THINKING ABOUT PROGRAM EVALUATION

A Unified Protocol
for
Federal "What Works" Clearinghouses

Douglas J. Besharov

prepared as part of

The White House Task Force for Disadvantaged Youth
Final Report
October 2003

Welfare Reform Academy
University of Maryland
American Enterprise Institute
1150 Seventeenth Street, N.W.
Washington, D.C. 20036
www.welfareacademy.org

Douglas J. Besharov is the Joseph J. and Violet Jacobs Scholar in Social Welfare Studies at the American Enterprise Institute, and a professor at the University of Maryland School of Public Policy.

Recommendations

Recommendation*: A task force of the relevant federal agencies should develop a government-wide approach to the evaluation of youth programs and policies, including agency-specific protocols. Because individual agencies have different needs, the protocols need not be identical, just sufficiently homologous so that materials and findings can be shared among agencies with relative ease.*

Recommendation: *Agency protocols should (1) identify the types of programs or policies whose evaluations should be assessed and (2) specify the evaluation methodologies that are acceptable (i.e., only random assignment experiments or also nonexperimental evaluations). The protocols should establish a priority-setting system for deciding which evaluations to assess. (For programs or policies that cut across agencies, joint or coordinated efforts should be considered.)*

Recommendation*: The agency protocols should establish a formal process of evaluation that specifies the criteria for assessment and the levels of evidence available. The process should be formalized, with written guidelines and data collection instruments, and it should be open and transparent and subject to outside review. The protocol should explicitly address whether nonexperimental evaluations and meta-analyses will be assessed—and under what conditions or with what limitations.*

Contents

I. Introduction 1

II. Evaluations to Assess 4
- Random assignment experiments 4
- Nonexperimental evaluations 8
- Programs or policies that "do not work." 12
- Meta-analyses 13

III. The Assessment Process 16
- Criteria for assessments 16
- Specified levels of evidence 18
- A formal process 18
- Open and transparent 19
- Outside review 19

Evaluating the Evaluations of Social Programs 21
- Program Issues 21
  - Program theory 21
  - Program implementation 22
- Causal Validity 22
  - Assessing the randomization 22
  - Assessing statistical controls in nonexperimental evaluations 23
- Data Issues 24
  - Sample size 24
  - Attrition 24
  - Data collection 24
  - Measurement issues 25
- Interpretation 26
  - Analytical models 26
  - Generalizability 26
  - Replication 26
  - Evaluator's description of findings 27
  - Evaluator's independence 27
- Policy Significance 27
  - Statistical significance/confidence intervals 27
  - Effect size 28
  - Sustained effects 28
  - Benefit-cost analysis 29
  - Cost-effectiveness analysis 29
  - Assessing meta-analyses 30

I. Introduction

***Recommendation**: A task force of the relevant federal agencies should develop a government-wide approach to the evaluation of youth programs and policies, including agency-specific protocols. Because individual agencies have different needs, the protocols need not be identical, just sufficiently homologous so that materials and findings can be shared among agencies with relative ease.*

Many federal agencies are developing research-based efforts to identify youth programs that "work," broadly called "'What Works' Clearinghouses." For example, the Department of Education maintains the "What Works Clearinghouse,"[1] the Department of Justice the "Blueprints for Violence Prevention" program,[2] the Substance Abuse and Mental Health Services Administration (SAMSHA) the "National Registry of Effective Programs" (NREP),[3] and the Office of Management and Budget (OMB) the "Program Assessment Rating Tool" (PART).[4]

Whatever the name of such efforts, the idea is the same: Social science findings should guide government decisions about which programs to support and at what funding levels, the content of technical assistance, and the additional research that is needed. In fact, making government decisions more evidence-based should be a major priority in this area. As recent OMB efforts demonstrate, relatively few youth programs supported by the federal government meet this test.[5]

Unfortunately for a "what works" effort, there is not a sufficiently large body of research (or evaluation) that has established the effectiveness of a relatively broad set of youth programs (or approaches) upon which important program choices can be based. Studies of youth programs

---

[1]U.S. Department of Education, What Works Clearinghouse, "About the WWC," available from: http://www.w-w-c.org/about.html, accessed August 5, 2003.

[2]University of Colorado, Center for the Study and Prevention of Violence, "Blueprints for Violence Prevention Overview," available from: http://www.colorado.edu/cspv/blueprints/index.html, accessed June 30, 2003.

[3]U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, "SAMHSA Model Programs: Effective Substance Abuse and Mental Health Programs for Every Community," available from: http://www.modelprograms.samhsa.gov, accessed August 19, 2003.

[4]Richard P. Emery, Jr. to Program Associate Directors, memorandum, 5 May 2003, "Completing the Program Assessment Rating Tool (PART) for the FY2005 Review Process," Office of Management and Budget, available from: http://www.whitehouse.gov/omb/part/bpm861.pdf, accessed August 20, 2003.

[5]The White House Task Force for Disadvantaged Youth, *The Federal Response to Disadvantaged Youth: April 2003 Preliminary Report* (Washington, D.C.: Author, April 2003), pp. 32-36, *stating*: "6 of the 28 youth programs (21 percent) rated were scored as 'ineffective' by the OMB PART system. Thirteen youth-related programs were rated as 'results not demonstrated.' Five were ranked as 'adequate.' Three were rated as 'moderately effective.' Only a single youth program (Consolidated Health Centers, which addresses disadvantaged youth as only one part of the population it serves) was given the highest rating, 'effective.'" [The White House Task Force for Disadvantaged Youth, *Preliminary Report on Findings for the Federal Response to Disadvantage Youth: April 2003* (Washington, D.C.: The White House Task Force for Disadvantage Youth), April 2003, p. 34.]

rarely have sufficient rigor and, when they do, they are usually of such limited applicability that they cannot be the *sole* basis of broad policy planning.

Jodie Roth and her colleagues describe the limitations of this research: "The review of the evaluation literature highlights the paucity of high quality outcome evaluations of programs fitting the youth development framework. As noted previously, little improvement in the state of program evaluation has occurred since the 1992 Carnegie Report . . . Nationally, strong interest in expanding adolescents' access to youth development programs exists. However, the current mismatch between the enthusiasm for these programmatic efforts and the empirical evidence calls into question the effectiveness of such efforts."[6] Rob Hollister adds, "what do we know about what works—our answer has to be: *not much*."[7]

Thus, if what works efforts in the field of youth development were limited to programs of *proven* effectiveness, as understood by rigorous social scientists, they would not be sufficient to guide government (or practitioner) decision making. But studies need not be perfect in order to be useful. Research projects entirely without flaws do not exist and, arguably, never will. Almost every evaluation is compromised by conceptual, programmatic, funding, time, or political constraints. No program is implemented with absolute fidelity to the original design. No sampling plan is without faults. Every data set is missing some observations and data. Analytical procedures are always misspecified to some degree. In other words, evaluation findings are only more credible or less so, and even poorly designed and executed evaluations can contain some information worth noting.

