urban, and suburban areas), or the deliberate sampling for heterogeneity (e.g., recruit diverse participants with respect to gender, age, income, ethnicity).

While some of these strategies may work in some studies and for some of the entities about which generalizations are desired, they are unlikely to work in most studies and for all generalizations of interest. With few exceptions, individual studies are often constrained by the unique selection of persons, settings, and times, rendering it impossible to draw generalized inferences about larger classes. In many studies, researchers do not have adequate access, budgets, or time to carefully select probability samples from target universes – if meaningful sampling frames for such target populations exist at all. Instead, researchers often select fixed levels from these populations, limiting inferences to such fixed instances. Or, researchers rely on convenience samples, in which cases inferences about a target populations can not be made based on sampling theory. Clearly, external validity is the Achilles' Heal of causal inferences based on an individual study.

8.3.4.4 Cronbach's Model-Based Reasoning for Justifying Internal and External Inferences Cronbach (1982; Cronbach, Nageswari, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) made two significant contributions to our understanding of generalizability. The first concerns the dependability of observations known in the literature on measurement theory as generalizability theory or G-Theory. G-Theory provides a framework for designing and investigating reliable observations by reinterpreting classical reliability theory (Nunnally & Bernstein, 1994) as a theory regarding the adequacy with which one can generalize from a sample of observations to a universe of admissible observations. The universe of admissible observations consists of observations that are interchangeable for the purposes of making a measurement decision. Observations are "dependable" or "generalizable" if they permit accurate inferences about the universe of admissible observations.

Cronbach's second contribution concerns the generalizability of program evaluations. Similar to G-theory, Cronbach defines a domain of admissible operations about which an investigator asks questions and would like to draw inferences. This domain consists of subjects or units (U), interventions or treatments (T), procedures for collecting data on outcomes (O), and the historical and cultural conditions or settings (S). To draw inferences about UTOS, an investigator collects data on instances of the various domains, referred to with lower case letters u, t, and o. Because researchers have little control over the

social and historical context of their research, they can rarely sample instances from S.

According to Cronbach's model, *internal inferences* are involved when making statements about UTOS on the basis of observations on utoS. Questions about the trustworthiness of these inferences are questions about internal validity. Note that Cronbach's internal validity is identical to Campbell's external validity when generalizations to a particular universe are of interest.

In addition to statements about UTOS, investigators and as consumers of research may be interested in domains that are different from the original. Cronbach calls this the domain of application and refers to it as *UTOS. The second generalizability question thus concerns inferences from utos to *UTOS. Statements about *UTOS involve external inferences or extrapolations if we would like to draw inferences about subjects populations, treatments, or outcomes not included in the original study. According to Cronbach's model, questions about the trustworthiness of external inferences are questions about external validity. Note that Cronbach's external validity concerns a generalizability question that Campbell did not consider in his model of external validity.

To justify internal and external generalizations, that is inferences from utos to UTOS or and from utos to *UTOS, Cronbach proposes reasoning by means of models. To justify internal inferences, models are constructed that simulate specific research problem. Models may be descriptive, explanatory, physical, mathematical, or logical, including the blueprints of an architect, the scale model of an engineer, the micromediational model of a microbiologist, for the mathematical model of a survey researcher. Conclusions are drawn in the model and then translated to the real world. Whether conclusions about UTOS are trustworthy depends on the extent to which the model is complete and credible.

Cronbach (1982) describes inferences about *UTOS "as a multi-track, if not trackless process" (Cronbach, 1982, p. 166) because different types of evidence and reasoning have to be combined. Any conclusion about *UTOS must be informed by the differences and similarities between *UTOS and UTOS. Clearly, the more *UTOS and UTOS differ the more has to be filled in to bridge the gap through complementary evidence and credible models to permit trustworthy projections. In general, external inferences about *UTOS are associated with considerably more uncertainty than statements about UTOS.

Cronbach's model-based justifications of generalized inferences include and go beyond the traditional justifications provided by sampling theory or causal explanation. Models may include the complex sets of mathematical equations used by economists to project the effect of increasing oil prices on inflation and unemployment rates. They may also involve informal heuristic models in which many of the premises may not be explicit and in which judgment and formal reasoning have to be combined. Regardless of the model, the credibility of the generalized inference rests on the extent to which the relevant research community accepts the assumptions it is build on. As Cronbach points out, the acceptance of generalized conclusions rests as much on social psychological

processes as it rests on the sheer strength of the empirical evidence with which different parts of a model can be supported.

8.3.4.5 Cook's Five Principles for Strengthening Causal Generalizations

Building on Campbell and Cronbach's work, Cook (1990, 1993) set out to examine how researchers have achieved generalizable causal relationships in the absence of strong causal models and probability sampling. While Cronbach provides a theoretical account of how generalizable claims are substantiated, Cook proposes five principles that researchers use to strengthen claims about the generalizability of causal claims. Cook's work is particularly interesting because it points to strategies and conditions that can be applied in planning individual research studies and be helpful when synthesizing findings from many individual studies.

