Review of General Psychology
2008, Vol. 12, No. 3, 236–252

1089-2680/08/$12.00 DOI: 10.1037/1089-2680.12.3.236

The Causal Generalization Paradox: The Case of Treatment Outcome Research

Graham L. Staines
Highland Park, New Jersey

Data from psychological experiments pose a causal generalization paradox. Unless the experimental results have some generality, they contribute little to scientific knowledge. Yet, because most experiments use convenience samples rather than probability-based samples, there is almost never a formal justification, or set of rigorous guidelines, for generalizing the study's findings to other populations. This article discusses the causal generalization paradox in the context of outcome findings from experimental evaluations of psychological treatment programs and services. In grappling with the generalization paradox, researchers often make misleading (or at least oversimplified) assumptions. The article analyzes 10 such assumptions, including the belief that a significant experimental treatment effect is likely to be causally generalizable and the belief that the magnitude of a significant experimental effect provides a sound effect size estimate for causal generalization. The article then outlines 10 constructive strategies for assessing and enhancing causal generality. They include strategies involving the scaling level of outcome measures, variable treatment dosages, effectiveness designs, multiple measures, corroboration from observational designs, and the synthesis of multiple studies. Finally, the article's discussion section reviews the conditions under which causal generalizations are justified.

Keywords: causal generalization, replication, purposive sampling, convenience samples, treatment outcome research

Large-scale nonexperimental studies such as cross-sectional surveys often use probability sampling, whose key components are a well-defined population, sampling units of this population with known probability (e.g., random sampling), and use of sample statistics (e.g., the mean) to estimate corresponding population parameters with known degrees of precision (expressed in terms of confidence intervals). In experimental research, the focus shifts from random (more properly, probability) sampling (which emphasizes the similarity between sample and population) to random assignment (which emphasizes the similarity among samples in different design conditions), where both selection procedures are based on known probabilities. In the two simplest situations, all cases in the population have an equal probability of being selected for the sample ([simple] random sampling), and all cases in the study have an equal chance of being assigned to any condition (random assignment). With sample equivalence (random assignment) taking priority over sample representativeness (random sampling), experimental research typically selects units, in part or in whole, at the convenience of the researcher. The researcher makes no attempt, or only a limited attempt, to ensure that this sample is an accurate representation of some larger group or population. Examples of convenience samples include patients in a clinic (clinical research), students in a classroom (educational research), and shoppers at a mall (market research).

Causal generalization most commonly refers to whether a study's findings can be extrapolated to a different set of circumstances (e.g., a different setting or different participants). In an experimental context, the question raised by causal generalization is whether a causal relationship demonstrated in an initial experiment holds over variations in persons, settings, treatments,

I gratefully acknowledge the assistance of Georg Matt in the preparation of this article.

Correspondence concerning this article should be addressed to Graham L. Staines, 321 North Fourth Avenue, Highland Park, NJ 08904. E-mail: glstaines@aol.com

and outcomes (for reviews, see Cook, 1990, 1993; Matt, 2003; Matt & Navarro, 1997; Shadish, Cook, & Campbell, 2002; West, Biesanz, & Pitts, 2000). To take an educational example, will the findings on the effects of a kindergarten Head Start program in one ethnic group in one city hold for the same program when administered to another ethnic group in a different city?

Data from such psychological experiments pose a causal generalization paradox. Unless the experimental results have some generality (or robustness), they contribute little to scientific knowledge. Yet, because most experiments use convenience samples rather than probability-based samples, there is almost never a formal justification, or set of rigorous guidelines, for generalizing the study's findings to other participant populations (Staines, 2007). Thus, although there is rarely any lasting interest in what happened to the particular participants in an experiment, information about these participants is all that an experiment provides.

As Shadish et al. (2002, p. 18) explained, causal generality, although a serious problem in all research, is the Achilles heel of experimentation: "The strength of experimentation is its ability to illuminate causal inference. The weakness of experimentation is doubt about the extent to which that causal relationship generalizes." They added that for research limited to single studies, "A conflict seems to exist between the localized nature of the causal knowledge that individual experiments provide and the more generalized causal goals that research aspires to attain" (p. 19). Furthermore, if scientists have designed their experiments primarily to maximize internal validity, they may have neglected the goal of causal generality. Absent a rigorous scientific argument, causal generalizations based on experimental studies rely substantially on scientists' educated guesses guided by assumptions and relevant evidence. The fallibility of researchers' predictions is thus an inherent component of causal generalizations (Matt, 2003).

For experimenters engaged in causal research to evaluate the success of professional interventions, the causal generality of their findings has high priority. Shadish et al. (2002, p. 19) observed that "policymakers may be interested in whether a causal relationship would hold (probabilistically) across the many sites at which it would be implemented as a policy, an inference that requires generalization beyond the original experimental study context." Experimenters are thus well aware that a scientific theory's value depends on its breadth of coverage of phenomena, and they assign considerable importance to generalizing causal inferences. After conducting an experiment that has high internal validity, therefore, investigators favor causal generalization of their significant findings (Essock, 2006).

Because the paradox of causal generality poses a severe challenge to researchers, many discussions have relied, whether explicitly or implicitly, on misleading or oversimplified assumptions. In this article, I first consider 10 such common but questionable assumptions about causal generality, then offer 10 constructive and defensible guidelines that assist investigators in assessing and enhancing the generality of their research findings. I focus on the causal generality of outcome findings from experimental evaluations of psychological treatment programs and services (Matt & Navarro, 1997). Although the emphasis is on treatment outcome research, many of the principles formulated apply, in increasing levels of generality, to all program evaluation, all field experiments, all psychological experiments, and all psychological research.

Flawed Assumptions

Assumption 1: Causal Generalizations Are the Product of Inductive Inferences

Causal generalizations are typically the product of inductive inferences in which general statements are made on the basis of specific observations (e.g., experimental data). To return to the earlier educational example, would the positive results of an experiment on the effects of a kindergarten Head Start program on the subsequent grammar school reading test scores of poor African American children in Memphis during the 1980s generalize to other test scores (e.g., mathematics) for poor children from other ethnic minorities in other cities at other times (Shadish et al., 2002)?

There are, however, two important exceptions. First, such generalizations can be derived deductively from theoretical statements of greater generality, a process in which no observations are involved. A second exception involves statistical

generalizations, for which the inferential process from probability samples to defined populations is again deductive. Much of the reasoning in mathematics and symbolic logic is deductive and takes the form of proofs. In statistics, the reasoning is likewise deductive (e.g., derivations of equations), but the conclusions based on analytic procedures are typically probabilistic (e.g., statistical decisions vulnerable to error, or estimating population parameters within ranges). What is central to the deductive methods for generalizing statistically to populations is the existence of information about sampling procedures and thus information about the likelihood, magnitude, and nature of errors (i.e., precision of estimation and bias).