The key issue is the extent to which a discerned fault reduces the credibility of an evaluation. But determining that is both a complicated and subjective process. The absence of a reasonably broad set of definitive findings means that most assessments must be based on incomplete data—and on numerous inferences that are derived from both evidence and theory from other areas of social welfare research. This makes all such assessments to some extent subjective, and at least somewhat tinged by the reviewer's social and political views. That is why the actual process of assessment is so important, and why it should include procedural and substantive safeguards.

Unfortunately, most policymakers and practitioners, as well as the general public, are ill-equipped to judge which faults are fatal, especially since they often must act before the traditional scholarly process can filter out invalid results. This is understandable, since assessing

---

[6] *Quoted in* Rob Hollister, *The Growth in After-School Programs and Their Impact* (Washington, D.C.: The Brookings Institution, February 2003), p. 12, available from: http://www.brookings.edu/views/papers/sawhill/20030225.pdf, accessed August 21, 2003.

[7] Rob Hollister, *The Growth in After-School Programs and Their Impact* (Washington, D.C.: The Brookings Institution, February 2003), p. 12, available from: http://www.brookings.edu/views/papers/sawhill/20030225.pdf, accessed August 21, 2003.

evaluation studies often requires both detailed knowledge of the programs involved and a high level of technical expertise. To help them better assess this research and glean the lessons it offers, there needs to be a government-wide effort to assess evaluations of youth programs and policies that uses generally accepted criteria for judging evaluations. (The criteria, of course, are not equally applicable to all evaluations.)

Federal agencies are beginning to develop such what works efforts. Reflecting their recent origins, most of these what works efforts are still in their formative stages, with individual agencies now grappling with how best to proceed. And, reflecting the fact that they have been developed largely from within specific federal agencies, they often lack common terminology and methodologies. To some extent, of course, such diversity is needed to reflect each agency's specific needs, disciplinary framework, and statutory and programmatic context. Nevertheless, greater commonality would facilitate the sharing of information among federal agencies and with the public, and would allow federal agencies to build on each other's efforts.

This paper describes (1) the basic elements of a unified protocol for federal what works clearinghouses and (2) how more detailed and agency-specific protocols might be developed.

II. Evaluations to Assess

Recommendation: *Agency protocols should (1) identify the types of programs or policies whose evaluations should be assessed and (2) specify the evaluation methodologies that are acceptable (i.e., only random assignment experiments or also nonexperimental evaluations). The protocols should establish a priority-setting system for deciding which evaluations to assess. (For programs or policies that cut across agencies, joint or coordinated efforts should be considered.)*

A first order question is what kinds of evaluations or research studies should be included in a what works clearinghouse. The short answer is all the evaluations that would give a full picture of what is known about the effectiveness of youth programs. That means evaluations with sufficient scientific rigor whether or not they show that a program or policy "works" or "does not work."

Each agency should develop a system for assessing the evaluations of the programs and policies under its jurisdiction, establishing a priority-setting system for deciding which evaluations to assess. For programs or policies that cross agencies, joint or coordinated efforts should be considered.

Random assignment experiments. Unfortunately, many studies that seek to evaluate the effectiveness of youth programs do not have causal validity, that is, their design does not support causal inferences.

Many social welfare programs look successful—to their own staffs as well as to outsiders—because their clients seem to be doing so well. A substantial proportion of trainees, for example, may have found jobs after having gone through a particular program. But did they get their jobs because of the program, or would they have gotten them anyway?

Any number of factors, however, could have caused the observed result: The economy may have improved, making more jobs available (and perhaps increasing the demand for low-skilled workers); the participants may have been especially amenable to help (or not amenable); or they may have gotten their jobs because of the passage of time (and perhaps the normal process of maturation). Determining what would have happened in the absence of the program or policy is the central task of impact evaluation. To do so, researchers try to establish the "counterfactual," that is, they try to see what happened to a similar group that was not subject to the program or policy.

Most social scientists believe that experimental designs are the best way to measure a program or policy's impact—because they can have high causal validity. In an experimental design, individuals, families, or other units of analysis are randomly assigned to either (1) a

"program group,"[8] whose members can take part in the program or are subject to the policy, or (2) a "control group,"[9] whose members do not. The experience of the control group, thus, is meant to represent what would have happened but for the program or policy, that is, the counterfactual.

If properly conducted, random assignment should result in statistically comparable program and control groups, that is, groups whose aggregate characteristics (measurable and unmeasurable) are comparable (within the limits of chance variation). This similarity means that the two groups are likely to be exposed to the same outside forces and to respond to those forces in similar ways, so that any subsequent differences in average outcomes can be attributed to the program or policy—to a known degree of statistical precision. This ability to rule out other causes gives randomized experiments a high degree of causal validity.[10]

Because experimental designs ordinarily do not require complex statistical adjustments to eliminate differences between program and control groups, they gain a credibility (and accessibility) that often gives their results substantial influence over policy. Policymakers can then focus on the implications of findings, rather than "become entangled in a protracted and often inconclusive scientific debate about whether the findings of a particular study are statistically valid."[11] For example, the evaluations of welfare-to-work programs conducted by Manpower Demonstration Research Corporation (MDRC) in the 1980s—which used experimental designs—are widely credited with having shaped the Family Support Act of

[8]We use the term "program group" because it seems to encompass all the other variations on the same concept, including "experimental," "treatment," and "intervention" groups—as long as they have been randomly assigned to the group. We do not use "experimental" group because in an evaluation of an ongoing program, the term would erroneously suggest that something new is being tested. In addition, the term would not apply in the case of a nonexperimental evaluation. Similarly, we do not use the terms "treatment" or "intervention" group, because the program or policy being tested may not be perceived as a "treatment" or "intervention" by those participating in an ongoing program.

[9]The terms "control" group and "comparison" group are sometimes used interchangeably. However, to emphasize the difference between randomized experiments and nonexperimental evaluations, we limit the use of "control" group to the nonprogram groups created by random assignment. We are careful to call the nonprogram groups in nonexperimental evaluations "comparison" groups.

[10]The ability of an evaluation to establish a cause-and-effect relationship between an outcome and a program or policy being evaluated is usually called "internal validity." This paper, however, adopts a variant of a revised term proposed by Donald Campbell: "causal validity." Campbell's full description of term was "local molar causal validity." *See* Donald T. Campbell, "Relabeling internal and external validity for applied social scientists," in W. M. K. Trochim (editor), *Advances in quasi-experimental design: New directions for program evaluation* (San Francisco, CA: Jossey-Bass, 1986), pp. 66-77.

