Appraising Evidence on Program Effectiveness

Norman A. Constantine and Marc T. Braverman

August 26, 2004

Citation:

This version of the chapter may not exactly replicate the final version published in the book. It is not the copy of record
Foundations and Evaluation: Contexts and Practices for Effective Philanthropy
Marc T. Braverman, Norman A. Constantine, and Jana Kay Slater (editors)
(available August 2004, Jossey-Bass)

Foreword
Richard T. Schlosberg III, former President and CEO, The David and Lucile Packard Foundation

Preface
The Editors
The Contributors

Introduction: Putting Evaluation to Work for Foundations
Jana Kay Slater, Norman A. Constantine, and Marc T. Braverman

Presidential Perspectives
Hodding Carter III, President and CEO, The John S. and James L. Knight Foundation
Michael M. Howe, President, The East Bay Community Foundation
Risa Lavizzo-Mourey, President and CEO, The Robert Wood Johnson Foundation

PART ONE: UNDERSTANDING FOUNDATIONS AS A CONTEXT FOR EVALUATION
1 Using Evaluation to Advance a Foundation’s Mission
Laura C. Leviton and Marian E. Bass
2 A Historical Perspective on Evaluation in Foundations
Peter Dobkin Hall
3 Foundations and Evaluation as Uneasy Partners in Learning
Mark R. Kramer and William E. Bickel
4 Building Strong Foundation-Grantee Relationships
Michael Quinn Patton, John Bare, and Deborah G. Bonnet
5 Evaluation as a Democratizing Practice
Jennifer C. Greene, Ricardo A. Millett, and Rodney K. Hopson
6 Integrating Evaluation into Foundation Activity Cycles
Laura C. Leviton and William E. Bickel

PART TWO: BUILDING CAPACITY FOR EVALUATION PRACTICE
7 Making Judgments About What to Evaluate and How Intensely
Melvin M. Mark and William Beery
8 Adapting Evaluation to Accommodate Foundations’ Structural and Cultural Characteristics
Ross F. Conner, Victor Kuo, Marli S. Melton, and Ricardo A. Millett
9 Field-Based Evaluation as a Path to Foundation Effectiveness
Patricia Patrizi and Edward Pauly
10 Strategies for Smaller Foundations
Marli S. Melton, Jana Kay Slater, and Wendy L. Constantine
11 Strategies for Comprehensive Initiatives
Debra J. Rog and James R. Knickman
12 Appraising Evidence on Program Effectiveness
Norman A. Constantine and Marc T. Braverman
13 Evaluative Thinking for Grantees
E. Jane Davidson, Michael M. Howe, and Michael Scriven
14 Communicating Results to Different Audiences
Lester W. Baxter and Marc T. Braverman

Index
Appraising Evidence on Program Effectiveness\textsuperscript{1,2}

Norman A. Constantine and Marc T. Braverman

Rarely do we find simple or conclusive answers to our questions about the effectiveness of a program intended to address a complex social problem. Yet well-designed and skillfully implemented evaluations can and often do provide compelling evidence in this regard. To make best use of this evidence requires a healthy dose of constructive skepticism, as well as an understanding of common sources of ambiguous results and misleading conclusions that frequently pervade program evaluations.

Our intent in this chapter is to provide some basic guidance on the critical appraisal of program effectiveness evidence. We begin with a discussion of the nature of ambiguous results and misleading conclusions, followed by an illustration based on an influential national study on adolescent virginity pledge programs. Next, we discuss essential tools for strengthening effectiveness evaluations — experimental and quasi-experimental designs, and program theory. Well-known evaluations from the areas of youth violence prevention and teen pregnancy prevention are used to illustrate challenges often encountered in the use of statistical significance tests to draw conclusions about program effectiveness. We then make a case for transparency and accessibility in providing full evaluation results for critique. Discussions follow of interpreting null or negative results and of the cumulative nature of knowledge about program effectiveness. The chapter closes with a reiteration of the importance of critically examining evidence about program effectiveness.


\textsuperscript{2} We gratefully acknowledge the constructive skepticism and helpful suggestions provided by Melvin Mark, Wendy Constantine, Mark Lipsey, and Joel Moskowitz, and the indispensable library support provided by Mieko Davis.
Although we invoke a few technical concepts in the course of our discussion, the ideas and examples presented rely primarily on the processes of carefully reasoned argument. We hope that this chapter will be useful to readers of diverse training and roles, and that it will support the critical appraisal of effectiveness evidence among evaluation professionals and consumers alike.

*Evaluation* can be defined broadly as *the systematic collection and use of information to answer questions about programs.* Depending on each unique program and stakeholder situation, different types of questions are commonly addressed within this overarching evaluation framework, including questions about the need for a program and about program development, implementation, improvement, effectiveness, and potential reach, among others. Several alternative perspectives have been articulated on what evaluation is and what it should be (Shadish, Cook, and Leviton, 1995; Mark and Smith, 2001; Cronbach, 1982). These perspectives vary on a number of dimensions, including the emphasis placed on direct questions about program effectiveness, that is, trying to determine whether a collection of program activities caused desired outcomes in a specific sample of participants in a particular setting and context, and to what extent results can be generalized to similar but nonidentical program activities, recipients, settings, and outcomes. Although certainly not the only types of questions that evaluation can deal with, these are the fundamental questions of program effectiveness evaluation.