Assumption 2: A Significant Experimental Treatment Effect Is Likely to Be Causally Generalizable

If the effect for a new treatment achieves the minimally acceptable significance level (i.e., $p < .05$), it is conventional to count it as a null hypothesis rejection, consider it an established finding, and treat it as publishable. In a field experiment, for example, Staines et al. (2004) compared two vocational counseling programs (experimental vs. standard) on the highest level of vocational activity achieved by unemployed methadone patients. Among other findings, the investigators reported that the advantage registered by the experimental group over the comparison condition was significant at $p < .05$ (two-tailed). At issue is whether such a finding is causally generalizable. Contrary to typical discussions of scientific findings, an isolated finding (i.e., one that has not yet been supported by replications) does not enjoy a high level of causal generality. In particular, an isolated finding often scores low on one precondition for causal generality, namely, exact (or literal) replicability (or reproducibility or repeatability). Exact replicability may be viewed as the most conservative (or narrow) form of causal generality because it minimizes the difference between the initial and follow-up studies. An acceptable level of exact replicability is necessary for causal generalization, and in general, the greater such replicability, the greater the causal generality.

The issue of exact replicability may be examined in terms of a simple research design. A posttreatment mean difference in a two-group experiment can be a real (or demonstrable) effect, or, alternatively, it can be a false positive result (i.e., a Type I error representing an accidental [or chance] difference). Fisher (1935, 1951, p. 14) noted that "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment [that] will rarely fail to give us a statistically significant result." Null hypothesis statistical testing offers a method of assessing replicability. According to Greenwald, Gonzalez, Harris, and Guthrie (1996, p. 176), "Two statistical tests can be said to replicate one another when they support the same conclusion (nonrejection or rejection in a specific direction) with respect to the same null hypothesis."

Replicability (or the power of an exact replication study) can largely be represented in terms of the initial study's p value. In other words, a p value resulting from null hypothesis testing is monotonically (but negatively) related to an estimate of a non-null finding's replicability (Greenwald et al., 1996). Exact replicability increases with (a) effect size, (b) sample size, and (c) the correlation between pretreatment and posttreatment scores on the outcome variable, but the contribution of these three factors to exact replicability is largely or completely captured by the relevant p value.

Although there are exceptions, such as large multisite clinical trials of treatments, exact replicability is easily and often overestimated by investigators whose research has generated a significant but isolated finding. As a result, there is reason to question the wisdom of immediately publishing an isolated finding. Assume, for example, that an initial experiment produces an effect with a two-tailed p value of .05. In an exact replication, when the expected effect size exactly matches the one needed to achieve $p = .05$, the probability of again rejecting the null hypothesis (in the same direction) is .5 (Greenwald et al., 1996). Assume further that an acceptable standard of replicability for an exact replication is .8 (i.e., the standard level of statistical power of .8). This level of replicability can be achieved only if the initial null hypothesis rejection had a p value of .005 or less (Greenwald et al., 1996). In practice, a researcher seeks a conceptual (as opposed to an exact) replication (i.e., a replication that is similar in terms of conceptually relevant components but

different in terms of conceptually irrelevant ones), and the resulting variations in the follow-up study entail that the p value of the original study's finding needs to be even smaller than .005 if the initial result is to be deemed conceptually replicable (i.e., able to achieve a null hypothesis rejection at $p < .05$ in a subsequent study). Table 1 summarizes the relevant significance levels for the original study and, likewise, for the follow-up study that is designed to produce a replication or generalization.

In short, when isolated findings score low on replicability (whether exact or conceptual), as many do, their level of causal generality is unlikely to be high, and researchers should not assume that they are probably generalizable. Stated differently, before investigators consider how to generalize a new null hypothesis rejection, they need to make sure that there is a replicable effect to generalize.

Assumption 3: The Magnitude of a Significant Experimental Effect Provides a Sound Estimate for Causal Generalization

Reports of experimental findings typically assume, whether implicitly or explicitly, that the obtained effect size, if significant, is the most accurate estimate of the true effect size and should be taken at face value. Such conventional assumptions, however, are misinformed. In the earlier example of vocational counseling for unemployed methadone patients (Staines et al., 2004), the effect size for the difference between experimental and control conditions on highest level of vocational activity obtained is not the most accurate estimate of the true effect size. The explanation lies in the fact that although some initial experiments will lead to null hypothesis rejections, many will not. Even if they involve causal relationships, studies that produce null hypothesis rejections are likely to capitalize on sampling error, so that an exact replication of an initial study will likely generate a smaller effect size. This shrinkage of effect magnitude for exact replications will generally be magnified in subsequent studies that test the causal generalization under other conditions (e.g., nonexact [or conceptual] replications). Thus, the true value of a causal effect is usually lower than the value obtained in the original study, meaning that the value obtained in the initial study typically contains an upward bias. Table 1 shows the effect sizes expected for replications and causal generalizations.

Assumption 4: The Arbitrary Nature of Convenience Samples Dictates That They Cannot Provide a Useful Basis for Causal Generalizations

A hypothetical study of the effect of motivational interviewing on alcohol-dependent patients at an urban hospital finds that at follow-up, drinking frequency is reduced by half a standard deviation more for the experimental group than for the control group. What grounds exist for generalizing this finding from a convenience sample to a specifiable population? In the absence of deductive probability sampling, one inductively based (or extrastatistical) approach to deriving causal generalizations from a convenience sample is to use purposive sampling strategies. Two purposive sampling methods, each of which lacks a statistical logic justifying formal generalizations, enhance causal generality: *representativeness* and *heterogeneity* (Matt, 2003; Shadish et al., 2002). The representativeness strategy implies that if clients in a treatment study are chosen to be typical of the

Table 1
Significance Levels and Expected Effect Sizes for Replication and Causal Generalization

Statistic	Exact replication	Conceptual replication	Causal generalization
Significance level required in initial study for subsequent replication or generalization	$p < .005$	Less than the p value required for exact replication	Less than the p value required for conceptual replication
Significance level required in follow-up study for replication or generalization	$p < .05$	$p < .05$	$p < .05$
Effect size expected in follow-up study for replication or generalization	Slightly less than the original study's effect size	Less than the effect size expected for exact replication	Less than the effect size expected for conceptual replication

population to which one wants to generalize, causal generality will be increased. An investigator's implementation of the representativeness strategy thus implicitly depends on his or her conception of the relevant population. In an evaluation of enhanced services for methadone patients, for example, selecting typical patients within a methadone clinic that is typical for a specified geographic area will facilitate the generality of the effectiveness findings.

The heterogeneity strategy calls for a high level of variability in the characteristics of the clients in the experiment. A heterogeneous sample will likely reflect many of the variations in the population to which one wants to generalize. Heterogeneity thus implies that causal generalizations will be more likely to be cases of interpolation and less likely to be cases of extrapolation; this is important because causal generality is enhanced more by interpolation than extrapolation. The rationale here is that interpolations are based on a pooling of two effect size estimates (i.e., one based on higher values and another based on lower values), whereas extrapolations rely on only one estimate (i.e., based on either higher or lower values; Matt, 2003). For example, interpolation and extrapolation can be applied to variations of treatment dosage in a psychological treatment evaluation. In predicting treatment effects for nonsampled values, interpolating to intermediate doses of treatment is less risky than extrapolating to extremely high or low doses. Sample heterogeneity is also an indirect route to sample representativeness. Sample heterogeneity is enhanced when minimally restrictive eligibility criteria are used so that a client sample can include more of the types of cases found in the relevant population. In trying to enhance internal validity, it may be tempting to exclude unusual, inaccessible, or uncooperative subgroups, but each such decision restricts causal generality (Westen & Bradley, 2005).