The Principle of Proximal Similarity. Campbell (1969) introduced the notion of proximal similarity in the context of construct validity. In the context of generalizability, Cook (1990) expands its definition to refer to the correspondence in manifest descriptive attributes between a class of persons, settings, causes, outcomes, and times about which generalizations are sought and the instances based on which empirical evidence about a causal relationship are available. The similarity is proximal because samples and universes match in observable characteristics and not necessarily in any of the more latent explanatory components that link a cause to an effect (Cook, 1990). Proximal similarity is clearly in the spirit of Brunswik's (1956) "situational representativeness".

Demonstrating proximal similarity to the critical standards of the research community is the first necessary condition for generalizing to target universes (i.e., first generalizability question). Proximal similarity is achieved by explicating and then matching the multidimensional content of the classes and instances involved in the generalization. But matching cannot be achieved on all components. Therefore, matching is most importantly achieved with those components that theoretical analysis suggests are central to the construct description.

Cronbach's notion of a domain of "admissible observations" may also be used to argue for proximal similarity. The idea is that a convenience sample of persons, treatments, and so forth may be considered representative if the instances included in the sample and the instances not included in the sample are equally acceptable or exchangeable. Shavelson and Webb (1981) even argue that under these conditions a sample should be considered random. What defines a domain of admissible or exchangeable instances depends on the set characteristics a researcher considers substantive irrelevant (i.e., exchangeable) and the set of characteristics deemed substantively relevant (i.e., prototypical). The latter defines the necessary conditions and the former the unnecessary (i.e., irrelevant) conditions of group membership.

The Principle of Heterogeneous Irrelevancies. A causal relationship will be easier to generalize if it has been replicated in multiple studies especially if these

replications involved different research teams, multiple populations, in multiple settings, with multiple implementations of treatments, and multiple outcome measures. Of interest are replications that are proximally similar with respect to conceptually relevant components but differ in all conceptually irrelevant ones. The principle of heterogeneous irrelevancies can strengthen causal generalization by examining whether the cause-effect relationship under investigation is robust or contingent upon a particular irrelevancy or set of irrelevancies. This is exactly what Cronbach et al. (1972) argue when generalizing from a sample of observations to a universe of admissible observations. In synthesizing findings across irrelevancies, we ask whether the irrelevancies make a difference and whether the causal relationship is obtained *despite* the irrelevancies.

The principle of heterogeneous irrelevancies provides a second necessary condition for generalizing to target universes (i.e., first generalizability question). Findings regarding the benefits of physical exercise in the elderly become trustworthier if they are robust across different type of exercise, different populations of elderly, for different levels of functioning, in different settings. As part of the new drug approval process, the FDA requires preclinical trials to involve at least two animal species to make heterogeneous the presumably irrelevant aspects of the genetic make-up (U.S. Food and Drug Administration, 1998). Clinical trials of new drugs have to be studied in different age and gender groups to determine the robustness (or lack thereof) across these groups. Perhaps the most elaborate application of this principle can be found in meta-analyses of psychotherapeutic interventions, demonstrating the robustness of effects (in causal direction) across a wide variety of different irrelevant characteristics of the researchers, the research design, the intervention, patients, and so forth.

The Principle of Discriminant Validity. The principle of discriminant validity calls for investigations that disentangle the many constituent components of a setting, cause, population, outcome, and time period, to determine the extent to which these components are necessary, sufficient, or irrelevant to the causal relationship under investigation. Through experimental manipulation and observational studies, the goal is to investigate treatment effects in subpopulations, in different settings, with different treatment components, and across different outcome constructs to identify the causal efficacious conditions and discriminate them from related though inefficacious conditions.

This approach does not help, however, if the variations in subpopulations, settings, and so forth are limited such that they all share a common bias. For instance, when all subjects are male or all outcome measures rely on self-report, treatment effects are confounded with gender and assessment method. To apply the principle of discriminant validity, treatment effects have to be studied across populations, treatments, outcomes, and settings with many levels, representing the range across which generalizations are desired.

Moderator effects play an important role in characterizing the boundaries of generalizability and identifying the conditions under which the direction or

strength of a relationship may vary. If this moderator involved a hypothesized substantive irrelevancies, its status now changes from that of a substantive relevancies and attempts should be made to better understand the role of this theory-relevant construct. Investigations of moderator effects are closely associated with the second generalizability question when generalizations across different UTOS to sub-UTOS are of interest.

Discriminant validity is a necessary condition for generalizing across universes. In combination with proximal similarity and heterogeneous irrelevancies, discriminant validity strengthens generalizations by identifying the limits of generalizability and the conditions under which effect changes in sign or magnitude. The principle of discriminant validity is invoked when researchers study dose-response relationships, examine treatment effects across different populations, or distinguish target outcome from side effects.