[11]Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives* vol. 9, no. 2, Spring 1995, p. 69.

1988.[12] So, too, in the 1990s, for the Abt Associates evaluation of the Job Training Partnership Act (JTPA) program.[13] As a result, experimental designs are increasingly used to evaluate a wide range of social programs and policies.[14]

The superior causal validity of randomized experiments has led most experts in program evaluation to call them the "gold standard" of evaluation.

Of course, the credibility of a particular randomized experiment depends on its being well implemented. Besides the problems attendant to all survey research, as Rossi, Freeman, and Lipsey state, "the integrity of a randomized experiment is easily threatened."[15] For example, differential attrition between the program and control groups could undermine their comparability. Or, program staff may not explain the program or policy that affects each group properly, possibly leading study participants to behave differently than if they knew the conditions that applied to them. And, the program may not be implemented as intended, so that the comparison between the groups does not represent the program's potential effects.

Sometimes, members of the program staff object to the denial of services built into the experimental design. If staff view the experiment as unethical or fear that members of the control group will complain, they may circumvent the procedures of the random assignment process, thus undermining the comparability of the program and control groups. This apparently happened, for example, in an evaluation of the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). In the evaluation, women in health clinics were recruited to participate in a study of the program and were to be randomly assigned based on their Social Security numbers. Proper random assignment should have resulted in an equal number of program and control group members. In one site, however, two-thirds of the women were assigned to the program group, suggesting that the staff who recruited the women falsified some

[12]Erica Baum, "When the Witch Doctors Agree: The Family Support Act and Social Science Research," *Journal of Policy Analysis and Management* vol. 10, Fall 1991, pp. 603-615.

[13]Larry L. Orr, Howard S. Bloom, Stephen H. Bell, Winston Lin, George Cave, Fred Doolittle, *The National JTPA Study: Impacts, Benefits, and Costs of Title II-A* (Bethesda, MD: Abt Associates Inc., Mar. 1994).

[14]David Greenberg and Mark Shroder, *Digest of Social Experiments* (Washington, D.C.: The Urban Institute Press, 1997).

[15]Peter H. Rossi, Howard E. Freeman, and Mark W. Lipsey, *Evaluation: A Systematic Approach 6*, 6th ed. (Newbury, CA: SAGE Publications, Inc., 1998), p. 303. *See, e.g.,* Anne Gordon, Jonathan Jacobson, and Thomas Fraker, *Approaches to Evaluating Welfare Reform: Lessons from Five State Demonstrations* (Princeton, NJ: Mathematica Policy Research, Inc., Oct. 1996).

Social Security numbers to allow women who otherwise would have been assigned to the control group to instead be assigned to the program.[16]

Seeking program participants for a randomized experiment may also result in a skewed sample.[17] For example, in the JTPA evaluation, extensive outreach was necessary because the assignment of applicants to the control group left unfilled slots in the program. The applicants brought into the program were not the same as those who were effectively "displaced" when assigned to the control group. Thus, the impacts on those who were in the program may not correspond to the impacts on those who would have been in the program in the absence of the experiment.

The JTPA evaluation also encountered difficulty enrolling sites. Although the Department of Labor attempted to obtain a nationally representative sample of sites, most refused to participate. And some that expressed a willingness to participate were rejected because of concerns related to their ability to administer a random assignment evaluation without disrupting their normal program operations. As a result, even though the study claims to be a "national" study of the program, the degree to which the findings are generalizable to the nation is questionable, especially if only sites with the best programs volunteered to participate.[18]

This brings up a less-appreciated problem with randomized experiments. Because randomized experiments can be so expensive to mount, they usually involve a relatively small number of sites. This means that the sites probably do not serve a representative sample of the nation as a whole, or even of discrete or identifiable subgroups or special populations. Hence, they tend to lack verifiable generalizability, often called "external validity," which is the applicability of a study's findings to larger populations of interest (which should be specified). Even an extremely well-designed evaluation with high causal validity is of limited usefulness to policymakers if its findings cannot be extrapolated to the program's total clientele, or at least to important, identifiable subgroups. In contrast, nonexperimental evaluations that rely on secondary data sources typically have a high level of generalizability, because they can estimate impacts over a broad range of program environments, participants, and time periods.[19]

---

[16]Michael J. Puma, Janet DiPietro, Jeanne Rosenthal, David Connell, David Judkins, and Mary Kay Fox, *Study of the Impact of WIC on the Growth and Development of Children. Field Test: Feasibility Assessment.* Final Report: Volume I (Cambridge, MA: Abt Associates Inc., 1991).

[17]James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* vol. 9, no. 2, Spring 1995, pp. 85-110.

[18]James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* vol. 9, no. 2, Spring 1995, p. 104.

[19]Robert A. Moffitt and Michele Ver Ploeg (editors), *Evaluating Welfare Reform in an Era of Transition* (Washington, D.C.: National Academy Press, 2001), p. 57.

Hence, the proposed protocol recommends an intensive inquiry into both (1) the quality of the randomization and (2) the evaluation's generalizability.

Nonexperimental evaluations. The other kind of evaluation is called "nonexperimental," and sometimes called a "quasi-experiment."[20] In nonexperimental evaluations, the counterfactual is established by identifying a "comparison"[21] group (for example, persons not participating in the program or from another site, another time, or a data set) whose members are not subject to the program or policy but are nevertheless thought to be similar to those in the program group.

The most common nonexperimental designs compare program participants before and after a program or policy change (pre-post comparison) or program participants to nonparticipants, to individuals from other geographic areas (comparison sites),[22] to individuals from different time periods, and to individuals drawn from secondary data sets. Aggregate data is often used to compare changes in outcomes over time or across geographic areas.

The major disadvantage of nonexperimental evaluations is that they have uncertain causal validity, at best. Put simply, the members of the comparison groups may differ substantially in some unmeasured or undetectable ways from those who have been exposed to the particular program or policy. Typically, nonexperimental designs employ statistical analyses to control for such differences, but how well they do so is open to sharp debate. As Rob Hollister cautions: "Without random assignment there is always the chance that there will be a concentration within the program participant group of those with characteristics that affect the outcome (e.g. the program participants may be more motivated than those who are in the comparison group). To the extent that those characteristics are measured it is possible to control for their effects with statistical models. It is the *unmeasured, or unmeasurable, characteristics (like motivation) which create the bias problem*."[23]

---

[20]Evaluations that do not involve random assignment are often called "quasi-experiments" since they often involve the comparison of the outcomes of program participants to those of a similar "comparison" group of nonparticipants. Unfortunately, using the term "quasi-experiment" tends to obscure the problem of uncertain causal validity inherent in all nonexperimental approaches. Hence, like an increasing number of commentators, we do not use the term "quasi-experiment" but rather use the term "nonexperimental" to cover both micro- and macro-econometric evaluations.