Although sample representativeness and heterogeneity strategies, when applied to convenience samples, may produce a more accurate estimate of the relevant population effect, the absence of any information on sampling error means that the extent of bias and the precision of estimation are unknown. Nonetheless, these purposive sampling strategies can be used for selection of factors other than experimental participants (e.g., treatment settings and therapists), whereas the notion of simultaneous probability sampling of participants, treatment settings, therapists, and so forth is essentially incoherent. Purposive sampling, then, is not the second best strategy for sampling all features of a treatment outcome experiment; it is the only viable one.

Assumption 5: Causal Generality Is Enhanced by Causal Explanations

To be sure, the presence of one and only one plausible and validated causal mechanism to explain a causal inference increases causal generality by strengthening the original finding in two ways: replicability (the follow-up study that validates the causal explanation typically provides a conceptual replication) and validity (the existence of a causal explanation validates the original treatment–outcome relationship). Additionally, as West et al. (2000, p. 56) explained,

> The causal explanation distinguishes the active from the inert components of our treatment package, and provides an understanding of the processes underlying our phenomenon of interest. These features permit us to specify which components need to be included in any new experimental context.

For example, it took 30 years to find a causal explanation for the relationship between the use of thalidomide as a sedative by pregnant women and the birth to infants with missing or stunted limbs. Thalidomide's ability to inhibit angiogenesis (i.e., blood vessel growth) caused limb defects in babies after maternal thalidomide usage. Thanks to this explanation, thalidomide can now be used to treat conditions characterized by uncontrolled angiogenesis (Matt, 2003).

In treatment outcome (and other behavioral) research, however, a single compelling explanation may be difficult to find (Cook, 1990, 1993). Although a single causal mechanism is often sufficient to provide an explanation, it is sometimes necessary to posit several partial causal mechanisms to formulate an explanation. When there is uncertainty regarding the relevant causal mechanism(s) (e.g., no causal mechanism has been identified, the causal mechanism identified forms only a partial explanation, or two or more competing causal mechanisms remain under consideration), the resulting doubt extends to causal generalizations. It is also often difficult to know how many causal mechanisms the explanation involves (e.g., whether a set of

partial causal mechanisms constitutes a complete explanation).

The uncertainties regarding the identification of the relevant causal mechanisms arise from several sources. First, the conceptualization of (causally relevant) unobservable variables may be imprecise (i.e., a theoretical issue). Second, the measures available to assess the unobservable variables may be unreliable, incomplete, or biased (i.e., a construct validity issue). Third, the level of measurement for unobserved variables is strictly ordinal at best, although measurement may approach the interval level. Fourth, the causal chain linking cause to effect may involve sequential mechanisms. To illustrate, in the study of the effect of vocational rehabilitation on employment outcomes among substance abusers, the intervention may lead to increased self-efficacy, which leads to participation in job training, which leads to achievement of competitive employment (Blankertz et al., 2004). The sequential issue raises questions about the number of links in the chain and the causal mechanisms involved in each link. In particular, the number of links may be partly a matter of judgment—that is, the extent to which the investigator chooses to subdivide the causal process into a set of stages. Given the difficulty in identifying the number and nature of causal mechanisms in psychological research, causal explanations may not enhance the generality of a causal effect to the desired extent.

Assumption 6: Self-Selection Bias Undermines Causal Generality

Now that informed consent is a staple of psychological research, research participants must be volunteers, and the generalization of causal effects from volunteers in a study to nonvolunteers (or to all clients) is a risk that plagues behavioral research because volunteers are unlikely to be sufficiently representative of nonvolunteers to permit generalization. Self-selection effects are illustrated in Coleman, Hoffer, and Kilgore's (1982) study of academic achievement in public and private (mostly Catholic) high schools. Comparisons among types of schools were confounded by selection factors. Far from being a random decision, whether a child attends a Catholic school is determined by parents and children (along with school admissions policies). Contributing factors include anticipated family income, religious views, the child's abilities, the school's reputation, and so on. To the extent that these factors independently affect high school achievement, they confound comparisons of achievement scores of parochial and public high school students (Rossi & Wright, 1984).

Still, there are circumstances in treatment services research when self-selection bias is a less serious problem. When interventions that have been tested experimentally with volunteers are implemented in regular clinical situations, they may or may not be mandatory for patients in the new setting. If they are mandatory, doubts inevitably arise about generalizing from volunteer samples to all patients because volunteers and nonvolunteers would be expected to produce different treatment outcomes. In particular, volunteers in a treatment evaluation study (who presumably display greater treatment receptivity) would be very likely to have higher treatment motivation than nonvolunteers. Whether their level of functioning is higher (or lower) than that of nonvolunteers likely depends on the type of treatment and the context. Although in some situations, it may be the more needy who volunteer, in substance abuse treatment studies those who volunteer to participate typically have higher levels of functioning than those who decline. In such cases, the combination of higher treatment motivation and better overall functioning makes volunteers more likely to be amenable to program participation and to achieve better outcomes.

In contrast, some experimentally tested interventions are introduced on a nonmandatory basis in regular clinical programs. These may include interventions that are advertised as mandatory but that are not forced on unwilling or uninterested clients because it would be counterproductive. For example, even when participation in vocational services is mandated at methadone clinics, as it increasingly is, only a minority of clients attend (Kidorf, King, & Brooner, 1999). It makes little sense to try to generalize from experimental samples of volunteers to the total population of clients because the effect of a program on hypothetical nonparticipants is not meaningful. In short, in the context of nonmandatory programs, the relevant sample inference is from those who volunteer for treatment in a research study to those who volunteer for treatment in a nonresearch situation. There

are likely to be some differences between these two groups that are a function of research requirements and the incentives to participate. Nonetheless, the selection bias in generalizing from volunteers for treatment in research to volunteers for treatment in a clinic should be lower than the bias in generalizing from research volunteers to all similar clients in a mandatory treatment situation.

Assumption 7: Judgments About Causal Generality Are Based on Similarity Between Studies

Discussions about causal generality often focus on the similarity between the initial study and some other situation. To illustrate, in asking whether the effects of a tough-love program for teaching unemployment skills to long-term unemployed people will succeed with parolees, an investigator has to consider whether the two populations are similar in causally relevant respects and different only in causally irrelevant respects (Matt, 2003). An analysis based on similarity, however, pertains only to targeted (causal) generality, that is, when a specific (but possibly hypothetical) population or circumstance is compared with the conditions in the original experiment. Targeted generality needs to be distinguished from overall (causal) generality, when the issue is whether one can generalize from an experiment's findings to any variation of participant population, study setting, outcome measurement, and so forth.