The Principle of Causal Explanation. While causal relationships are concerned with establishing whether a causal link exists, causal explanations are concerned with identifying *how* or *why* a causal connection occurs. They involve specifying the full set of conditions promoting the cause-effect connection, which often entails identifying the mediational forces set in motion when the treatment varies and without which the effect would not occur.

Causal explanations strengthen empirical generalizations. However, they are not sufficient nor are they necessary conditions for generalizations. Cook (1990, 1993) concludes that the role of causal explanations for justifying generalizations may be overrated. He argues that given the paucity of strong causal explanations and the nature of many problems investigated in the behavioral sciences, it is often unrealistic – if not unethical – to expect and wait for complete causal explanations before attempts are made at causal generalizations. For instance, understanding the micro-mediating processes of a new drug on a molecular level has great significance for making predictions about potential effects in humans. However, it is neither a necessary nor a sufficient condition for the FDA to consider a drug safe and effective. The FDA approval rests primarily on the body of empirical evidence regarding the drug's safety and effectiveness for a particular indication in particular populations and at particular dosage levels. At the same time, a complete causal explanation for the operation of a drug is not sufficient to justify that it is safe and effective for marketing. The recent FDA approval of thalidomide was only given after comprehensive clinical trials despite the fact that the causal mediating mechanisms are quite well understood.

The Principle of Empirical Interpolation and Extrapolation. Populations, treatments, outcomes, settings, and times can vary along many different dimensions. Some of these dimensions are quantitative, in which case special opportunities arise for justifying certain generalizations. Examples for such quantitative dimensions are age, income, weight, family size, treatment dosage or duration, and quantitative outcomes. If we consider these dimensions as fixed factors and collect data at strategically spaced levels (Abelson, 1995), we create the oppor-

tunity to describe the quantitative relation between effect and dose, duration, age, income, and so forth. Assuming valid characterizations of these quantitative relationships, we can derive interpolations and extrapolations about levels of these factors that have not yet been studied. This form of empirical extrapolation and interpolation may strengthen inferences about novel conditions for which no empirical data are available. Thus, empirical interpolations and extrapolations are at the center of the third generalizability question (i.e., generalizations to novel universes).

Interpolation is involved when characteristics can be ordered along a quantitative dimension (e.g., dosage) and when inferences about this characteristic are desired at a level that falls between two known levels. For instance, this is the case when we infer the effects of a drug dosage at 450 mg based on individual studies of the effects at 200, 300, 400, 500, and 600 mg. The narrower the gap and the more data points are available below and above the gap to be interpolated, the more confident are we about our interpolation because the dose-effect relationship is less likely to change abruptly over the interpolated gap.

A similar rationale holds for extrapolations, where we have studied treatment effects at levels 5, 6, 7, 8, 9, 10 and are interested in generalizing to treatment effects at levels 2 and 4, or 12 and 20. Again, the shorter the gap across which we extrapolate and the wider the range of levels across which we studied the relationship, the more confident we are in the extrapolation inferences. Moreover, the wider the range of levels across which we have studied the relationship, the more confident we are that we have identified the proper mathematical model to make the extrapolation (e.g., linear or logarithmic). The shorter the extrapolation leap, the less likely it is that the relationship between level and effect does not change abruptly.

The extrapolation inference will always be weaker than the interpolation inference because we have collected evidence regarding the nature of the relationship from only one direction. From this perspective, interpolations can be sought of as two extrapolations that can be pooled to yield a better estimate.

Interpolations and extrapolations are model-based predictions, whose validity hinges on the assumption that the model holds in between the levels to be interpolated and at the levels to be extrapolated. There are many examples in the natural and behavioral sciences where such assumptions are patently false and relationships change abruptly or take on new qualitative forms. For instance, the physical properties of water change dramatically at specific temperatures. Many pharmaceutical compounds have beneficial effects across a certain range of dosage but may have no effect below and lethal effects above that range. Similarly, in certain problem solving tasks motivation and performance are positively related up to a point at which increases in motivation lead to a decline in performance.

8.4 COOK'S PRINCIPLES APPLIED TO META-ANALYSIS

As a tool for "communal testing of generality" (Abelson, 1995), meta-analysis holds great promise for justifying generalized inferences regarding all three generalizability questions. Matt and Cook (1994) have argued that the generalizability of meta-analytic inferences is particularly justified when

a) the universes about which generalizations are desired are well matched by the instances represented in individual studies (i.e., proximal similarity),
b) individual studies share substantively relevant features but are heterogeneous with respect to irrelevant features (i.e., heterogeneous irrelevancies), and
c) studies can be disaggregated to investigate substantively meaningful subclasses (i.e., discriminant validity).