[21]*See* footnote 9, *supra*.

[22]Sometimes the comparisons across geographic areas also follow individuals over time.

[23]Rob Hollister, *The Growth in After-School Programs and Their Impact* (Washington, D.C.: The Brookings Institution, February 2003), p. 8, available from: http://www.brookings.edu/views/papers/sawhill/20030225.pdf, accessed August 21, 2003.

Researchers deal with selection bias through careful regression analysis that statistically controls for a variety of background variables. Examples of common background characteristics include age and sex of the child, maternal age and education, and family status. But even in the best circumstances, such methods cannot capture all of the differences between the two groups, because no data set has information on all the characteristics that may affect the outcomes being examined. All studies are missing some variables of interest.

As a result, there is always some unresolvable uncertainty regarding the causal validity of nonexperimental evaluations. Here is how Gary Burtless of the Brookings Institution put it:

> Our uncertainty about the presence, direction, and potential size of selection bias makes it difficult for social scientists to agree on the reliability of estimates drawn from nonexperimental studies. The estimates may be suggestive, and they may even be helpful when estimates from many competing studies all point in the same direction. But if statisticians obtain widely differing estimates or if the available estimates are the subject of strong methodological criticism, policymakers will be left uncertain about the effectiveness of the program.[24]

Consequently, many literature reviews, meta-analyses, and what works efforts completely exclude nonexperimental evaluations from their assessments.[25] We think this goes too far.

[24]Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives* vol. 9, no. 2 (Spring 1995), pp. 63-84, p. 72.

[25]*See, e.g.*, Matthew Stagner, Jennifer Ehrle, and Jane Reardon-Anderson, "Systematic Review of the Impact of Mandatory Work Policies on Family Structure," (Washington, D.C.,: The Urban Institute, February 24, 2003).

First, there are relatively few experimental evaluations of youth programs,[26] so that relying *solely* on them would provide very little information about promising programs or approaches. In fact, even studies labeled as randomized experiments often, on closer examination, turn out to be something less. For example, a random assignment evaluation of the Memphis Extended-Day Tutoring Program was apparently subverted when children assigned to the program group who did not attend the program (or had low attendance rates) were added to the control group. Rob Hollister observes, "Adding to the control group members of the group initially assigned to the program group but selected out because of non-attendance, or some other reason, seriously undermines the strength of the initial random assignment in avoiding selection bias."[27]

Second, because of ethical issues, randomized experiments ordinarily cannot be used to evaluate full-coverage programs, while nonexperimental ones can. Randomized experiments that involve denying a service to someone who is otherwise entitled to it cannot be conducted. In the late 1980s, for example, the state of Texas implemented a random assignment evaluation to test the impact of twelve-month transitional child care and Medicaid benefits. When the study began, the program group was receiving a benefit (the transitional services) that was otherwise unavailable. Hence, denying the same benefit to the control group did not raise an ethical issue. But a year later, nearly identical transition benefits became mandatory under the Family Support Act. At that point, the control group was to be denied what had become part of the national,

---

[26]For a summary of the major evaluations of youth programs, *see* Rob Hollister, *The Growth in After-School Programs and Their Impact* (Washington, D.C.: The Brookings Institution, February 2003), p. 10. *See also* Peter Benson and Rebecca N. Saito, "The Scientific Foundations of Youth Development," *Youth Development: Issues, Challenges and Directions*, Fall 2000, available from: http://www.ppv.org/pdffiles/ydv/ydv_4.pdf, accessed August 25, 2003; Richard M. Catalano, Lisa Berglund, Jeanne A. M. Ryan, Heather S. Lonczak, and J. David Hawkins, *Positive Youth Development in the United States: Research Findings on Evaluations of Positive Youth Development Programs, Social Development Research Group* (Seattle, WA: Social Development Research Group, University of Washington, November 13, 1998), available from: http://aspe.hhs.gov/hsp/PositiveYouthDev99/, accessed August 25, 2003; Mark Dynarski, Suzanne James-Burdumy, Wendy Mansfield, Daniel Mayer, Mary Moore, John Mullens, and Tim Silva, *A Broader View: The National Evaluation of the 21st Century Learning Centers Program, Design Report: Volume 1* (Washington, D.C.: Mathematica Policy Research, Inc., March 2, 2001, available from: http://www.mathematica-mpr.com/PDFs/broadviewvol1.pdf, accessed August 25, 2003; Jacquelynne S. Eccles and Janice Templeton, "Community-Based Programs for Youth: Lessons Learned from General Developmental Research and From Experimental and Quasi-experimental Evaluations," *Urban Seminar on Children's Health and Safety*, John F. Kennedy School of Government, Harvard University, 2001; Olatokunbo S. Fashola, "Review of Extended-Day and After-School Programs and Their Effectiveness," Report 24, Center for Research on the Education of Students Placed at Risk, Johns Hopkins University, 1998; and Jodie Roth, Jeanne Brooks-Gunn, Lawrence Murray, and William Foster, "Promoting Healthy Adolescents: Synthesis of Youth Development Program Evaluations," *Journal of Research on Adolescence*, vol. 8, no. 4, 1998, pp. 423-459. *See generally* Peter H. Rossi, Howard E. Freeman, and Mark W. Lipsey, *Evaluation: A Systematic Approach 6*, 6th ed. (Newbury, CA: SAGE Publications, Inc., 1998).

[27]Rob Hollister, *The Growth in After-School Programs and Their Impact* (Washington, D.C.: The Brookings Institution, February 2003), p. 9, available from: http://www.brookings.edu/views/papers/sawhill/20030225.pdf, accessed August 21, 2003.

legally guaranteed benefit package. In the face of complaints, the Secretary of Health and Human Services required the control group to receive the benefits, thereby undercutting the experiment.[28]

Third, when used with extreme care, and in the context of other evaluations, especially randomized experiments, they can provide supportive and enriching information about the program or policy being evaluated.

Randomized experiments are also often unable to discern certain types of outcomes. For example, experimental designs may not capture significant "entry effects." Entry effects occur before individuals are ever assigned to an experimental or control group. For instance, a school attendance requirement for welfare mothers may deter some from applying for welfare, whereas significantly expanding child care and other benefits may attract others to it. If random assignment occurs at the point of applying for welfare, these behavioral effects would not be captured in the evaluation.[29]

The impact of community-wide changes in norms or expectations may also be difficult to detect. For example, they may change the culture of the welfare office, leading caseworkers to treat all clients—program and control—differently, which would not be captured by a simple program-control comparison of outcomes.[30]

Nevertheless, the great uncertainty about the causal validity of nonexperimental evaluations requires the utmost caution in there use. Some what works efforts have a special rule limiting their use. For example, the Department of Education's "*Study Design and Implementation Assessment Device" (Study DIAD)* has a question about the comparability of control/comparison groups for "randomized," "quasi-experimental," and "regression discontinuity" designs.[31] The question permits a clear "yes" answer for both randomized and regression-discontinuity designs, but only a "maybe yes" answer for nonexperimental designs.