**Ambiguous Results and Misleading Conclusions**

At heart, questions about program effectiveness are about causation: Did a program *cause* the intended results? And answering questions about causation is seldom a straightforward task. This complexity can open the door to results that are ambiguous and conclusions that are spurious or misleading. We distinguish here between *ambiguous results* as
those that are open to multiple conflicting conclusions and *misleading conclusions* as those that are inappropriately supported by either dubious results or the questionable use of results.

Ambiguous results are common in program effectiveness evaluation, as in all areas of scientific inquiry—social and natural sciences alike. Depending on the nature of the questions asked, the resources available, the skill of the evaluation team, and many other factors, effectiveness evidence and conclusions will be of greater or lesser certainty, but rarely does any one study conclusively answer an effectiveness question. It is the nature of science that confidence in a hypothesized relation (such as between a program and a potential effect) must develop over time and through replications, and always remains subject to possible modification or refutation. Science advances iteratively, often through skepticism, criticism, and debate. The same is true for evaluation. Under the right circumstances, good outcome evaluations do provide useful evidence to inform program investment decisions, especially when results are interpreted within the context of a strong theoretical grounding together with previous evaluations of similar programs. Depending on the particular situation, the necessary level of confidence required for our results and conclusions will vary. Over time and through additional studies of similar programs, these conclusions might be either weakened or strengthened; at times, a program investment decision might require reconsideration.

Divergent values and ideological views within our society about social problems and their solutions can foster the distortion of effectiveness evidence, especially when the evidence is ambiguous. For example, after decades of evaluation on programs to prevent adolescent pregnancy and sexually transmitted infections, deep ideological divisions among proponents of abstinence-only versus comprehensive sexuality education programs appear to have hardened. Both sides appeal to program evaluation “evidence,” often ambiguous at best, to make their case (see several illustrations that follow). Another motivation for the misuse of evidence is the real or imagined pressure on grantees and external evaluators to show positive results and definitive conclusions that will please the funders (Braverman and Campbell, 1989; Moskowitz,
Biases that might result can range from the largely unconscious to the blatantly obvious. And we are all potentially susceptible — evaluators, program staff, foundation program officers, and other evaluation consumers.

To further complicate these challenges, modern statistical and methodological tools are becoming increasingly difficult for nonspecialists to understand. Most evaluators can’t keep up with the full range of methods available, while statisticians often specialize in esoteric areas or techniques. When complex statistical methods are employed, evaluation consumers might be inclined to “trust the expert” as a coping strategy and assume on faith that the evaluation results and conclusions are valid. In truth, complex statistics can sometimes strengthen a good evaluation design, but the principle doesn’t hold for a weak one. As Light, Singer, and Willett (1990, p. v) note, "You can’t fix by analysis what you’ve bungled by design." Fortunately, evaluation designs can be easier for a nonexpert to understand and critically appraise than are complex statistical methods; in many cases, this will be the most important area to assess when considering the validity of any conclusions about effectiveness.

Causation, Correlation, and Alternative Explanations

The essential nature of causation and the types of evidence necessary to demonstrate a causal relation, such as the effect of a program on an outcome, have been long debated (e.g., Bunge, 1979; Mackie, 1980; McKim and Turner, 1997). In spite of these ongoing debates, it is safe to say that a pragmatic view of causation is most appropriate to intervention effectiveness studies. Most program outcomes of interest are the result of numerous and interacting causes—some that are potentially changeable (such as home environment) and some that are much less so (such as genetic influences). What we expect of the best interventions is to partially influence some outcomes, under specific conditions and circumstances, by modifying one or more causal factors. But how do we know when this has happened?
According to the nineteenth-century philosopher John Stuart Mill, at least three criteria must be invoked in justifying casual claims: (1) association (or correlation—the cause is related to the effect), (2) temporality (the cause comes before the effect), and (3) elimination of plausible alternative explanations (other plausible explanations for an effect are considered and ruled out). The key is that all three are necessary, yet sometimes the second and often the third of these criteria are neglected in the design and interpretation of evaluation studies. And even when the third criterion is explicitly addressed by the evaluation, it is often arguable whether or not a sufficient number of the most likely plausible explanations have been considered.

It is widely recognized that correlation does not necessarily imply causation, yet erroneous causal attributions are commonly made based on association or correlation alone. Consider the potential conclusion that adolescents’ levels of psychological attachment to their families are a cause of observed differences in problem behavior levels, based on a correlation between these two variables. Although this conclusion might in fact be valid, the correlation alone does not provide sufficient supportive evidence for its validity; any number of alternative explanations could fit the observed relationship. For example, lower levels of problem behavior might strengthen family attachment rather than the other way around. Or a third factor, such as patterns of parental conflict, might independently influence both attachment and problem behavior.

When both of Mill’s first two conditions (association and temporality) hold, it can be even more tempting to erroneously infer causation without considering other plausible explanations. As an example, consider the National Longitudinal Study of Adolescent Health (commonly known as the Add Health study) (Resnick and others, 1997). This large correlational study has yielded uncountable associations among adolescent behaviors, background conditions, health outcomes, and other factors. And because it was longitudinal, involving linked measurements

---

3 In the spirit of constructive criticism and rational debate, written responses will be sought from the authors of all studies critiqued in this chapter, and posted at www.crahd.phi.org/evidence.
over time from the same participants, some of these associations have been examined for the *temporality* expected for a cause-and-effect relation. Yet little effort has been invested in addressing the third critical criterion for causality: identifying and ruling out plausible alternative explanations.