8.4.1 Meta-Analysis and the Principle of Proximal Similarity

Psychotherapy outcome studies are perhaps the best reviewed body of empirical research using meta-analytic methods (Matt & Navarro, 1997). Following Smith and Glass' (1977) initial meta-analysis of about 500 psychotherapy outcome studies (see also Smith, Glass, & Miller, 1980), more than 50 additional meta-analyses had been conducted by 1992, with many more since then. Glass and colleagues set out to investigate whether psychotherapy in general is beneficial. In the framework presented above, this question implies the desire to generalize to the universe of interventions labeled psychotherapy (T), the universe of persons receiving treatment (U), the universe of settings in which the treatment takes place (S), the universe of outcomes used to assess effects (O), and the historical period during which psychotherapy has been practiced (Tm). Smith and Glass (1977) estimated that psychotherapy treatment group patients did about eight-tenths of a standard deviation better on the outcome variables than did patients in the control groups. Overall, the empirical generalization appears warranted that psychotherapy works.

How can such a general conclusion be justified? The justification begins with investigating the proximal similarity between target universes and instances included in the meta-analysis. Evidence has to be generated that the broad universes of UTOSTm were well represented by samples included in the meta-analysis. The goal is not an exact match or probabilistic representation as sampling theory suggests. Instead, the goal is to achieve an approximate match on prototypical components with multiple operational representations. Or, in Cronbach's (1982) terms, one has to argue that the instances of UTOSTm included in the meta-analyses are exchangeable with the instances not included. In Smith and Glass' meta-analysis (1977), "psychotherapy" included a variety of orientations and techniques such as psychodynamic, behavioral, cognitive, interpersonal, hypnosis, bibliotherapy, eclectic, and others. Similarly, outcomes included a multitude of measures, ranging from global in-

dices of adjustment to frequency counts of a specific symptom and from standard trait inventories to ad hoc therapist ratings. It appeared that key prototypical characteristics of "psychotherapy" were represented in the sample of studies included in Smith and Glass' meta-analysis.

8.4.2 Meta-Analysis and the Principle of Heterogeneous Irrelevancies

Once a case has been made for proximal similarity, meta-analysts have to generate evidence that a causal connection is not completely confounded with any specific characteristic of u, t, o, s, tm. This calls for the application of the principle of heterogeneous irrelevancies. The greater the number of irrelevancies across which primary studies differ, the greater the chance that a causal connection is not completely confounded. The assumption that substantive irrelevancies are heterogeneous should not be made lightly. It has to be based on evidence that mono-operation biases are not present across the domains of which generalizations are desired.

Lack of heterogeneity was not a problem in the Smith and Glass meta-analyses. Smith and Glass coded primary studies to collect data on many substantively relevant and irrelevant study characteristics, including year and source of publication, professional affiliation of authors, age, gender, socioeconomic status of participants, type and reactivity of outcome measures, type of control conditions, sample size, duration of therapy, experience of therapist, recruitment of subjects, and setting in which therapy took place.

Overall, the heterogeneity in the sample of studies appeared to match the heterogeneity of the domain about which inferences are desired. That is, psychotherapy outcome studies are conducted by many different research teams, researchers with training in different disciplines and with different professional affiliations. Researchers use different approaches for recruiting subjects, implementing treatments, and measuring outcomes. Similarly, psychotherapy is conducted across a wide range of settings, including private practices, schools, community mental health centers, and university-affiliated hospitals.

Within specific subclasses of treatments, heterogeneity was reduced, giving rise to potential mono-operation biases. For instance, some of the subclasses of interventions relied more heavily on small sample sizes, student volunteers, school settings, short therapies, and certain types of outcome measures. Heterogeneity in studies is welcomed if it matches the heterogeneity of the domain about which inferences are desired. However, any restriction would limit conclusions regarding the generalizability, as was the case with certain subclasses of psychotherapeutic interventions (Matt & Navarro, 1997).

8.4.3 Meta-Analysis and the Principle of Discriminant Validity

While proximal similarity and heterogeneous irrelevancies are at the center of generalizations to target domains (i.e., first generalizability question), the principle of discriminant validity plays a key role when generalizing across subdomains. Rather than lumping heterogeneous studies together, meta-analysts

can test whether the heterogeneity in treatment effects is larger than expected due to chance alone (Hedges & Olkin, 1985). Equally important is the decision whether to rely on a fixed, random, or conditional random effects model (Hedges, 1994; Hedges & Vevea, 1998). This decision is influenced by whether a researcher is interested in drawing inferences about a few clearly defined (i.e., fixed) subclasses of a domain or to the entire domain consisting of a large number of admissible parts.

Depending on the statistical model chosen and sample size permitting, heterogeneous domains of U, T, O, S, Tm may be disaggregated to identify more homogeneous subdomains. This starts the exploration of potential interaction effects, that is, substantively relevant and substantively irrelevant characteristics that moderate the size or direction of treatment effects. For instance, such analyses may indicate that different types of interventions or different outcome constructs (e.g., a substantively relevant heterogeneity) are associated with different effect sizes but that treatment effects are robust with respect to the type of setting or subject recruitment.