---

[28]U.S. Department of Health and Human Services, Office of the Secretary, "Press Statement," undated, quoting Secretary Louis Sullivan as stating, "No person receiving welfare in the state of Texas has been denied or will be denied benefits to which they are entitled."

[29]Robert Moffitt, "Evaluation Methods for Program Entry Effects," in Charles Manski and Irwin Garfinkel (editors), *Evaluating Welfare and Training Programs*, (Cambridge, MA: Harvard University Press, 1992), pp. 231-52.

[30]Irwin Garfinkel, Charles F. Manski, and Charles Michalopoulos, "Micro Experiments and Macro Effects," in Charles Manski and Irwin Garfinkel (editors), *Evaluating Welfare and Training Programs*, ed. (Cambridge, MA: Harvard University Press, 1992), pp. 253-76.

[31]U.S. Department of Education, What Works Clearinghouse, "Study Design and Implementation Device (Version 1.0)," July 2003, p. 14, available from: http://www.w-w-c.org/DIAD_Final.doc, accessed August 20, 2003.

As a result, nonexperimental designs can not receive as high a ranking as an randomized experiment, all else being equal.

Therefore, the protocol should explicitly address whether nonexperimental evaluations will be assessed—and under what conditions and with what limitations.

Programs or policies that "do not work." Some literature reviews and what works efforts include only those programs or policies that seem to work, and they exclude evaluations that show no statistically significant effects, small effects, or effects that do not last.

For example, one assessment of the evidence on the impact of youth development programming included only those evaluations that showed "evidence of behavioral outcomes."[32] Rob Hollister laments that "this means that well designed evaluations that found *no statistically significant impact* were not reported. I believe the exclusion of evaluations where there was no statistically significant impact was a mistake, as it is important for us to learn what *doesn't* work as well as what does work."[33] We agree.

First, knowing what to do about programs and policies requires knowing both what works and what does not work. Excluding evaluations that did not work presents an incomplete and misleading picture of what is known about a particular program or policy. Thus, if there were ten studies of a particular approach to youth training and only one study found that the program "worked" (based on a 10 percent level of statistical significance), including only that finding in a review of what works would give the approach too much credibility because just by chance, one in ten studies are likely to show significant effects.[34] In other words, it is important to know whether a particular program or policy that has been tested "worked" in ten studies out of ten, or just one out of ten. Including just those studies with positive effects obscures this point.

Second, even evaluations that find no effect or a negative effect can offer important lessons. Was the program poorly implemented? Did control/comparison group members have

---

[32] *Quoted in* Rob Hollister, *The Growth in After-School Programs and Their Impact* (Washington, D.C.: The Brookings Institution, February 2003), p. 12. Hollister was referring to Richard M. Catalano, Lisa Berglund, Jeanne A. M. Ryan, Heather S. Lonczak, and J. David Hawkins, *Positive Youth Development in the United States: Research Findings on Evaluations of Positive Youth Development Programs, Social Development Research Group* (Seattle, WA: Social Development Research Group, University of Washington, November 13, 1998), available from: http://aspe.hhs.gov/hsp/PositiveYouthDev99/, accessed August 25, 2003.

[33] Rob Hollister, *The Growth in After-School Programs and Their Impact* (Washington, D.C.: The Brookings Institution, February 2003), p. 8, available from: http://www.brookings.edu/views/papers/sawhill/20030225.pdf, accessed August 21, 2003.

[34] This problem is exacerbated by what is commonly referred to as the "file-drawer problem," the publication bias of journals for studies that show significant results can offer a one-sided view of the evidence. The idea is that for every study with significant results that is published, there may be many more with insignificant results languishing in file drawers unpublished.

easy access to similar services? Were there defects in the evaluation that could affect the findings? Were there effects for some subgroups, even if there are no statistically significant effects overall? Answers to these questions can lead to refinements and further testing. Before ruling out a particular approach, it is also important to determine whether similar findings have come up in replications. If so, the research can point to program approaches or policies that should not be replicated. This information can be just as important to policymakers and practitioners, as they examine ways to improve their policies.

What works and what does not, however, is not a simple dichotomy. Researchers often use tests of statistical significance and effect sizes to make such determinations, but there has been some carelessness in this regard. A review of criminal justice evaluation studies by David Weisburd and his colleagues found that many evaluators mistakenly report statistically nonsignificant results as meaning a program or policy had no effect.[35] Such results, however, do not mean the program or policy had no effect, only that the effect was not strong enough to reach statistical significance. Of course, when a program has statistically significant negative effects, one can conclude that it doesn't work.

There are many ways evaluations could be classified, based on whether the findings have positive or negative effects, whether they are statistically significant, and whether the effect sizes indicate meaningful effects. Assuming that the study meets the criteria for a sound evaluation, programs that work could include those with statistically significant findings and effect sizes exceeding some minimum threshold. Programs that do not work could include those with negative findings, statistically insignificant findings, or small effect sizes. Programs that do not satisfy the evaluation criteria could be classified under "results not demonstrated." Regardless of how the findings of any one study are classified, however, it is important that all studies be included in an assessment of what works (and what doesn't).

Meta-analyses. Another kind of evaluation that could be included in an assessment of the effects of a program or policy is a meta-analysis. A meta-analysis is a statistical procedure for combining the results from individual studies, even those with conflicting findings and different evaluation approaches, into a single study with an integrated set of findings.[36] It can lead to stronger findings of effects, because it often combines evaluations with small samples and thereby increases the statistical power of the analysis.

---

[35]David Weisburd, Cynthia M. Lum, and Sue-Ming Yang, "When Can We Conclude That Treatments of Programs 'Don't Work'?," *The Annals of the American Academy of Political and Social Science*, vol. 587, May 2003, pp. 31-48.

[36]Morton Hunt, *How Science Takes Stock: The Story of Meta-Analysis* (New York, NY: Russell Sage Foundation, 1997). *See, e.g.*, Mark W. Lipsey, "Juvenile Delinquency Treatment: A Meta-Analytic Inquiry into the Variability of Effects, in Thomas D. Cook, Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis, and Frederick Mosteller (editors), *Meta-Analysis for Explanation: A Casebook* (New York, NY: Russell Sage Foundation, 1992), pp. 83-127.