A compelling illustration is provided by a widely publicized Add Health study conclusion that virginity pledge programs “cause virginity,” that is, delay initiation of sexual intercourse (Bearman and Bruchner, 2001). Complex statistical methods, such as survival analysis and logistic regression, were used to reach this conclusion. Several qualifications regarding the program setting were appropriately discussed, most notably that to have an effect, the pledge must occur in a community of other pledgers that is neither too small nor too large relative to the total student population in the school. The authors, however, neglected sufficient consideration of plausible alternative explanations. Foremost among these would be the possibility that a pre-existing disinclination to initiate sex might have been a primary causal factor behind both signing the pledge and delaying intercourse. If true, this alternative explanation implies that signing the virginity pledge serves as a marker to identify those youth who delay intercourse for any number of other reasons and that, in the absence of pledging, the pattern of sexual initiation would be largely unchanged. This alternative arises from the likelihood of a strong self-selection effect, meaning that participants determine for themselves whether they will be part of the intervention group (in this case, those who pledged) or the control group (those who did not). It is likely that pre-existing differences between those who chose to pledge and those who didn’t—most notably differences in the intent to delay intercourse—are not only related to the intervention group assignment but are arguably among its most important determinants.

A statistical adjustment procedure intended to remove the effect of self-selection was described in an appendix to the article, but this procedure was both logically and statistically inadequate (see Pedhazur and Schmelkin, 1991, pp. 295-296 for a discussion on the futility of this type of adjustment). Instead, the researchers’ conclusions regarding a pledge effect suggest
a criterion for causality of *post hoc ergo propter hoc* (“after this, therefore because of this”), a fundamental fallacy of logic, known since classical times, that involves inferring a causal relation on the basis of correlation and temporality alone.

This study, nevertheless, has generated extensive media coverage and policy discussion (see, for example, Boyle, 2000; Nesmith, 2001; Schemo, 2001; Willis, 2001) and has had a substantial influence on federal policy about sexuality education. Prior to this study, the U.S. Department of Health and Human Services had required as performance measures for the evaluation of federally-funded abstinence education programs “the proportion of program participants who have engaged in sexual intercourse” and the birth rate of female program participants (Federal Register, 2000). Two years later, on the heels of extensive media attention to Bearman and Bruchner’s (2001) study, these sexual behavior and birth rate measures were replaced with the “proportion of youth who commit to abstain from sexual activity until marriage” (Department of Health and Human Services, 2002). Thus, virginity pledging has become the primary behavioral outcome to be measured.

If one reads the various critiques and summaries of the pledge study and its conclusions, it is remarkable to find no mention of the obvious plausible alternative explanation of a pre-existing disinclination among pledgers. Instead, the critiques tend to focus on the limited conditions under which the intervention is believed to be effective and the negative side effects observed (for example, that pledgers who break the pledge were less likely to use contraception than nonpledgers). Yet in considering the original question—Do virginity pledges cause initiation of sexual intercourse to be delayed?—the answer remains that they might or might not. This particular study adds little or nothing to our knowledge of this wished-for effect.

**Experimental and Quasi-Experimental Designs**

The pledge study example sets the stage for a brief review of experimental and quasi-experimental study designs. A more comprehensive overview of this topic is provided by
Reichardt and Mark (1998), and a definitive coverage can be found in Shadish, Cook, and Campbell (2002). The critical missing design element in the pledge study was a controlled manipulation, that is, random or other controlled assignment of the pledge program to some schools or classrooms and not others. Random assignment would be characteristic of a true experimental design, whereas nonrandom assignment strategies could be part of a quasi-experimental design. By contrast, the virginity pledge study design was purely correlational, in that no manipulation of intervention delivery across schools, classrooms, or other units took place. With a good experimental or quasi-experimental design, the plausible alternative explanation for the virginity effect could have been ruled out or rendered unlikely, and then the potential effectiveness of pledging could have been examined more appropriately. The admonition “no causation without manipulation” (commonly attributed to Paul Holland) might be somewhat exaggerated for effect, yet it is a useful heuristic for raising a red flag whenever one encounters claims of program effects based on self-selected participation in an intervention program. Correlational designs do have a variety of appropriate and important uses, such as developing hypotheses to be tested in subsequent studies. However, their utility in eliminating plausible alternative explanations is limited.

Both experimental and quasi-experimental designs are intended to help address common threats to internal validity, that is, causal attribution of effects to the program. Eight of these threats are discussed by Shadish, Cook, and Campbell (2002), including the selection-related threat illustrated by the virginity pledge study. Briefly, these are

- Selection (pre-existing differences between intervention and control groups that could explain the effect)

4 The skeptical reader might ask whether path analysis or structural equation modeling approaches to the analysis of correlational data can be used to demonstrate causal relationships. We briefly note that these analysis can be applied to any type of design, correlational, experimental, or quasi-experimental; that they typically suffer from measurement error and model misspecification error problems; and that in any case, “… the model does not “confirm” causal relationships. Rather it assumes causal links and then tests how
• Ambiguous temporal ordering (which variable occurred first?)
• History (extraneous events occurring during the intervention that could explain the effect)
• Maturation (naturally occurring changes in participants over time that could explain the effect)
• Regression (natural movement on subsequent measurements toward the overall group average -- applicable for groups selected on the basis of extreme scores)
• Attrition (differential loss of participants in the groups studied)
• Testing (practice or other effects based on exposure to the assessment instrument)
• Instrumentation (changes in the function or meaning of the measures used over time or between groups)

Some or all of these threats to validity can be addressed by various randomized-experimental and quasi-experimental designs. However, the two types of designs differ in important ways and are appropriate in different circumstances.