The principle of discriminant validity is applied in meta-analyses when studies are stratified to investigate moderator conditions (e.g., gender, treatment types) or to distinguish important cognate constructs from each other (e.g., functional disability, life satisfaction, self-esteem, symptoms, adjustment). As mentioned above, Smith and Glass' meta-analysis spawned a large number of additional meta-analyses on psychotherapy effects (Matt & Navarro, 1997). The purpose of these additional meta-analyses was to explore whether psychotherapy effects generalize across different types of interventions, outcomes, settings, and populations. While there is evidence that certain conditions moderate the magnitude of psychotherapy effects, none of the meta-analyses identified conditions associated with harmful effects (Matt & Navarro, 1997). That is, the beneficial effects of psychotherapy generalize across wide range disorders, types of interventions, outcomes, settings, and time periods.

The principle of discriminant validity was also applied in meta-analytic investigations of the placebo effect in psychotherapy (Matt & Navarro, 1997). At issue is the question whether a treatment group, which is presumed to be receiving psychotherapy, is actually receiving both psychotherapy and a host of nonspecific placebo interventions. The latter include, for example, the mere attention given by a caring person, the expectation of improvement brought about simply by being seen by a mental health professional with credentials for dealing with psychological problems, or by the simple fact of talking with another human being about the problem. Such placebo effects have proven to be so powerful in medicine that they need to be controlled by using double-blind designs and introducing placebo control groups.

To examine the role of placebo effects in psychotherapy, at least eight meta-analyses compared experimental studies in which patients in psychotherapy were compared against patients who received placebo treatment that did not include the presumed active therapy ingredient (Bowers & Clum, 1988; Casey & Berman, 1985; Clum, Clum, & Surls, 1993; Kazdin, Bass, Ayers, & Rodgers, 1990; Landman & Dawes, 1982; Lyons & Woods, 1991; Matheny, Aycock, Pugh,

Curlette, & Silva, 1986; Miller & Berman, 1983). Pooling estimates across these meta-analyses suggest that about 20% of the total psychotherapy effect could indeed be attributed to nonspecific treatment components (i.e., placebo effect; $d = .18$). However, the remaining 80% of the total effect can be attributed to the unique treatment components of psychotherapy ($d = .68$). Thus, the total psychotherapy effect appears to be a combination of specific and nonspecific treatment effects ($d_{\text{Total}} = .68 + .14 = .82$).

8.4.4 Meta-Analysis and the Principle of Empirical Interpolation and Extrapolation

Empirical interpolation and extrapolation are most closely linked to the third generalizability question, in which inferences about novel universes are desired. Because individual studies are often limited in the range of u, t, o, s that are included, combining different studies may be of great benefit if each study investigates a different level of a quantitative dimension.

It is common in meta-analyses to model dose-response relationships where different studies contribute effect estimates at different dosages. Similarly, meta-analyses frequently examine the stability of treatment effects over time by combining data from studies in which effects were assessment at different time intervals after treatment ended. Such models may then be used to interpolate or extrapolate effects at levels not studied.

Recently, Shadish, Matt, Navarro, and Phillips (2000) applied an extrapolation strategy in a meta-analysis of psychotherapy outcomes in "the lab" (i.e., efficacy conditions) versus "the clinic" (i.e., effectiveness conditions). Briefly put, the "lab vs. clinic" debate arose because the vast majority of psychotherapy outcome studies are conducted in ways that are not very representative of the conditions under which therapy is actually conducted by practicing therapists. For example, lab studies often are conducted at universities with clinically inexperienced graduate student therapists who are trained intensively and specifically in a single treatment that is then applied uniformly to a highly selected patient population with a narrow range of problems that the treatment is deliberately designed to help. Such therapy is quite different from the real world of clinic therapy in which experienced therapists work in busy clinics giving an eclectic array of therapy to patients with diverse problems. If this criticism is true, then there might be serious doubts about whether the results of psychotherapy meta-analyses generalize to clinically representative conditions. Indeed, in a preliminary examination of this issue limited to child psychotherapy, Weisz, Weiss, and Donenberg (1992) concluded that the very few studies of clinic therapy that they could locate showed little or no effects compared to no treatment control. Shadish et al. (1997) revisited the same issue, asking the authors of published meta-analyses to identify studies in their data bases that were conducted outside of research labs. They found that the effects of studies approximating clinical practice were about the same as those conducted in research labs. However, like Weisz et al. (1992), Shadish et al. (1997) found very few studies of clinic therapy, and concluded that the gener-

alizability of psychotherapy meta-analysis results to clinically representative settings was a topic that still needed far more study.