Meta-analysis is frequently used in medical research, where interventions and outcome variables tend to be clearly defined, but where clinical trials often involve relatively small samples. As a result, some successful treatments may not appear to be effective, because their evaluation samples were too small to detect anything but the biggest impacts. (This is known as a Type II error, erroneously accepting a finding of no effect.) Even if they have statistically significant effects, they may have such wide confidence intervals, that the possible range of effectiveness is very large and different studies may have very different findings. The added statistical power that comes with a meta-analysis can transform a series of evaluations with no statistically significant effects into an overall finding with a statistically significant effect and, because the confidence intervals become smaller, the overall result tend tends to look more precise. As Mark Lipsey observes in a meta-analysis of juvenile delinquency treatment programs based on many studies with small samples, "The sample sizes used in this literature (median around 60 in each experimental group) do not yield sufficient statistical power for an individual study to find statistical significance for effects sizes in the range of .10-.20 standard deviation units."[37]

A meta-analysis involves several steps. First, the purpose of the analysis and the questions to be addressed are determined. Second, the evaluations that address the purpose are identified. Third, the data from each evaluation are collected and coded. This includes information on the outcomes to be examined, as well as the characteristics of the evaluations and programs themselves. Fourth, the outcomes are transformed into a common metric—an effect size—so that they can be compared across evaluations. (An effect size is the standardized difference between program and control/comparison group mean outcomes.) Finally, the data are statistically combined to determine overall program effects.

Assessing meta-analyses raises special challenges. First, many meta-analyses combine all available evaluations, whether methodologically strong or not, and whether published in a peer-reviewed journal or not. As Richard Berk, a professor in the department of statistics and sociology at the University of California, Los Angeles, and Peter Rossi, S.A. Rice Professor Emeritus at the University of Massachusetts (Amherst), caution: "everything depends on the quality of underlying studies. If they have weak validity overall, even the fanciest of meta-analyses cannot save the day. Meta-analysis cannot correct for fundamental flaws in the original research."[38] Some researchers attempt to control for the quality of the evaluations within the meta-analysis itself by weighting each according to its methodological strength, while others exclude methodologically weaker ones altogether. Either approach, however, requires

[37]Mark W. Lipsey, "Juvenile Delinquency Treatment: A Meta-Analytic Inquiry into the Variability of Effects, in Thomas D. Cook, Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis, and Frederick Mosteller (editors), *Meta-Analysis for Explanation: A Casebook* (New York, NY: Russell Sage Foundation, 1992), p. 126.

[38]Richard A. Berk and Peter H. Rossi, *Thinking About Program Evaluation 2* (Thousand Oaks, Calif.: Sage Publications, 1999), p. 105.

developing clear and objective criteria for dealing with the quality of the evaluations considered. But because this is ultimately a subjective process, the findings remain uncertain.

Second, the programs in a meta-analysis often have different design features, often are aimed at different target groups, and often are carried out in different environments, making it more difficult to discern which aspects of the programs studied are responsible for their effects. This may be particularly true with most social programs, where the implementation, services provided, and other program characteristics can vary tremendously from program to program. Proponents of the meta-analysis approach argue that the analysis can take these differences into account by including them in the statistical model, but this adds another layer of uncertainty and subjectivity to the process.

For these reasons, meta-analyses require the highest level of scrutiny before they are included in a what works effort.

III. The Assessment Process

Recommendation*: The agency protocols should establish a formal process of evaluation that specifies the criteria for assessment and the levels of evidence available. The process should be formalized, with written guidelines and data collection instruments, and it should be open and transparent and subject to outside review. The protocol should explicitly address whether nonexperimental evaluations and meta-analyses will be assessed—and under what conditions or with what limitations.*

Evaluations can go wrong in many ways. Some have such obvious faults that almost anyone can detect them. Other flaws can be detected—and properly assessed—only by experts with long experience and high levels of judgment. Hence, the proposed protocol recommends an intensive inquiry into the quality of the evaluation.

Criteria for assessments. In recent years, various evaluations have sought to determine the effectiveness of particular youth programs. Many of these evaluations provide important information about the impact of such programs, but most also have serious flaws that sharply limit their usefulness. Hence, the proper use of these evaluations requires distinguishing relevant and valid findings from those that are not.

Whether an evaluation uses an experimental or nonexperimental design, a host of questions must be answered before deciding that its findings should be accepted. This inquiry should be based on the generally accepted criteria for judging evaluations.[39] The main areas of inquiry include:

- **Program "theory"**: Does the program or policy make sense in light of existing social science knowledge?

- **Program implementation**: If the program was not implemented as intended, how might the evaluation have been affected?

- **Assessing the randomization**: Was random assignment accomplished successfully? If not, how serious were the problems?

- **Assessing statistical controls in nonexperimental evaluations**: How comparable are the program and comparison groups? Were the possibilities of selection bias and omitted variables considered?

- **Sample size**: Is the sample large enough to yield reasonably precise estimates?

[39]Douglas J. Besharov, Peter Germanis, and Peter H. Rossi, *Evaluating Welfare Reform: A Guide for Scholars and Practitioners* (College Park, MD: University of Maryland, School of Public Affairs, 1997).

- **Attrition**: Was the level of attrition measured and were statistical adjustments used to control for any potential attrition-related biases?

- **Data collection**: Were the necessary data available and reliably collected?

- **Measurement**: Were the key variables valid and could they be measured reliably?

- **Analytical models**: Are the data summarized and analyzed by means of appropriate statistical models?

- **Generalizability**: Are the study's findings applicable to broad populations of programmatic or policy interest ("external validity")? If not, how does this limit the usefulness of the findings?

- **Replication**: Has the evaluation been replicated elsewhere and, if so, are the findings consistent?

- **Evaluator's description of findings**: Are the findings presented accurately? Are they even-handedly presented, describing the limitations of the analyses and considering alternative interpretations?

- **Evaluator's independence**: Are the evaluators involved in the program's development or operations? Do they have a stake, even indirect, in the findings?

- **Statistical significance/confidence intervals**: Were statistical significance tests reported? What level of significance was used?

- **Effect size**: Were effect sizes calculated for all impact estimates and placed in the context of other programs or policies that have similar goals?

- **Sustained effects**: Were program impacts measured after the evaluation was completed? Was the length of the follow-up period sufficient to determine if the effects were sustained?

- **Benefit-cost analysis**: Were the major benefits and costs identified? Were benefits and costs identified for all affected parties, such as program participants, taxpayers, and society as a whole?

- **Cost-effectiveness analysis**: Were the major costs associated with achieving specific outcomes identified?

Appendix 1 contains a detailed listing of the questions that might be asked under each category. The final protocol should be developed by each federal agency with the collaborative process described below.