*Randomized experiments* (sometimes referred to as true experiments, randomized controlled trials, or randomized field trials) involve the random assignment of units (for example, persons, schools, clinics, or communities) to intervention versus comparison conditions. This is done to control or minimize some of the major potential threats to validity and can be especially powerful in reducing or eliminating selection effects.

*Quasi-experiments* do not involve random assignment to intervention and comparison conditions but employ some combination of other design features to help rule out alternative explanations of observed effects. These include nonrandom assignment strategies such as matching or stratifying, scheduling of multiple pre- and post-intervention measurements, multiple treatment and comparison groups, and intervention timing. A variety of quasi-experimental designs involving one or more of these strategies have evolved over time (Campbell and strong they would be if the model were a correct representation of reality.” (Shadish, Cook, & Campbell,
Stanley, 1966; Cook and Campbell, 1979) and are discussed in depth by Shadish, Cook, and Campbell (2002).

It is sometimes stated that randomized experiments are the “gold standard” for evaluating program effectiveness. In some ways, this might be true, but with important contextual qualifications (perhaps “bronze standard” would be a more appropriate designation). In the real world of intervention program environments, quasi-experimental designs are often the more cost-effective choice—sometimes the only logistically possible approach—and frequently can provide a basis for equally or more valid causal conclusions than a randomized experiment would provide. For example, an experiment in which some participants refuse to accept their randomized assignment, as often occurs, can yield a weaker design than a good quasi-experiment. And a randomized experiment is often contraindicated until a program has sufficiently matured. As Shadish and colleagues (2002, p. 277) caution: “Premature experimentation can be a great waste of resources—indeed it can undermine potentially promising interventions for which there has not been time to develop recruitment procedures, identify and fix implementation problems, and serve the clientele long enough to make a difference.”

Conversely, in some situations a smaller-scale randomized experiment might appropriately be conducted under ideal circumstances at one site prior to implementing a larger-scale, quasi-experimental multisite design under more realistic circumstances (see Glasgow, Lichtenstein, and Marcus, 2003). The key is that both randomized and quasi-experimental designs offer the potential to reduce the likelihood of some or many alternative explanations for a suspected effect, but to be realized, this potential requires skillful application under the right circumstances and conditions. It is the evaluator’s responsibility to argue and sufficiently document the case that an appropriate design has been skillfully applied. And it is the evaluation consumer’s responsibility to critically appraise this argument, together with the
associated results and conclusions. In spite of the potential power and elegance of both randomized and quasi-experiments, neither offers immunity to ambiguous results and misleading conclusions. To assume that results and conclusions are valid simply because they arise from a particular type of evaluation design perpetuates a common but inadequate substitute for critical appraisal.

Examples of Randomized and Quasi-Experimental Designs

The Hutchinson Smoking Prevention Project (Peterson and others, 2000) illustrates an exemplary (and expensive) long-term randomized trial of a school-based tobacco-use-prevention intervention. In this rigorous evaluation, forty Washington school districts were randomly assigned to the intervention or the control condition. More than eight thousand third-grade students were enrolled and followed until two years beyond high school. High program implementation fidelity was achieved, and 94 percent of the enrolled students were retained through the last follow-up assessment. No significant differences were found in smoking rates between intervention and control students, and the authors conclude with confidence that “there is no evidence from this trial that a school-based social-influences approach is effective in the long-term deterrence of smoking among youth” (p. 1979).

Although the Hutchison study provides a compelling illustration of the randomized trial as “gold standard,” the gold here might refer not only to the quality of the study but to its cost as well. It would be a mistake, unfortunately, to believe that this level of rigor is typical among randomized evaluations in school and community intervention studies.

As a more fiscally modest example, consider the Around the Clock Mobile Crisis Intervention evaluation (Reding and Raphelson, 1995). This was a much less expensive but wonderfully elegant local evaluation of a community intervention program. It was based on a creative combination of quasi-experimental design elements. Here, the design elements of multiple measurements over time (interrupted time series), intervention timing (introduction and
withdrawal of the intervention), and an intact comparison group were efficiently combined to provide compelling evidence of intervention effectiveness (discussed in more detail in Shadish, Cook, and Campbell, 2002, pp. 188–189). Arguably, the conclusions of this quasi-experimental evaluation are just as convincing as are those of the randomized Hutchison study, for a small fraction of the cost. Of course, the costs of these two studies were not set arbitrarily but depended on the evaluation questions, the nature of the interventions, and their contexts.

Program Theories and Social Science Theories

Strictly speaking, an outcome evaluation need only consider the effects of an intervention on one or more outcomes, without addressing how or why these consequences occurred. Yet outcome evaluations that are built on a clearly specified program theory can have many advantages (Lipsey, 1993; Cook and Shadish, 1994). Perhaps most important, a good program theory “enables evaluators to eliminate rival hypotheses and make causal attributions more easily” (House, 2001, p. 311). Furthermore, as Hughes (2000, p. 324) explains,

A willingness to entertain rival interpretations, an ability to place knowledge within broader contexts, and an openness to new ways of conceptualizing problems are essential to scientific inquiry. Theory serves these functions as well as directs inquiry, unifies and systematizes knowledge, and makes sense of what (might) otherwise be inscrutable empirical facts.