The methods for adducing evidence differ for overall versus targeted generality. As elaborated below, evidence of overall generality is largely limited to focusing on characteristics, especially strengths, of the original study. These include design issues (e.g., the various types of experimental validity [e.g., internal, statistical conclusion, or construct], statistical power, and scaling level of outcome measures; Brewer, 2000; Cook & Campbell, 1979) and also the nature of the statistical findings obtained (e.g., effect size, significance level, and availability of a validated causal explanation). Evidence of targeted generality includes an additional source of information, namely, the similarity between the initial situation studied and the situation targeted for generalization, in which similarity can refer to surface similarity (Campbell, 1986; Pawson & Tilley, 1997; Shadish et al., 2002), similarity via interpolation versus extrapolation (Matt, 2003; West et al., 2000), or structural similarity via causal explanations (Shadish et al., 2002; Vosniadou & Ortony, 1989). If assertions of targeted generality are supported by similarities between the two situations, it may not be necessary to have strong evidence of overall generality. Thus, the different factors that determine overall versus targeted generality may produce an asymmetry: Whereas evidence of overall generality may support targeted generality, the specific evidence of targeted generality may not provide much support for overall generality.

Assumption 8: Positive Findings Are Generalizable but Negative Findings Are Not

As suggested by the phrase "the error of affirming the null hypothesis" (Greenwald et al., 1996), it is common but inaccurate to assume that positive findings (i.e., presence of a substantial and statistically significant treatment effect, typically but not necessarily in the predicted direction) are generalizable, but that negative findings (i.e., absence of any evidence of a substantial and significant treatment effect [i.e., failure to reject the null hypothesis]) are not. Although causal generality is usually discussed in the context of positive findings, there is no need for any such restriction. Just as generalizing information about promising interventions is important, generalizing findings about the ineffectiveness of certain treatments can generate realistic expectations, conserve scarce resources, and encourage the development of better treatments.

For example, Rossi and Wright (1984, pp. 334–335) reviewed the major field experiments of the Great Society Program that were conducted in the 1960s and the early 1970s:

> These experiments covered a wide variety of topics: income maintenance plans intended to replace the existing welfare benefits system; housing allowances that might stimulate the market to produce better housing for the poor; health insurance plans that would not create perverse medical-care price effects,

only to conclude that "a reasonable summary of the findings is that the expected value of the effect of any program hovers around zero" (p. 341). Acknowledging such a pattern of negative

findings can usefully encourage efforts to determine the reasons for program failure and, likewise, can discourage efforts to repeat ineffective (and costly) types of interventions (Rossi, 1987).

Still, positive findings are often more generalizable than negative ones. One reason is that negative findings may arise from sources other than a flawed experimental hypothesis—what Greenwald et al. (1996, p. 177) called "the researcher's use of invalid research operations." Such weaknesses include problems with research design, treatment implementation, and measurement. By comparison, findings that support an experimental hypothesis suggest that the research methodology is adequate. In short, because positive findings suffer from less interpretational ambiguity, they can be more confidently generalized than negative findings. This greater informational value of positive findings is one reason why editors of scientific journals are more likely to accept papers that report positive results.

The overall advantage that positive findings have over negative findings needs to be qualified, however, because the relative generality of positive versus negative findings varies as a function of the receptivity of participants to the experimental treatment. In fact, treatment receptivity and positive versus negative findings (i.e., statistical significance) interact disordinally to influence causal generalization (West et al., 2000). More important, whereas generalizing positive outcomes will be most convincing when circumstances make achieving a significant treatment effect difficult, generalizing negative outcomes will have greatest force when circumstances are favorable to achieving a significant treatment effect. This overall principle follows logically (but trivially) from the meaning of treatment receptivity, but applying it inductively to new research situations requires assumptions about appropriate indicators of receptivity. On the basis of previously observed patterns, that is, researchers have expectations about the attributes that increase (or decrease) treatment receptivity in specific contexts. Yet investigators' assumptions about treatment receptivity in any particular context can always be mistaken.

To illustrate, client treatment motivation and overall functioning level, both of which typically increase receptivity to substance abuse treatment (Staines et al., 2003), have implications for causal generality that depend on whether the treatment findings are positive. If a study achieves positive results, low initial levels of treatment motivation and overall functioning among study participants augur well for the generality of the positive findings because the intervention may be more effective with other clinical populations that, for the most part, operate at higher levels of motivation and functioning. (Interestingly, this is one of the few cases in which an effect size might increase in a follow-up study.) By similar reasoning, causal generality of the positive results would be lower if the client sample had greater motivation and higher functioning. (This is the issue discussed earlier in terms of self-selection bias.)

Conversely, if an adequately powered study produces nonsignificant results, high initial levels of motivation and functioning among study participants would augur well for the generality of the negative findings because the intervention would likely be less effective with populations at lower levels of motivation and functioning. Similarly, causal generality of the negative results would be lower if client motivation and functioning were lower. It should be noted that not all client-based predictors of treatment receptivity are dynamic characteristics such as treatment motivation and level of functioning. For example, researchers in the substance abuse treatment field have generally agreed that certain stable patient attributes also predict poorer treatment outcomes. These include young age of onset of substance use, number of prior admissions to substance abuse treatment, number of years of addiction, and frequency of illicit substance use (Long, Williams, & Hollin, 1998; McLellan, Luborsky, Woody, O'Brien, & Druley, 1983).

To take an example of the relevance of treatment receptivity to causal generalization, unemployed methadone patients are long-term, hard-core drug addicts, who typically register poor outcomes when offered vocational (or other psychosocial) interventions (e.g., Coviello, Zanis, & Lynch, 2004; Lidz, Sorrentino, Robison, & Bunce, 2004). Yet Kidorf, Neufeld, and Brooner (2004) combined stepped-care approaches with behavioral reinforcement to motivate employment in opioid-dependent outpatients receiving methadone. A review of medical and billing records revealed that among patients

unemployed at intake ($N = 110$), the great majority (84%) had achieved full-time employment by follow-up—an unusually high success rate. Kidorf et al.'s positive findings for their stepped-care approach are more generalizable to less severely addicted populations of substance abuse patients than are the negative findings generally reported for other vocational interventions with unemployed methadone-maintained patients.

When viable, meta-analysis is also a useful strategy for addressing negative results. As is elaborated below, because meta-analysis increases statistical power, it can minimize Type II errors (i.e., failures to detect an effect). Furthermore, meta-analysis estimates effect size (or sizes) and can show treatment effects to be weak (i.e., small or even zero) without relying solely on the dichotomous results of null hypothesis testing (i.e., multiple failures to reject the null hypothesis).