Recently, Shadish et al. (2000) reinvestigated this issue and conducted a new meta-analysis of 90 psychotherapy outcome studies, differing in the degree to which they approximate prototypical conditions of clinical practice. They concluded that therapy effects do not deteriorate over the range of clinical representativeness that was present in the 90 outcomes studies. Shadish et al. (2000) also found that effects increase with larger dose, and when outcome measures are specific to treatment. Thus, clinic therapies may be able to produce larger effects by providing longer and more intensive treatments. Moreover, some clinically representative studies used self-selected treatment clients who were more distressed than available controls, and these quasi-experiments underestimated therapy effects. Given the range of clinical representativeness in existing outcome studies, Shadish et al. (2000) extrapolated effects of an ideal study of clinically representative therapy. This projection suggests that effects are similar to those reported in past meta-analyses of studies conducted in research settings.

8.4.5 Meta-Analysis and the Principle of Causal Explanation

There are two major strategies with which causal explanation may strengthen generalized inferences based on meta-analysis. The first strategy involves decomposing domains to isolate those components that are involved in the generation and moderation of a treatment effect. Meta-analyses contribute to causal explanation to the extent that the studies allow meaningful decomposition of treatments, outcomes, persons, or settings. For instance, decomposing treatment effects in specific and nonspecific components contributes to the causal explanation of how psychotherapy effects come about. Similarly, differentiating between different settings in which services are delivered (i.e. clinic vs. lab), between levels of therapist experience and training, between types and modalities of psychotherapy, or between effects on target behaviors and peripheral symptoms contribute to a better understanding of the causal effects of the intervention.

A second strategy involves identifying the causal mediating processes that are set in motion by a treatment. Although there is nothing in theory that would prevent meta-analyses to strengthen inferences about causal mediating processes, such contributions are unlikely to be made routinely. The reason for this is that our best explanatory models are typically phrased at levels far below that of a meta-analysis aggregating across studies. For psychotherapy this may involve theories of behavior change at the level of an individual person or family. For meta-analyses to synthesize studies of micro-mediating processes at that level, theories must have reached a level of maturity and consensus such that multiple studies of the same theory and of the same micro-mediating processes are available. In many disciplines within the social and behavioral sciences, researchers tend to pursue different causal explanations

and measure different explanatory processes, providing little opportunity to accumulate multiple studies from different research groups.

There are some more mundane constraints for meta-analyses of micro-mediating processes. There is the paucity of detail about treatment components in many publications. Unless additional detail can be obtained by contacting the original researchers, a meta-analysis can often contribute very little to further causal explanation at the level of micro-mediating processes. There is also a considerable amount of selective reporting that takes place when publishing study findings. Thus, a study may highlight the significant contribution of some micro-mediating processes but fail to report the nonsignificant contribution of others. Such a reporting bias (Matt & Cook, 1994) favors Type I errors and limits generalizability of meta-analytic conclusions. Finally, reported methods often reflect what was planned rather than what was achieved, in which case meta-analysts are in a poor position to accurately describe the conditions under which a particular process was observed. There are a few notable exceptions in behavioral science meta-analyses (e.g., Harris & Rosenthal, 1984; Becker, 1992; Devine, 1992; Shadish, 1992) and there may be more in the natural sciences where theories have reached more mature levels.

8.5 CONDITIONS THAT FACILITATE GENERALIZED CAUSAL INFERENCES

Generalizations based on single studies are usually weak. This is the case because an individual study can only do so much to make heterogeneous the substantive irrelevancies and explore potentially interacting moderator conditions. At the same time, the fact that many studies have been conducted on a particular topic does not guarantee valid empirical generalizations. What follows is an outline of some of the conditions that promote generalizable causal inferences.

8.5.1 Individual Programs of Research

Programs of research are collections of multiple connected studies conducted by one or more researchers or research groups on a particular research question or family of questions. Studies conducted as part of a research program play a crucial role in generating generalizable knowledge, particularly if they involve different research designs, different populations, interventions, measures, and settings. Multiple studies generated by such research programs facilitate the generalization of findings by probing the robustness of causal relationships in the face of substantively irrelevant aspects, by identifying the substantive factors that moderate an effect, and by elaborating explanatory processes. They provide evidence based on which interpolations and extrapolations can be based and the proximal similarity between sample and universes can be justified.

8.5.2 Integrative Reviews

Different from individual programs of research, integrative reviews involve collections of primary studies that are not necessarily well connected or orchestrated by a network of researchers. Instead, such studies often cover many different research programs, published over many decades, by researchers from different countries on different continents, subscribing to different research paradigms. While integrative reviews are less likely to advance causal explanations – a particular strength of individual research programs – their major contribution is likely to come from exploring heterogeneous irrelevancies across the large diversity of populations, outcomes, interventions, historical and cultural contexts, and so forth. Because integrative reviews are likely to involve large numbers and diverse characteristics, they provide a particularly good opportunity to explore the robustness of a causal relationship, to investigate factors that may moderate its direction or size, and to interpolate and extrapolate to new domains.