In general, a program theory is the specification of what must be done to achieve the desired program goals, what side effects might also be produced, and the mechanisms through which these goals and effects are generated (Chen, 1990). More specifically, a complete program theory specifies several critical elements: (1) the problem condition to be addressed, (2) the intended intervention populations and relevant background circumstances, (3) the critical program components and their interrelationships, (4) the key aspects of the process by which the program achieves its effects, (5) the intended and potential unintended effects of the
intervention, and (6) the interrelationships among these effects (Lipsey, 1993). To help illustrate a program theory, the hypothesized interrelationships among the various theory elements are sometimes portrayed in a diagrammatic format, commonly referred to as a causal diagram.

A program theory is sometimes based on an existing social science theory of change. Glanz, Rimer, and Lewis (2002) provide a comprehensive review of social science theories applicable to health behavior and health education interventions. As an illustration, the theory of reasoned action (Fishbein and Middlestandt, 1989) specifies a network of influences among beliefs, attitudes, perceived social norms, behavioral intentions, and behaviors, and has been used to design and evaluate intervention programs in HIV-AIDS prevention and other areas. Constantine and Curry (1998) documented the process of a school-based violence prevention evaluation in which a program theory was elicited through a series of participatory activities with program staff and other stakeholders, and overlaid on a higher-level theoretical framework derived from the theory of reasoned action.

More commonly, however, the program theory is built around the operational assumptions and expectations of the program's designer, with little or no explicit connection to an existing social science theory. Instead, it reflects a thoughtful analysis of how and in what contexts the program is expected to work and what intermediate events need to happen if it is to be successful. This type of locally developed program theory is often referred to as a logic model or theory of change.

Uses and Abuses of Statistical Significance Testing

In a recent critique, Gorman (2002) describes a randomized trial of the Second Step violence prevention curriculum published in the Journal of the American Medical Association (Grossman and others, 1997). Based on this evaluation, the program was certified as an exemplary program by a U.S. Department of Education expert panel (Safe, Disciplined, and Drug-Free Schools Expert Panel, 2001). The department's guidelines for a program to be
classified as “exemplary” include the criterion of “at least one evaluation that has demonstrated an effect on substance abuse, violent behavior, or other conduct problems one year or longer beyond baseline” (Safe, Disciplined, and Drug-Free Schools Expert Panel, 1999, p. 5). As a result of this certification, a program is recommended for use in federally funded Safe, Disciplined, and Drug Free Schools programs and exempted from requirements of further outcome evaluations at the local level. The problem that Gorman points out, however, is that the results of the published study are much more consistent with chance findings than with evidence of program effectiveness. The intervention and control groups were compared for statistically significant differences on a total of twenty outcomes (for example, teacher-reported aggressive behavior, parent-reported social skills, etc.) at one-year post-baseline. With the statistical significance level (alpha) set at .05, only one of the twenty comparisons showed the groups to be significantly different from each other – exactly the result that would be expected by chance alone if, in fact, there were no effects of the program.

As another example, consider Reducing the Risk -- a prevention curriculum designed to reduce the incidence of teen pregnancy and sexually transmitted infections (Barth, 1989), developed in part through foundation funding. This curriculum also was recognized by a federal review panel, in this case through the Centers for Disease Control and Prevention as a “Program That Works.” This recognition required that a program had been “proven effective in reducing HIV risk behaviors” (Centers for Disease Control and Prevention, 2001). The “Programs That Work” designation was assigned to Reducing the Risk, based on an outcome evaluation (Kirby, Barth, Leland, and Fetro, 1991) that included thirty-two initial significance tests of risk-behavior outcome comparisons (involving nine behavioral outcome measures and various combinations of time of measurement and subgroups based on prior sexual experience levels). Of these, three were significant at the .05 alpha level, involving only the students who were sexually inexperienced at the time of program entry. Thirty-six additional outcome comparisons were conducted to test effects within gender, ethnicity, and risk-level subgroups for
the three significant outcomes; of these additional tests two were significant at the .05 alpha level.

Up to this point, sixty-eight significance tests of potential differences between the intervention and control groups were conducted, yielding five statistically significant results. Once again, this is very close to what would be expected by chance in the face of no program effects and given the specified statistical significance criterion of .05. Moreover, a reader who carefully studies the article will be able to identify numerous additional comparisons that were tested but were not systematically reported, further compromising any conclusions of program effectiveness based on this study.

All of this is not to say that these two evaluations show their respective programs to be ineffective. Absence of evidence is not evidence of absence (this theme is explored further in the section on interpreting null or negative effects). In fact, the Reducing the Risk evaluation yielded an overall pattern of differences between the intervention and control groups suggestive of potential behavioral effects within some subgroups that would be worthy of further study. Yet for neither evaluation do the results justify conclusions of “proven effectiveness” according to the standards of evidence associated with the statistical significance testing approach employed, nor by the criteria of the federal review groups that certified the two programs.