Assumption 9: Causal Generality Is Properly Conceptualized in Terms of Cronbach's UTOS Formulation of Factors in an Experiment

Cronbach (1982) provided a useful classification of the four basic components of experiments: *units* (i.e., participants), *treatments* (interventions), *observations* (i.e., outcome measures), and *settings* (contexts), or UTOS. Matt (2003) pointed out that a possible fifth entity is time, which, if separated out from the concept of settings and added to the preceding list, draws attention to a study's historical context. To illustrate, Cronbach's UTOS formulation applies to the evaluation of psychotherapy as follows: units refer to patients; treatments refer to psychotherapeutic methods; observations refer to measures of symptoms, functioning, and so forth; and settings refer to clinics (or practitioners' offices).

The Cronbach formulation needs to be augmented, however, for treatment services research. Given that psychological treatments also involve providers in the form of therapists (or counselors), Cronbach's classification system needs to be expanded to five categories: clients, treatments (or methods), therapists (or counselors), outcome measures, and (treatment) settings. Causal generality, in short, may refer to one or more of these five categories. In comparisons between alternative treatments, it is important to separate the contribution of the provider (e.g., the counselor or therapist) from the contribution of the treatment method because therapists have different types and levels of skills and should not be viewed as interchangeable, just as research methods should not be considered interchangeable unless research has shown that they produce similar results (Elliott, Stiles, & Shapiro, 1993). Notwithstanding the well-documented evidence of treatment outcome differences among psychotherapists (Crits-Christoph et al., 1991; Elkin,1999; Martindale, 1978; Perry & Howard, 1989), many comparative treatment studies fail to distinguish between provider effects and treatment method effects. When therapist effects are erroneously interpreted as therapy effects, the rate of Type I errors is elevated. That is, the size, and the presence, of any treatment method effects reported will be exaggerated. If the appropriate analyses are not used, such confounds compromise the study's internal validity and, hence, its causal generality as well (Staines, Cleland, & Blankertz, 2006).

Assumption 10: Causal Generalization Is the Logical Next Step for Building on an Experimental Causal Finding

Causal generalization is a logical next step for building on an experimental finding, but it is not the only logical option. The major alternative to formulating a causal generalization is, as described above, to conduct a follow-up study designed to test an explanation of the observed effect. For example, if a study demonstrates the efficacy of a new drug, additional research designed to understand the micromediating processes of the drug on a molecular level can facilitate predictions about its beneficial effects for different health conditions (Matt, 2003). These two alternatives (formulating a causal generalization and testing a causal explanation) deserve clarification.

When an experiment is followed by a causal generalization, the generalization typically acts as an interim step between studies. In its simplest form, the basic research sequence has three steps from the perspective of the investigator of the follow-up study: (a) Study 1, (b) causal generalization, and (c) Study 2. First, the original study produces a causal finding. Second, the

initial investigator and other researchers make inductively based assumptions about which characteristics of the research study are causally relevant (i.e., affect whether the causal effect occurs) and which are causally irrelevant. They hypothesize that the findings generalize across variations in the purportedly irrelevant characteristics, and they raise questions about whether the findings generalize across variations in the characteristics judged to be relevant. Third, a follow-up study is designed. It may challenge the assumption about which study characteristics are irrelevant by testing whether a change in one or more of these characteristics produces a different result. Alternatively, it may accept the assumption about irrelevant characteristics and test whether the result holds up when one of the purportedly relevant characteristics is varied. As multiple follow-up studies accumulate, this three-step sequence helps to define the boundaries of the causal generalization. In short, causal generalizations are both a product of the initial study and a contributor to the design of follow-up studies. In evidentiary terms, causal generalizations are the weak link in the three-step process because they are assumptions rather than data-based findings. The sequencing of studies and causal generalizations is usually more complicated than in this simple example.

An alternative research approach is to follow the initial study with one designed to test a proposed causal explanation or compare two competing explanations. The follow-up study would usually be quite similar to the original one (i.e., often a conceptual replication). More important, it can strengthen or modify the causal generalization formulated on the basis of the original study. That is, a plausible and validated causal explanation both strengthens the original causal inference (and hence its overall causal generality) and refines information as to where the effect is likely to be repeated (i.e., modifies targeted generality).

Constructive Guidelines

As noted, this article provides an alternative to the flawed assumptions in the form of constructive strategies for assessing and enhancing overall (as opposed to targeted) causal generality. It builds on prior analyses of causal generalization, in particular Matt's (2003) review of taxonomies of principles for justifying and strengthening causal generalizations (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Cronbach, 1982; Cook, 1990, 1993; Shadish et al., 2002). I outline 10 such constructive strategies: (a) rigorous research design, (b) a priori tests of significance, (c) statistical power, (d) scaling level of outcome measures, (e) variable treatment dosages, (f) effectiveness designs, (g) multiple methods, (h) corroboration from observational studies, (i) synthesizing multiple studies, and (j) generalizing from multiple studies.

These 10 strategies, which are compatible with each other and can be pursued simultaneously by different investigators, exhibit important similarities and differences. Whereas the initial 7 strategies involve only the original research study, the remaining 3 concern the combination of results from more than one study. The fact that replicability enhances generality is relevant to whether the strategies are related to causal generality directly or have an indirect relationship that is mediated by replicability. Whereas all 10 strategies inform the assessment of causal generality, 6 also increase generality (i.e., rigorous research design, a priori tests of significance, statistical power, scaling level of outcome measures, variable treatment dosages, and effectiveness designs). These 6 produce more valid estimates (e.g., more precise and less biased) of the size, form, and significance of the treatment–outcome relationship and thus directly increase causal generality. Four strategies use multiple tests of the treatment–outcome effect (whether within the original study or across multiple studies) to produce more information on replicability and, thus, indirectly on causal generality (multiple methods, corroboration from observational studies, synthesizing multiple studies, and generalization based on multiple studies). Three of these 4 also use multiple tests to determine under what (other) conditions the original effect holds and thus directly provide information on causal generality (multiple methods, corroboration from observational studies, and generalization based on multiple studies). One strategy (statistical power) increases the likelihood of demonstrating replicability and thus indirectly provides more information about causal generality. These relationships between the 10 strategies and the concepts of replicability and

causal generality are summarized in Table 2 and described below.

Strategy 1: Rigorous Research Design

Compared with other research designs, experiments produce the most causally general results. Randomization, when combined with other components of methodological rigor (internal validity, statistical conclusion validity, and construct validity), produces more precise and less biased estimates of treatment–outcome effects, and such estimates are likely to be more generalizable than estimates based on less rigorous designs.

Strategy 2: A Priori Tests of Significance

To determine the initial study's replicability accurately, statistical tests of effects should be limited to independent, a priori predictions and adjusted (e.g., Bonferroni corrected) for alpha inflation. They should be further adjusted if multiple (correlated) tests are being conducted a posteriori. Any fishing expeditions for significant findings in the original study raise the Type I error rate well above the nominal alpha level of .05, and thus lead to overestimation of the design's replicability (Matt, 2003).

Strategy 3: Statistical Power

To provide the most accurate information about the replicability of the initial experiment's findings, research designs should have enough power (e.g., sufficient cases) to generate low p values (e.g., $p < .005$) if effect sizes are of the expected magnitude.