While meta-analyses have many potential benefits for facilitating generalized inferences, they cannot provide a panacea for generalization questions. As is the case with quasi-experimental designs of primary studies, the non-experimental nature of meta-analyses makes it necessary to carefully examine and rule out plausible alternative explanations to claims of causal generalization (Matt & Cook, 1994).

8.5.3 Critical Multiplism

Focused research programs and integrative reviews will provide little evidence to support causal generalizations if individual studies rely on the same subject populations, recruitment strategies, interventions, measures, and settings. Such shared research design and interventions characteristics are likely to produce the same method biases masquerading and confounding underlying causal effects. Synthesizing individual studies that share the same biases yields biased meta-analytic findings, leaving the meta-analytic evidence of little use to support any of the principles discussed above.

To provide a rich foundation for causal generalization, critical multiplist methods should be applied whenever possible at the level of the individual study, the focused research program, or the integrative review (Cook, 1982; Shadish, 1993). To mention a few, such methods call for the investigation of multiple methods to assess outcome, multiple settings to investigate interventions, several smaller studies rather than a single large study, multiple subject population, multiple research groups, and so forth.

8.5.4 Public Debates

By their nature, causal generalizations are at the center of many public policy debates. Such debates play a crucial role in forcing out hidden assumptions and assuring that important stakeholders have been taken into account.

Because of public debates, new sources of bias or important new contextual conditions may be discovered. Public debate will draw attention to the personal and public costs and benefits of more stringent or liberal policies. In addition, public debates help establish the extent to which more liberal or more conservative generalizations may be implemented. In the case of California's public policy regarding second-hand smoke exposure, public debates (including a public referendum) led to the adoption of a more liberal generalization regarding the health effects of low-level second-hand smoke exposure. In this instance, the public health benefits of "overgeneralizing" (i.e., second-hand smoke exposure is toxic at any level) were found to outweigh the risk of policies mandating ventilation systems to reduce second-hand smoke exposure levels even though such harsher policies interfere with the civil rights of smokers.

8.6 CONCLUSIONS

Rarely – if ever – do researchers and consumers of research believe that findings apply only to the specific circumstances of a specific study. Instead, we believe – or behave as if – findings apply to larger domains of persons, interventions, outcomes, settings, and times than were included in a study. At issue are three types of empirical generalizations with respect to persons, treatments, outcome, settings, and times. The first type of generalization involves inferences to target universe based on specific samples, a form of inductive empirical generalizations. The second involves inferences across target universes or across sub-universes, a form of deductive empirical generalizations. The third involves generalized inferences to novel universes, closely related to empirical interpolation and extrapolation.

Empirical generalizations are best achieved through the synthesis of multiple studies, conducted by many research teams, with different populations, in different settings, with multiple operationalizations of interventions and outcomes. One form of research synthesis, meta-analysis, has particularly great promise to facilitate generalized inferences. Meta-analysis provides no shortcuts or guarantees for generalizations. However, it does provide research design and analytical tools to conduct principled investigations of generalizability claims and increases the likelihood of better inferences compared to individual studies.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum.

Babbie, E. R. (1995). *The practice of social research* (7th ed.). Belmont: Wadsworth.

Becker, B. (1992). Models of science achievement: Forces affecting male and female performance in school science. In T. D. Cook, H. Cooper, D. S. Corday, H. Hart-

mann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A case-book* (pp. 209–281). New York: Russell Sage Foundation.

Bhaskar, R. (1978). *A realist theory of science* (2nd ed.). Atlantic Highlands, NJ: Humanities Press.

Bowers, T. G., & Clum, G. A. (1988). Relative contribution of specific and nonspecific treatment effects: Meta-analysis of placebo-controlled behavior therapy research. *Psychological Bulletin*, *103*, 315–323.

Brunswik, E. (1947). *Systematic and representative design of psychological experiments.* Berkeley, CA: University of California Press.

Brunswik, E. (1952). *The conceptual framework of psychology.* (Int. Encycl. unified Science, v. 1, no. 10.). Chicago, IL: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Campbell, D. T. (1969). Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Casey, R. J., & Berman, J. S. (1985). The outcome of psychotherapy with children. *Psychological Bulletin*, *98*, 388–400.

Clum, G. A., Clum, G. A., & Surls, R. (1993). A meta-analysis of treatments for panic disorder. *Journal of Consulting and Clinical Psychology*, *61*, 317–326.

Cook, T. D. (1982). Postpositivist critical multiplism. In L. Shotland & M. M. Marks (Eds.), *Social science and social policy.* Newbury Park, CA: Sage.

Cook, T. D. (1990). The generalization of causal connections. In L. Sechrest, E. Perrin, & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data.* (DHHS Pub. No. 90–3454). Washington, DC: U.S. Department of Health and Human Services.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them* (pp. 39–82). (New Directions in Program Evaluation, No. 57). San Francisco: Jossey Bass.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design & analysis issues for field settings.* Boston, MA: Houghton Mifflin.