The practice of testing large numbers of potential outcomes for statistical significance, while ignoring the increasing likelihood of finding spurious effects as more tests are added, is often referred to as a scattershot approach or a fishing expedition. When only the significant results are reported and the nonsignificant tests are held back, the term cherry picking is sometimes applied (see, as an example, the replication evaluation of Reducing the Risk [Hubbard, Giese, and Rainey, 1998]). And published fishing expeditions are often cherry picked at a later stage in summary reports or other distillations provided by external review panels, program advocates, or program critics.
These problems are not unique to the Second Step or Reducing the Risk evaluations.
For example, Dar, Serlin, and Omer (1994) report that only 21 percent of 111 reviewed studies of psychotherapy effectiveness compensated even minimally for performing multiple significance tests. We used these particular evaluations as illustrations because of the direct influence they have had on public policy and the fact that they are erroneously considered by many to be methodologically rigorous outcome evaluations with unambiguous conclusions of program effectiveness.

Appropriate Strategies for Multiple Significance Tests

There are three well-developed strategies for addressing the multiple significance test problem: (1) conduct fewer and more focused significance tests, (2) employ a more stringent statistical significance criterion (that is, the alpha level) for those that remain, and (3) supplement statistical significance testing with alternative indicators of effects.

Conduct Fewer and More Focused Significance Tests

The best way to avoid engaging in a fishing expedition is to use the program’s theory to inform the specification of a small number of tests of theory-derived hypotheses regarding expected differences on particular outcome measures. The theory also should guide the designation, in advance of conducting the analyses, of a few selected tests as primary, allowing the others to be designated in advance as secondary or exploratory. The primary tests are those involving the outcomes the program is betting on—the ones that will be most important in reaching conclusions about a program’s effectiveness. The program theory also can be used to specify particular subgroups of the sample in which to conduct additional tests of significance, rather than testing all possible combinations of subgroups. Subgroup tests should be used cautiously and, in general, only conducted in two situations: (1) following a significant full-group test or (2) if suggested by the program theory or past results and planned in advance of seeing the data, in place of full-group tests.
Employ a More Stringent Statistical Significance Criterion

To control for the likelihood of erroneously obtaining statistically significant results by chance, it is sometimes recommended to make the significance criterion more conservative by dividing the desired alpha level by the number of significance tests to be performed. Thus if one were conducting twenty tests and desired an alpha level of .05, each test would then be conducted at a more conservative level of .0025 (.05/20). This strategy, known as the Bonferonni correction, works well in studies with large samples where statistical power is strong. In smaller studies, however, it can overcorrect and excessively reduce the test's power of finding a potential effect to be significant (Shaffer, 1995).

The overcorrection problem can be addressed by applying corrections separately to smaller subsets of significance tests that are logically defined. This might involve content-specific families of tests. For example, if a study employs thirty measures equally divided among knowledge, attitude, and behavior outcomes, these three content categories could define three families of ten outcome tests each. Corrections then could be based on the ten tests per family rather than the total thirty tests, thereby reducing the risk of overcorrection.

Another way to define families of tests would be through nesting of subgroups following a significant full sample test. For example, if for a particular outcome a full sample difference is found to be statistically significant, then within subgroup tests for this same outcome could be corrected only for the number of subgroups, again reducing the risk of overcorrection. The logic and methods involved in these and other approaches to correcting for multiple significance tests is discussed by Shaffer (1995). In general, these approaches tend to be good compromises between uncorrected multiple significance tests and the overcorrection that often results from a blanket Bonferonni correction across all tests.

Supplement Statistical Significance Testing with Alternative Indicators of Effects

Another important strategy involves going beyond the usual reliance on statistical significance tests and the arbitrary .05 alpha level to focus on other indicators of program
effects. A movement has been growing among methodologists to do so (for example, Cohen, 1994; Wilkinson and the American Psychological Association Task Force on Statistical Inference (1999), although its effect on practice is developing slowly. At a minimum, an effect size should be calculated to indicate the actual size of an effect in a metric that can be compared across studies. A commonly used effect size is the difference between the intervention and control group means, divided by the combined group standard deviation (Cohen, 1969). Discussions of this and other measures of effect size can be found in Lipsey and Wilson (2001) and Rosenthal and DiMatteo (2001). A confidence interval should be provided around each effect size, indicating its likely lower and upper bounds, given a specified degree of confidence, typically 95 percent (which is equivalent to a .05 alpha level). Another advantage of providing effect sizes is that this facilitates later studies, or meta-analyses, that attempt to quantify average findings across multiple studies of the same or similar interventions. It is hard to imagine an outcome evaluation for which reporting effect sizes and confidence intervals would not be appropriate.

It is also important to consider the practical meaning of a particular effect in terms of its clinical, programmatic, or policy significance. For example, one supplemental measure that might be employed to help assess clinical significance is the proportion of individuals in the intervention group versus the control group who reach a specified level of normal or healthy functioning (Jacobson and others, 1999). This provides a fundamentally different type of knowledge than does the statement that a significant difference between the two groups was found on a measure of healthy functioning.

In most outcome evaluations involving multiple statistical significance tests, evaluation consumers should expect to see some combination of the strategies discussed above. Often all three can be applied together. This is illustrated in the Infant Health and Development Program (1990) study, a foundation-funded national randomized evaluation of a child development intervention. Eight primary outcome variables were selected early in the evaluation (and prior to
any data analysis) from a pool of several hundred measured outcomes, and primary versus secondary tests were clearly distinguished in the published report. A correction was applied to the desired significance level of .05 for these eight primary tests, yielding a more stringent corrected criterion of .006. Finally, effect sizes were calculated for all outcomes, and these were given equivalent emphasis to the statistical significance tests in the report.