Strategy 4: Scaling Level of Outcome Measures

Ratio-level measures of outcomes increase causal generality directly by addressing the problem of arbitrary metrics in criterion measures. Blanton and Jaccard (2006, p. 28) defined a metric as arbitrary "when it is not known where a given score locates an individual on the underlying psychological dimension or how a one-unit change on the observed score reflects the magnitude of change on the underlying dimension." The use of arbitrary metrics in outcome measures in different studies may overstate the comparability of criterion measures. Such usage may overlook noncomparabilities that are a function of sample characteristics. These noncomparabilities are compounded when different studies use different arbitrary metrics. More important, treatment evaluations

Table 2
Direct and Indirect Strategies for Assessing and Enhancing Causal Generality of Treatment Effects

Strategy	Involves original study only vs. multiple studies	More accurate effect estimate from original study increases generality directly (enhancement)	Information about conditions under which effect holds increases information about generality directly (assessment)	Information about replicability increases information about generality indirectly (assessment)
Rigorous research design	Original	Yes	No	No
A priori significance tests	Original	Yes	No	No
Statistical power	Original	Yes	No	Yes
Scaling level of outcome measures	Original	Yes	No	No
Variable treatment dosages	Original	Yes	Yes	No
Effectiveness designs	Original	Yes	No	No
Multiple methods	Original	No	Yes	Yes
Corroboration from observational studies	Multiple	N/A	Yes	Yes
Synthesizing multiple studies	Multiple	N/A	No	Yes
Generalization based on multiple studies	Multiple	N/A	Yes	Yes

Note. N/A = not applicable.

can include ratio-level measures that are more directly interpretable and that make any noncomparabilities across studies transparent. A vocational counseling study can use ratio-level criteria such as time spent in employment, dollars earned in employment, and so forth. These ratio measures facilitate both replicability and causal generality because they increase comparability of the outcome measures between the initial study and subsequent studies. Additionally, if employment outcome measures need be adjusted to eliminate noncomparabilities arising from variations in labor market conditions (e.g., unemployment rates and cost of living), available data can be used to correct ratio-level measures of criteria.

Strategy 5: Variable Treatment Dosages

A parametric study (Kazdin, 1998) that uses interventions that cover much of the range of possible treatment dosages will be better able to estimate the treatment outcome effect than a study that aims at a constant treatment dose. For example, in Kidorf et al.'s (2004) stepped care service delivery intervention designed to motivate employment in outpatients on methadone, counseling was provided at three levels of intensity. In general, if treatment is differentially effective at different doses (e.g., if there is a direct linear relationship between dosage and outcomes), predictions about outcomes in other situations may be more accurate. This is another example of the advantages of using multiple methods when assessing causal generality. It is also an example of the advantages of basing causal generalizations more on interpolation and less on extrapolation.

Strategy 6: Effectiveness Designs

Kazdin (1998, p. 39) viewed efficacy and effectiveness designs in treatment evaluation studies as defining opposite ends of a continuum:

> *Efficacy* refers to treatment outcomes obtained in controlled psychotherapy studies that are conducted under laboratory and quasi-laboratory conditions (e.g., treatment is specified in manual form, recruited subjects are homogeneous and may show a narrow range of problems, and treatment delivery is closely supervised and monitored). *Effectiveness* refers to treatment outcomes obtained in clinic settings in which the usual control procedures are not implemented.

Interestingly, efficacy enhances replicability more than causal generality, whereas effectiveness enhances causal generality more than replicability. This partial tradeoff between replicability and causal generality arises because efficacy and effectiveness studies prioritize internal validity and external validity somewhat differently. Efficacy studies focus primarily on internal validity, and effectiveness studies concentrate on external validity and thus causal generality. Effectiveness studies are especially likely to have greater causal generality than efficacy studies if there is also evidence of replication—that is, both types of studies produce similar findings. For example, recent meta-analytic analyses (Shadish, Matt, Navarro, & Phillips, 2000; Shadish et al., 1997) have shown that psychotherapy outcomes are quite similar for efficacy studies (conducted in research laboratories) and effectiveness studies (conducted in clinics that approximate typical conditions of clinical practice). It is worth pointing out that the greater causal generality of effectiveness studies reflects the contributions of representativeness and diversity (i.e., purposive sampling) along multiple dimensions.

Strategy 7: Multiple Methods

Use of multiple methods in the study's design (i.e., critical multiplism; Cook, 1985; Matt, 2003; Shadish, 1994) tests for causal generality both indirectly (because the multiple methods test for replicability) and directly (because the variations in methods provide limited evidence concerning whether the causal relationship holds across different circumstances). The (indirect) contribution of multiple methods to causal generality via replicability requires clarification. A study that demonstrates the treatment effect in multiple ways (e.g., multiple sites, multiple clinics at each site, and multiple clinical staff at each site; outcome measures covering multiple domains; and multiple outcome measures in each domain) is less vulnerable to shrinkage because its overall pattern of findings would be less dependent on sampling error. Included here would be evidence of the convergent and divergent validity of its measures. For example, the use of different modalities in measuring outcomes (e.g., using biological specimens to supplement self-reports and collateral reports of illicit substance use) can ensure that findings are not biased by being

modality specific (i.e., mono-operation bias). The multimethod approach represents a within-study (or internal) replication. It can be expected to avoid site-specific sampling error, but it does not test for as many sources of sampling error as a true replication.

Multiple methods indicate whether treatment effects tend to be replicable (consistent null hypothesis rejections), nonreplicable (no null hypothesis rejections) or interactive (some null hypothesis rejections and some nonrejections). For example, if study findings are duplicated on all relevant outcome measures, there is evidence of replicability. If the results are duplicated on some but not all outcome measures, interaction between treatment and outcome measures is indicated. Also, even if the results for the total sample are significant, it is important to compare study findings across subgroups to test whether the evidence favors replication, nonreplication, or interaction. More specifically, possible interactions between the experimental treatment and a variety of potential moderator variables should be tested to see whether main effects hold for all subgroups (replication), hold for some subgroups but not others (interaction), or are converted into disordinal (or qualitative) interactions (i.e., treatment effects that change sign depending on the value of the moderator variable). Such interaction tests should provide information about the extent to which treatment effects are general versus contingent on other factors.

Strategy 8: Corroboration From Observational Studies

Consistent with Westen and Bradley's (2005) advocacy of multiple forms of evidence, observational data may provide complementary evidence (Matt, 2003), which can augment the generality of causal inferences both indirectly (via replicability) and directly (via null hypothesis rejections under different study conditions). In the indirect case, if experimental findings are corroborated by results from observational studies (i.e., correlational analyses of treatment effects), replicability (and, thus, generality) is enhanced because the probability that the original finding represents a false positive is substantially reduced (e.g., Reif, Horgan, Ritter, & Tomkins, 2004). Additionally, observational studies often include more cases and thus have more statistical power to detect potential causal relationships, and they may use probability sampling that allows statistical generalization to defined populations. The major weakness of the observational designs concerns internal validity; that is, these designs cannot rigorously rule out alternative causal explanations of findings that support the experimental hypothesis. This is less of a concern if the finding (i.e., a null hypothesis rejection) has already been obtained using a rigorous experimental design. Berk (2005, p. 16) summarizes the corroboration argument in his advocacy of "a mix of true experiments, quasi-experiments, and observational studies so that the comparative advantages of each can be exploited."