Cormier, S. M., & Hagman, J. D. (1987). *Transfer of learning: Contemporary research and applifications.* San Diego: Academic Press.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Fransisco: Jossey Bass.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *British Journal of Statistical Psychology*, *16*, 137–163.

D'Amato, R. J., Loughnan, M. S., Flynn, E., & Folkman, J. (1994). Thalidomide is an inhibitor of angiogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *91*, 4082–4085.

Devine, E. C. (1992). Effects of psychoeducational care with adult surgical patients: A theory-probing meta-analysis of invention studies. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A case-book* (pp. 35–82). New York: Russell Sage Foundation.

Estes, W. K. (1978). *Handbook of learning and cognitive processes.* Hillsdale, NJ: Lawrence Erlbaum.

Groeben, N., & Westmeyer, H. (1975). *Kriterien psychologischer Forschung* [Criteria for psychological research]. München: Juventa.

Hammond, K. R. (1948). Subject and object samling – A note. *Psychological Bulletin*, *45*, 530–533.

Hammond, K. R. (1951). Relativity and representativeness. *Philosophy of Science*, *18*, 208–211.

Harris, M. J., & Rosenthal, R. (1984). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, *97*, 363–386.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Hommel, B., & Prinz, W. (1997). *Theoretical issues in stimulus–response compatibility.* New York: Elsevier.

Kazdin, A. E., Bass, D., Ayers, W. A., & Rodgers, A. (1990). Empirical and clinical focus of child and adolescent psychotherapy research. *Journal of Consulting and Clinical Psychology*, *58*, 729–740.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart and Winston.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Landman, J. T., & Dawes, R. M. (1982). Psychotherapy outcome. Smith and Glass' conclusions stand up under scrutiny. *American Psychologist*, *37*, 504–516.

Lenz, W. (1962). Thalidomide and congenital abnormalities. *The Lancet (Volume 1)*, 45.

Lyons, L. C., & Woods, P. J. (1991). The efficacy of rational emotive therapy: A quantitative review of the outcome research. *Clinical Psychology Review*, *11*, 357–369.

Matheny, K. B., Aycock, D. W., Pugh, J. L., Curlette, W. L., & Silva, K. A. (1986). Stress coping: A qualitative and quantitative synthesis with implications for treatment. *Counseling Psychologist*, *14*, 499–549.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (p. 503-520). New York: Russell Sage Foundation.

Matt, G. E., & Navarro, A. M. (1997). What meta-analyses have and have not taught us about psychotherapy effects: A review and future directions. *Clinical Psychology Review*, *17*, 1-32.

McBride, W. G. (1961). Thalidomide and congenital abnormalities. *The Lancet (Volume 2)*, 1358.

Miller, R. C., & Berman, J. S. (1983). The efficacy of cognitive behavior therapies: A quantitative review of the research evidence. *Psychological Bulletin*, *94*, 39–53.

National Academy of Sciences. (1997). *Technology transfer systems in the United States and Germany: Lessons and perspectives.* Washington, DC: National Academic Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

Pfeiffer, R. A., & Kosenow, W. (1962). Thalidomide and congenital abnormalities. *The Lancet (Volume 1)*, 45–46.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach.* Oxford: Clarendon Press.

Shadish, W. R. (1992). Do family and marital therapy change what people do? A meta-analysis of behavioral outcomes. In T. D. Cook, H. Cooper, D. S. Corday, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A case-book* (pp. 129–208). New York: Russell Sage Foundation.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. In L. Sechrest (Ed.), *Program evaluation: A pluralistic enterprise* (pp. 13–57). (New Directions in Program Evaluation, No. 60). San Francisco: Jossey-Bass.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*, 512–529.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, S., Crits-Christoph, P., Hazelrigg, M., Jorm, A., Lyons, L. S., Nietzel, M. T., Thompson Prout, H., Robinson, L., Smith, M. L., Svartberg, M., & Weiss, B. (1997). The effects of psychotherapy conducted under clinically representative conditions. *Journal of Consulting and Clinical Psychology*, *65*, 355–365.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133–166.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: The John Hopkins University Press.

U.S. Food and Drug Administration. (1998). *The Center for Drug Evaluation and Research (CDER) Handbook.* Washington, DC: Food and Drug Administration.

Verheul, H. M., Panigrahy, D., Yuan, J., & D'Amato, R. J. (1999). Combination oral antiangiogenic therapy with thalidomide and sulindac inhibits tumour growth in rabbits. *British Journal of Cancer*, *79*, 114–118.

Webster. (1986). *Webster's third new international dictionary of the English language unabridged.* Springfield, MA: Merriam-Webster.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic. Effects of child and adolescent psychotherapy. *American Psychologist*, *47*, 1578–1585.

Zadeh, L. A., & Yager, R. R. (1987). *Fuzzy sets and applications: Selected papers.* New York: Wiley.