**Transparency and Accessibility**

A preliminary draft of the virginity pledge study discussed earlier was released by the authors in July, 2000—six months prior to its official publication date (January 2001) in the *American Journal of Sociology*. However, this printed journal was not mailed to libraries or subscribers until June of 2001, creating, in effect, a one-year interval in which the report was extensively discussed in the national media and among policy advocates on both sides of the issue, yet without access to the final published version of the article. Many discussants were limited to commenting on the press release or the few selected details reported in the media reports.

As another example, consider a legislatively mandated evaluation of a statewide teen pregnancy prevention program in California (Cagampang and others, 2002). An executive summary of this foundation-funded evaluation was widely disseminated by the state of California six months prior to the public release of the full report. This evaluation contained several policy-relevant but questionable claims of positive program effects, yet by the time the full report was released the opportunity for meaningful discussion and debate had largely passed (see Constantine, 2003, for a methodological critique of this evaluation and its conclusions). These two cases are not isolated examples. Indeed, it can be argued that this type of situation, involving release of evaluation conclusions to the media or directly to policymakers prior to availability of the documentation necessary to critically appraise the conclusions, is becoming increasingly prevalent and might already represent the norm. Why might this be so?
During the last decade, health and social program evaluation has grown to unprecedented levels. Academic journals have proliferated, while other means of disseminating findings, including electronic methods, have grown exponentially. At the same time, the popular media appear to have insatiable appetites for reporting abbreviated findings from “the latest studies”. When it comes to information, however, more often means less (Brown and Duguid, 2000). In many ways, we are drowning in information and increasingly challenged to use this information effectively. At a minimum, this requires separating the proverbial wheat from the chaff, that is, critically appraising the validity of evaluation conclusions and the evidence they rest upon. Yet when the popular media serve as the gatekeepers, critical appraisal of validity is often replaced by intuitive appraisal of newsworthiness or, increasingly, of potential entertainment value. And the piecemeal release of summaries or selected results can deprive legitimate stakeholders of the information needed to critically appraise results and conclusions.

There are no easy answers to this challenge. Publication of an evaluation in a peer-reviewed scientific journal provides a stamp of approval, yet peer review can be a notoriously slow process, with low inter-reviewer agreement commonly found among peer reviewers (Cicchetti, 1991; Cole, 1992; MacCoun, 1998; Peters and Ceci, 1982). And, in any case, peer-reviewed journal publication is beyond the means of many smaller evaluations. Some foundations use their own external or internal review panels. Depending on the ability and objectivity of the reviewers, this can be a very constructive approach. Often, however, we must appraise conclusions without the benefit of prior professional and rigorous review.

Program evaluation results released to (and reported by) the news media or policymakers without the benefit of a published and readily available peer-reviewed article, or minimally a publicly available comprehensive report, should raise a red flag. At the same time, it is important to recognize that peer or other professional reviews are not a sufficient condition to establish the validity of reported conclusions. Executive summaries might be all that a busy professional is willing and able to read, but summary authors are remiss if they do not provide a
Web link or other convenient access to a full report that includes a thorough description of the context, methods, assumptions, limitations, and complete results of the evaluation. Ultimately, the evaluation consumer must take responsibility for critically appraising evaluation results, or else finding a trusted and qualified colleague to provide this service. In either case, timely access to a comprehensive report is essential.

**Interpreting Null or Negative Results**

So far, we have discussed some of the issues involved in interpreting program evaluations that appear to provide evidence in support of program effectiveness. There are also challenges associated with making sense of evaluations that find no evidence of the intended effects (null results) or effects that are opposite of those intended (negative results). It can seem natural to conclude from such a study that the program is not effective, and critics of a controversial program might be especially inclined to embrace this conclusion. However, such a stance is often premature. In the face of no-effect findings, a careful examination of the study is warranted to understand which alternative explanations are most likely (Lipsey, 1993).

Several possible reasons for a failure to find positive results lie with the evaluation study itself. First, a weak evaluation design can obscure the effectiveness of a program. A common design weakness is the use of unreliable measures or measures lacking sufficient validity or sensitivity for their evaluation use. As Lipsey (1988) concluded in a review of a random sample of 122 published evaluation studies, “In most studies the quality of the outcome measure cannot be assumed, nor is it typically demonstrated in the study itself” (p. 15). Another common design weakness is insufficient statistical power. Power relates to the likelihood of an actual program effect being accurately captured in a test of statistical significance (as opposed to a “false negative” finding). Power problems often result from too few participants, as well as from use of measures with low reliability or validity. A study with inadequate power is likely to fail to demonstrate the statistical significance of actual effects or to yield effect size estimates that are
imprecise. It has long been recognized that insufficient power is a serious and widespread
day problem in the evaluation literature, even among published studies (Lipsey and others, 1985).

A second set of problems relates to the implementation of the evaluation: the evaluation
might have been well designed but poorly conducted. An example would be inadequate training
of the study’s data collectors, especially if the data collection process requires substantial skill,
such as for conducting individual or group interviews. Another example applies to evaluations
that have long timeframes, involving multiple waves of data collection over time. These
evaluations are vulnerable to attrition (participants dropping out of the evaluation) from the later
data points, and to the extent that such attrition occurs, it potentially weakens the ability of the
study to identify program effects (alternatively, the study might yield spurious positive results). In
evaluations with multiple data-collection points, efforts must be made to motivate continued
participation and to track participants who move, as well as to analyze and disclose the potential
effects of the attrition on evaluation results (see Constantine and others, 1993).