To illustrate, evidence of possible corroboration from an observational study comes from secondary analysis of data from the nationally representative Alcohol and Drug Services Study (1996–1999). Drawing on this dataset, Reif et al. (2004) studied 297 non–methadone outpatient clients with an identifiable need for employment counseling, as defined by a proxy measure of unemployment throughout the year before admission. The investigators compared those who received such counseling (met need, 42%) with those who did not (unmet need, 58%). Although met-need clients had significantly longer treatment duration and greater likelihood of employment postdischarge than unmet-need clients, the two groups were equally likely to complete treatment and to be abstinent at follow-up. Treatment effects, that is, were found for some but not all outcome measures. These observational findings could be compared with data from experimental studies of vocational counseling interventions for similar clients. The observational study would test the causal generality of the experimental results indirectly via possible replication of the findings for the relevant outcome measures. It could also augment information about causal generality directly because instead of representing just a different set of conditions, the treatment sample is nationally representative.

Strategy 9: Synthesizing Multiple Studies

Meta-analysis, a major methodological tool for obtaining causally general findings, uses two statistical models to synthesize results: fixed effects and random effects (Hunter & Schmidt,

2004). The fixed-effects model, the simplest and more commonly used meta-analytic model, combines data from a set of studies to yield aggregate findings. The model assumes that the same effect size underlies all the studies (i.e., the assumption of homogeneity). The fixed-effects model is thus most justifiable when the meta-analysis is designed to summarize the data in a given set of studies descriptively (i.e., to maximize internal validity and statistical conclusion validity). This fixed-effects model can be well suited to combining findings from multiple primary studies of the effects of psychological treatment (or other behavioral interventions). The technique pools results to increase the precision of treatment outcome estimates. It increases statistical power when individual studies are prone to Type II errors and have wide confidence intervals (Gilbody, Song, Eastwood, & Sutton, 2000). Additionally, examination of the individual studies provides information on replicability (and thus indirectly on causal generality).

Ideally, to establish internal validity, all individual cases from multiple studies of the same phenomena would be pooled in one analysis to maximize statistical power and precision. Various noncomparabilities among studies typically make pooling of individual cases impossible. Meta-analysis offers the next best statistical strategy for combining findings from a set of similar studies (i.e., substituting study-based data for individual-level data). In treatment services research, for example, meta-analysis is most effective in establishing internal validity when the studies pooled are of high quality and have documented equivalence in samples, diagnostic procedures, comorbidity, randomized treatment assignment, reliable and valid measures with equivalent reactivity, comparable settings, skilled therapists, and so forth (Klein, 2000). Such equivalence conditions, however, are rarely met in treatment services research (or in other research domains).

Strategy 10: Generalization Based on Multiple Studies

The random-effects model, by comparison, synthesizes results from multiple studies to yield causally general estimates of treatment effects; that is, it permits generalization to studies not included in the meta-analysis. The model allows for population parameters to vary across studies (i.e., the assumption of heterogeneity). Moderator variables can be identified so that subpopulations with different effect sizes can be defined. It is worth observing that if moderator effects are large, the overall meta-analysis may not be particularly informative. The random-effects model maximizes external (as opposed to internal) validity, and therefore facilitates causal generality. It is also important to point out that if, as happens all too frequently, the fixed-effects model is used when population parameters vary across studies, confidence intervals are erroneously narrow, and all significance tests have Type I biases (i.e., elevated above the nominal alpha level of .05). Although the logic of the random-effects model in meta-analysis is sound, the approach has definite limitations. Results are contingent on the number, representativeness, and heterogeneity of the convenience sample studies available for analysis. Furthermore, the search for moderator variables (e.g., subject characteristics) is confined to those that can be and are identified at the study level as opposed to only the individual-participant level.

Discussion

An analysis of overall (vs. targeted) causal generality requires an understanding of replicability. Exact replicability is a necessary but not a sufficient condition for inductively based causal generality. Because exact replicability is the narrowest instance of, and also a prerequisite for, causal generality, evidence of exact replicability enhances causal generality. Stated differently, low exact replicability detracts from causal generality. Researchers, who are understandably keen to generalize their findings to other situations, can easily overlook issues of replicability. An adequately powered initial study must achieve a p value of .005 (or less) to achieve exact replicability at the .05 level. Conceptual replicability at the .05 level requires an even lower p value in the original study, and causal generality at the .05 level typically requires a still lower p value. Additionally, because of sampling error, the effect sizes of follow-up studies are expected to be smaller than the original one (i.e., effect shrinkage).

This analysis of the replicability and generality of causal findings exemplifies a broader

methodological tradition in psychological research. Various pairs of methodological concepts, that is, are best conceptualized in terms of continuums. Replicability and causal generality are one such pair. They are usefully represented as the extremes (or anchors) of a continuum, with exact replicability referring to follow-up studies identical to the original study and causal generality referring to follow-up studies quite different from the original. Conceptual replication refers to a follow-up study that is designed to retest the original experimental hypothesis but that allows some variation from the original study's circumstances (e.g., changes in measures, sampling technique, field environment, and experimental design), where such deviations from the original study are not expected to alter the results appreciably (i.e., convert a null hypothesis rejection into a nonrejection). Conceptual replicability is thus located on the continuum but closer to the exact replicability end. Reliability and validity of measurement represent a second set of paired methodological concepts. Campbell and Fiske (1959) defined measurement reliability and validity in terms of a continuum, where reliability refers to correlations between measures that are the same (or very similar) and validity refers to correlations between measures that are substantially different. The replicability–generality and reliability–validity continuums are similar. The left-hand anchor of each continuum (replicability and reliability) refers to the repetition of an assessment, whereas the right-hand anchor of each continuum (generality and validity) refers to departures from the original assessment. In addition, efficacy and effectiveness form a third continuum that is also related to the replicability–generality continuum. Whereas efficacy studies, which emphasize internal validity, enhance replicability, effectiveness studies, which emphasize external validity, enhance causal generality.

Because an isolated causal finding is a weak basis for causal generalizations, there is reason for caution in basing a causal generalization solely on such a finding. Yet investigators who report an isolated causal finding often overestimate its magnitude, replicability, and generality. Multiple studies (along with multiple methods within individual studies) are therefore needed to bolster causal generalizations. Although a single study can provide estimates of exact replicability, that is, multiple studies can provide evidence concerning actual (conceptual) replications, which can be summarized using the fixed-effects model of meta-analysis. Multiple studies can also be used to produce causal generalizations based on the random-effects model. Accordingly, determining the extent of causal generality requires a major scientific investment, often involving the research of multiple investigators. "As Campbell and Stanley (1963) noted, we usually 'learn how far we can generalize an internally valid finding only piece by piece through trial and error' (p. 19), typically over multiple studies that contain different kinds of persons, settings" (Shadish et al., 2002, p. 86). Nowhere are such collaborative efforts as extensive as in research on treatment outcomes (Matt, 2003; Matt & Navarro, 1997). Furthermore, the credibility of a causal generalization depends on the degree to which researchers in the relevant field endorse the assumptions on which it is built, thereby making social processes and supporting empirical evidence both determinative of its acceptance (Matt, 2003).