Specified levels of evidence. Given the wide range in the quality of evaluations, and the limited number in many important areas of youth policy, the assessment process needs to distinguish among levels of evidence and, based on the level of evidence, the appropriate use of the findings. The categories might be, for example, "strongly supported by research," "somewhat supported by the research," "no research on the subject," "somewhat negated by the research," and "strongly negated by the research." There might also be additional categories like "supported by theory and available data" and "negated by theory and available data."

The Department of Education's "What Works Clearinghouse," for example, has a *Cumulative Research Evidence Assessment Device (CREAD)* that distinguishes among the levels of evidence by assessing program evaluations on their "construct validity," "internal validity," "external validity," and "statistical conclusion validity." The Evidence Report Team rates the confidence of each program on each of these criteria as "confident," "somewhat confident," and "somewhat unconfident."[40]

OMB's "PART" also includes a section on "Program Results/Accountability," which asks questions about a program's ability to meet short- and long-term performance measures, be cost effective, and show positive results through independent evaluations. The program's ability to meet these criteria is rated as "yes," "large extent," "small extent," "no," or "NA."[41]

Based on the assessment, there might be a recommendation to apply the findings to specific agency activities (such as funding decisions and dissemination), replicate the research, or redirect the research. OMB's "PART," for example, is intended to influence budget and policy decisions.[42] The Justice Department makes grants to support the implementation and replication of "Blueprints for Violence Prevention" programs.

A formal process. The assessment process needs to be institutionalized, with standard procedures to encourage unbiased treatment. This includes clear rules about collecting and interpreting data concerning the research, ordinarily pursuant to written instruments. (To the

[40]Harris M. Cooper and Jeffrey C. Valentine, *What Works Clearinghouse Cumulative Research Evidence Assessment Device* (version 0.6), Washington, D.C.: U.S. Department of Education, 2003.

[41]Richard P. Emery, Jr. to Program Associate Directors, memorandum, 5 May 2003, "Completing the Program Assessment Rating Tool (PART) for the FY2005 Review Process," Office of Management and Budget, p. 47-50, available from: http://www.whitehouse.gov/omb/part/bpm861.pdf, accessed August 19, 2003.

[42]Office of Management and Budget, *Budget of the United States Government, Fiscal Year 2004* (Washington, D.C.: Government Printing Office, 2003), p. 4, available from: http://www.whitehouse.gov/omb/budget/fy2004/pdf/budget.pdf, accessed August 19, 2003.

extent feasible, the information or documents to be collected should include the evaluation's sampling plan, data collection plan, evaluation plan, all interim and final reports (including appendices), and any other publications by the evaluators themselves or by other commentators writing about the evaluation.) There should also be provision for the systematic sharing of information among agencies.

The "What Works Clearinghouse," for example, follows a standardized process. Each year, it selects general "topic areas," or categories that it would like to evaluate during the coming year (*e.g.*, "Programs for Increasing Adult Literacy" or "Curriculum-Based Interventions for Increasing K-12 Math Achievement").[43] It then accepts nominations of interventions and evaluations to be reviewed that fall under the topic areas. The Evidence Report Team then evaluates each intervention using the *Study Design and Implementation Assessment Device (Study DIAD)* and then the *Cumulative Research Assessment Device (CREAD)*, and writes an Evidence Report synthesizing this information.[44]

Open and transparent. The assessment process needs to be open and transparent to outsiders, with the presentation of detailed data about the research and the reasons for the assessment given.

Thus, the National Institutes of Health's (NIH) Consensus Development Conference (CDC) Program convenes conferences to discuss "controversial issues in medicine important to health care providers, patients, and the general public."[45] These conferences include a two-day session, open to the public, with presentations of research by scientific experts and a period of questions and comments by the attendees. The end result of the conference is a "consensus statement that advances understanding of the technology or issue in question . . . and that will be useful to health professionals and the public."[46] The results of this report are released to the public during a press conference, and are then made available in web and print versions.

Outside review. The process should be subject to review by outside experts, post-publication debate, and revision. The principal investigator of any evaluation assessed should have the opportunity to submit materials or comments.

---

[43]U.S. Department of Education, "What Works Clearinghouse: Evidence Report Topics," available from http://www.w-w-c.org/topicnom.html, accessed June 30, 2003.

[44]U.S. Department of Education, "Introduction to the What Works Clearinghouse Evidence Report Process and the Role of Scientific Standards," March 5, 2003.

[45]National Institutes of Health, "About the Consensus Program: Frequently Asked Questions," available from: http://consensus.nih.gov/about/faq.htm, accessed August 5, 2003.

[46]National Institutes of Health, "Guidelines for the Planning and Management of NIH Consensus Development Conferences," available from http://consensus.nih.gov/about/process.htm, accessed August 5, 2003.

As part of NIH's Consensus Development Conference, the information presented by scientific experts is reviewed and evaluated by a panel of between nine and sixteen members from outside NIH, ranging from other scientists in the field to health professionals.[47] The panel then drafts the consensus statement based on the research presented, presents it to conference attendees for comment, and revises the statement prior to release of the findings. Final panel revisions continue following the conference, and the statement is published by NIH's Office of Medical Applications of Research (OMAR) and often by a medical journal.[48]

[47]National Institutes of Health, "Guidelines for the Planning and Management of NIH Consensus Development Conferences," available from http://consensus.nih.gov/about/process.htm, accessed August 5, 2003.

[48]National Institutes of Health, "Guidelines for the Planning and Management of NIH Consensus Development Conferences," available from http://consensus.nih.gov/about/process.htm, accessed August 5, 2003.

Appendix
Evaluating the Evaluations of Social Programs[49]

A thorough "evaluation of an evaluation" of a social program or policy requires an intensive assessment of the program or policy as well as the evaluation. The following questions can serve as a guide to the process.

Program Issues

Program theory. Underlying every program should be some theory or model of how the program is conceived to work and how it matches the condition it is intended to ameliorate. Understanding the program theory helps to establish the evaluation's analytical framework and data needs. Hence, an evaluation of the program should describe the underlying social problem it is intended to address and how the causal processes described in the model are expected to achieve program goals. Important questions include:

- What is the underlying social problem the program or policy is meant to address? Is it adequately described?
- Is the program or policy being evaluated adequately described?
- Does the program or policy make sense in light of what is known about the social problem being addressed, broader social science theory, and previous evaluations of similar programs or policies?
- What are the program or policy's desired outcomes?
- Are the hypothesized causal processes by which program or policy is intended to achieve its goals clearly stated?
- To what extent are the desired outcomes aligned with what the program or policy might accomplish?
- Have potential side-effects of the program or policy (both positive and negative) been identified?

[49]Source: Douglas J. Besharov and Peter Germanis, "Guide to Evaluating Evaluations of Social Programs," (June 2003). Used with permission.