Several other possible explanations for no-effect results are related to the program itself.
First, a potentially effective program might have been poorly implemented—a possibility that
occurs often in the field. Program staff might not have been adequately trained, participants
might not receive the full intervention, or there might be large variations in implementation
across different program sites. Such problems produce an inadequate test of the basic program
theory.

Intervention programs rarely remain in a constant state over time. In their development and
implementation, they typically progress through a maturation process (Berk and Rossi, 1999).
Immediately after the planning phase, in the initial period of implementation there is often a
considerable amount of unstructured experimentation going on. During this time, the program
can be evolving rapidly. A program that grows in this way will not serve its clientele most
effectively until it has reached a certain stage of maturity, at which point its operations become
more stable. It would be of questionable value to subject this type of program to a full-scale
effectiveness evaluation during its first year, although this sometimes happens. In interpreting
the meaning and usefulness of an outcome evaluation, two questions related to a program’s
developmental trajectory should be asked:

- Was this evaluation conducted at an appropriate time in the development of the
  program?
- What is the relevance of this evidence for understanding the program in its current
  form?

These questions do not speak to the quality of the evaluation itself but rather to the
relationship of the evaluation to the program and to its developmental course.

Finally, the null effect finding might instead result from a failure of the program theory—
the concept underlying the program might be inadequate, even if program implementation and
evaluation implementation are strong. Failure of program theory can result from an insufficient
understanding of the critical intervening processes, leading to an emphasis on the wrong kinds
of program activities. This can also result from inadequate specification of the program’s most
appropriate target population.

Following findings of no effects, a comprehensive review of the results from the
perspective of the program theory can inform the program designers in making necessary
adjustments to the program, its delivery, or its evaluation so that an improved version can be
tested. Alternatively, they might decide that the challenges are too fundamental to justify the
investment of further resources in another round of program refinement and evaluation.
Whatever course of action is followed, the decisions leading up to it should involve a full
consideration of all relevant factors.
Understanding Results in a Broad Context

In most areas of scientific inquiry, the replication of results across multiple studies is an important component of the knowledge-generation process. Similarly, evaluation studies should be interpreted in light of the larger context of what else is known about the intervention and the program theory behind it. An individual study represents one particular set of decisions about how to measure the critical variables and how to time the observations, as well as one implementation, for better or worse, of the observation procedures. Furthermore, if the study involves only one program delivery site, it also represents only a single instance of program implementation. These factors—measurement, timing, program implementation, and others—often can be varied without undermining the basic theory of the program. Thus a single evaluation study, compelling though it might be, should be considered only one piece of a gradual process of knowledge accretion. Conclusions from an evaluation cannot be considered robust until there is evidence that the findings would be similar under varying contexts and conditions of program implementation, or documentation of the specific conditions, settings, and populations associated with effectiveness.

Sometimes there may be a large discrepancy between the findings of a given evaluation and the findings that were expected, based on past studies. A study may show dramatic results when previous studies have shown little effect, or it may show no results when other evaluations were positive. These discrepant patterns can be potentially interesting and revealing. They call for a more careful examination of the findings and the program theory, often suggesting the need for further investigations.

With these considerations in mind, one important question that arises when interpreting findings is where to go next. A good study typically raises new questions as surely as it answers existing ones. Sometimes it will be useful to pursue new or refined evaluation questions, systematically varying aspects of the program or the evaluation design. In other cases, a study may be so consistent with previous investigations that it is deemed fully justified to make
important program-related decisions on the basis of what has been learned up to the present point. In all these cases, examination of the evaluation findings within a broad context, that is, beyond the single study, provides a stronger level of confidence for whatever decisions or courses of action are eventually pursued.

Conclusions

In this chapter we have covered a range of issues around appraising evidence about program effectiveness and using this evidence with validity and integrity. Our discussion has been necessarily incomplete. There are many other issues that an evaluator must master and an evaluation consumer needs to understand in order to do full justice to the business of making sense of effectiveness evidence and conclusions. We have presented some of the most important and frequently encountered issues — those that often lead to ambiguous results or misleading conclusions. Our aim is to raise awareness of some of these challenges, to motivate foundation evaluators and evaluation consumers to ask the right questions about effectiveness evidence and conclusions, and to provide a basic understanding to help support constructive skepticism and critical appraisal.

Program effectiveness evaluation is built largely on social science methodology. As with most science-based endeavors, it provides imperfect answers to our questions about reality. Yet these answers can improve over time with continued study and debate. For program effectiveness evaluation to realize its full potential, it is incumbent upon evaluators and evaluation consumers to demonstrate competence in logical and critical approaches to appraising evidence. And this includes, at its core, asking the right questions and considering the answers with appropriate skepticism. As Patton (2002, p. 11) advises,

*Evaluator and the people we work with have to critically and thoughtfully examine the evidence that is purported to be “scientific” and draw their own conclusions...*

*Consumers of evaluations need to attend to “truth in packaging.” Look beyond the label*
or assertion that some proposal is “science-based” to examine where the evidence comes from and what it really shows.

Only when we heed this advice are we able to reap the powerful potential benefits that effectiveness evaluation has to offer.
References


S. F. Schneider (Eds.), *Primary prevention of AIDS: Psychological approaches* (pp. 93–110). London: Sage.