References

Berk, R. A. (2005). *Randomized experiments as the bronze standard.* Retrieved March 6, 2008, from http://repositories.cdlib.org/ccpr/olwp/CCPR-030–05/

Blankertz, L., Magura, S., Staines, G. L., Madison, E. M., Spinelli, M., Horowitz, E., et al. (2004). A new work placement model for unemployed methadone patients. *Substance Use & Misuse, 39,* 2239–2260.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41.

Brewer, M. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3–16). New York: Cambridge University Press.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 66–77). San Francisco: Jossey-Bass.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Coleman, J. S., Hoffer, T., & Kilgore, S. (1982). *High school achievement: Public, Catholic, and private schools compared.* New York: Basic Books.

Cook, T. D. (1985). Post-positivist critical multiplism. In R. L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 21–62). 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; pp. 9-31). Washington, DC: U.S. Department of Health & 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* (New Directions in Program Evaluation No. 57, pp. 39–82). San Francisco: Jossey-Bass.

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

Coviello, D. M., Zanis, D. A., & Lynch, K. (2004). Effectiveness of vocational problem-solving skills on motivation and job-seeking action steps. *Substance Use & Misuse, 39,* 2309–2324.

Crits-Christoph, P., Baranackie, K., Kurcias, J., Beck, A. T., Carroll, K., Perry, K., et al. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1,* 81–91.

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

Elkin, I. (1999). A major dilemma in psychotherapy outcome research: Disentangling therapists from therapies. *Clinical Psychology: Science and Practice, 6,* 10–32.

Elliott, R., Stiles, W. B., & Shapiro, D. A. (1993). Are some psychotherapies more equivalent than others? In T. R. Giles (Ed.), *Handbook of effective psychotherapy* (pp. 455-479). New York: Plenum Press.

Essock, S. M. (2006). Enhancing generalizability: Stepping up to the plate. *Psychiatric Services, 57,* 141.

Fisher, R. A. (1951). *The design of experiments* (6th ed.). Edinburgh: Oliver & Boyd. (Original work published in 1935).

Gilbody, S. M., Song, F., Eastwood, A. J., & Sutton, A. (2000). The causes, consequences and detection of publication bias in psychiatry. *Acta Psychiatrica Scandinavica, 102,* 241–249.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Kazdin, A. E. (1998). *Research design in clinical psychology* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Kidorf, M., King, V. L., & Brooner, R. K. (1999). Integrating psychosocial services with methadone treatment: Behaviorally contingent pharmacotherapy. In E. C. Strain & M. L. Stitzer (Eds.), *Methadone treatment for opioid dependence* (pp. 166–195). Baltimore: Johns Hopkins University Press.

Kidorf, M., Neufeld, K., & Brooner, R. K. (2004). Combining stepped-care approaches with behavioral reinforcement to motivate employment in opioid-dependent outpatients. *Substance Use & Misuse, 39,* 2215–2238.

Klein, D. F. (2000). Flawed meta-analyses comparing psychotherapy with pharmacotherapy. *American Journal of Psychiatry, 157,* 1204–1211.

Lidz, V., Sorrentino, D. M., Robison, L., & Bunce, S. (2004). Learning from disappointing outcomes: An evaluation of prevocational interventions for methadone maintenance patients. *Substance Use & Misuse, 39,* 2287–2308.

Long, C. G., Williams, M., & Hollin, C. R. (1998). Alcoholism treatment: Intake variables as outcome predictors. *Addiction Research, 6,* 295–305.

Martindale, C. (1978). The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology, 46,* 1526–1530.

Matt, G. E. (2003). Will it work in Munster? Meta-analysis and the empirical generalization of causal relationships. In H. Holling, V. Böhning, & R. Schulze (Eds.), *Meta-analysis* (pp. 113–139). Berlin: Springer.

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.

McLellan, A. T., Luborsky, L., Woody, G. E., O'Brien, C. P., & Druley, K. A. (1983). Predicting response to alcohol and drug abuse treatments. *Archives of General Psychiatry, 40,* 620–625.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London: Sage.

Perry, K., & Howard, K. I. (1989, June). *Therapist effects: In search of the therapists' contribution to psychotherapy outcome.* Paper presented at the annual meeting of the Society for Psychotherapy Research, Toronto, Ontario, Canada.

Reif, S., Horgan, C. M., Ritter, G. A., & Tomkins, C. P. (2004). The impact of employment counseling on substance user treatment participation and outcomes. *Substance Use & Misuse, 39,* 2391–2424.

Rossi, P. (1987). The iron law of evaluation and other metallic rules. *Research in Social Problems and Public Policy, 4,* 3–20.

Rossi, P., & Wright, J. D. (1984). Evaluation research: An assessment. *Annual Review of Sociology, 10,* 331–352.

Shadish, W. R. (1994). Critical multiplism: A research strategy and its attendant tactics. In L. B. Sechrest & A. J. Figueredo (Eds.), *New directions for program evaluation* (pp. 13–57). San Francisco: Jossey-Bass.

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

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, G., Crits-Christoph, P., Hazelrigg, M. D., et al. (1997). Evidence that therapy works in clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65,* 355–365.

Staines, G. L. (2007). Comparative outcome evaluations of psychotherapies: Guidelines for addressing eight limitations of the gold standard of causal inference. *Psychotherapy: Theory, Research, Practice, Training, 44,* 161–174.

Staines, G. L., Blankertz, L., Magura, S., Bali, P., Madison, E. M., Spinelli, M., et al. (2004). Efficacy of the customized employment supports (CES) model of vocational rehabilitation for unemployed methadone patients: Preliminary results. *Substance Use & Misuse, 39,* 2261–2285.

Staines, G. L., Cleland, C. M., & Blankertz, L. (2006). Counselor confounds in evaluations of vocational rehabilitation methods in substance dependency treatment. *Evaluation Review, 30,* 139–170.

Staines, G. L., Magura, S., Rosenblum, A., Fong, C., Kosanke, N., Foote, J., & Deluca, A. (2003). Predictors of drinking outcomes among alcoholics. *American Journal of Drug & Alcohol Abuse, 49,* 203–218.

Vosniadou, S., & Ortony, A. (1989). Similarity and analogical reasoning: A synthesis. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 1–17). New York: Cambridge University Press.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–88). New York: Cambridge University Press.

Westen, D., & Bradley, B. (2005). Empirically supported complexity: Rethinking evidence-based practice in psychotherapy. *Current Directions in Psychological Science, 14,* 266–271.

Received April 12, 2007
Revision received August 30, 2007
Accepted September 4, 2007 ■