Program implementation. The key to assessing the success or failure of a program is how well it is implemented. For, no matter how well an evaluation is designed and carried out, if the program is not implemented well, its impact findings may be unreliable and of little use for policy making. Hence, an evaluation should describe the degree to which the program is implemented in accordance with original plans and the nature and extent of any deviations—including a description of the services provided, the "dosage" received (that is, the amount of time of participation), and other relevant information. Important questions include:

- Are the implementation and operation of the program or policy adequately described?
- Was the program or policy implemented as intended?
- Was the program or policy properly applied to the program and control groups? Did the study participants know whether or not the program or policy applied to them?
- Are implementation problems described?
- Are key aspects of program performance, such as the number of people receiving services and the types of services provided, described?
- Was there significant variation in the delivery of program services for either all participants or important subgroups?
- What services were provided to or otherwise received by the control group?
- If the program was poorly implemented, how might the evaluation have been affected?

Causal Validity

Assessing the randomization. The causal validity of an experimental evaluation depends on whether participation in the program and control groups was successfully randomized. This can be judged by reviewing the processes of randomization and the apparent comparability of the resultant groups. Important questions include:

- Is the evaluation as a whole adequately described?
- Is the random assignment adequately described?
- Was random assignment accomplished successfully? If not, how serious were the problems?
- Were statistical tests undertaken to assess the comparability of the baseline characteristics of experimental and control groups? What did they indicate?

- Did the assignment procedures result in any nonrandom additions or removals from the program or control group? If so, what statistical tests were conducted to determine how such procedures may have affected the comparability of the groups? How successful were they?

- Was there post-assignment attrition? Did it disproportionately affect either the program or control group? If so, what statistical adjustments were used to control for any potential bias? How successful were they?

- Was there crossover between program and control groups? If so, what statistical adjustments were used to control for any potential bias? How successful were they?

- Was there contamination between program and control groups? If so, what statistical adjustments were used to control for any potential bias? How successful were they?

- Was the unit of random assignment appropriate for answering the research questions?

- Did the random assignment procedure alter the program's normal enrollment procedures, thereby affecting the characteristics of those receiving program services?

- Could the program have had entry or community effects prior to random assignment?

Assessing statistical controls in nonexperimental evaluations. The causal validity of a nonexperimental evaluation depends on comparability of the program group and the comparison group. This can be judged by reviewing the process by which the comparison group was selected or created, the apparent comparability of the resultant groups, and the statistical models used to control for differences between the groups. Important questions include:

- Is the evaluation as a whole adequately described?

- How was the comparison group selected (or constructed)?

- How comparable are the program and comparison groups? Were statistical tests employed?

- Was the possibility of selection bias and omitted variables considered? How well were they handled? Have data on the characteristics that help identify and control for selection biases been collected?

- Were confounding variables identified?

- Were the statistical models well-specified, that is, were the variables included substantively relevant?

- Were sensitivity analyses conducted to examine the sensitivity of a model's findings to its assumptions?

- Were the statistical models in the proper functional form, that is, was the model appropriate to the statistical properties of the data being analyzed?

Data Issues

Sample size. For a program to be considered successful, it must have statistically significant impacts. The larger the evaluation sample, the more likely it is that effects will be detected. A large sample also makes it possible to conduct subgroup analysis and can offset some of the effects of attrition. Important questions include:

- Is the sample large enough to yield reasonably precise impact estimates, both overall and for important subgroups?

- Were the implications of sample size discussed, particularly in relation to the size of confidence intervals, which show the margin of error around an estimate?

Attrition. All evaluation projects are vulnerable to attrition, that is, members of the research sample leaving the study. People leave studies because they lose interest, move away, or simply believe they no longer need the service. If a large proportion of the initial sample leaves the evaluation (or never starts it), if the attrition is greater in either the program or control/comparison group, or if the exits are concentrated in particular subgroups, the evaluation's finding can be compromised. Important questions include:

- What is the extent of attrition?

- Did attrition vary over time, by outcome, by data source, or by program or control group?

- Did attrition affect the statistical power of the evaluation?

- How did attrition affect the representativeness of the sample?

- How did attrition affect the comparability of experimental and control groups?

- Were statistical adjustments made to deal with attrition? How successful were they?

Data collection. No matter how well designed, an evaluation is ultimately dependent on the completeness and accuracy of its data. This includes data on both study participants and the

program or policy being evaluated. The data for any evaluation typically come from multiple sources, including administrative records and special surveys, and cover periods beginning before program implementation and extending into periods after program completion. Too much missing or inaccurate data can undermine the credibility of the evaluation's findings. Important questions include:

- Are the data collection sources and procedures described?
- Are the data sources appropriate for the questions being studied?
- Are the data complete? What steps were taken to minimize missing data? For example, for survey-based findings, what procedures were used to obtain high response rates?
- Are the data accurate? How was accuracy verified?
- What statistical or other controls were used to correct for potential bias resulting from missing or erroneous data? How successful were they?
- What are the implications of missing or erroneous data for the findings?

Measurement issues. The variables used in an evaluation should be based on valid and reliable measures. A measure is "valid" if it measures what it is supposed to measure. A measure is "reliable" if repeated measures produce the same result. Important questions include:

- Were all appropriate and relevant variables measured?
- Were all key measures properly defined?
- Were all key variables measured with instruments of accepted validity and reliability?
- Were there any changes in data collection methods that could produce changes in outcomes?
- Were outcomes measured at a time appropriate for capturing the program or policy's true impact?
- Have potential sources of measurement error been identified? Were the measurements affected by response and recall biases? Did subjects misinterpret questions or otherwise provide erroneous responses?
- Were there Hawthorne effects; that is, did the act of measurement affect the outcome?

Interpretation

Analytical models. The data from an evaluation often can be analyzed in several ways, each of which may lead to somewhat different interpretations. A strong analytical model is needed so that the data can be analyzed and interpreted in a way that is consistent with the program's purpose and the evaluation's design. This is particularly important for nonexperimental evaluations, where the statistical models should include adequate specification (the variables included are substantively relevant) and proper functional form (the model is appropriate to the statistical properties of the data being analyzed). Important questions include:

- Were appropriate statistical models used?
- Were the models used tested for specification errors?
- Was the analysis performed at the same level as the randomization?
- Was the unit of random assignment appropriate for answering the research questions?

Generalizability. Often called "external validity," generalizability is the applicability of a study's findings to larger populations of interest (which should be specified). Even an extremely well-designed evaluation with high causal validity is not useful to policymakers if its findings cannot be extrapolated to the program's total clientele, or at least to important, identifiable subgroups. Important questions include:

- How were the evaluation sites and the evaluation sample selected? How representative is the sample of the broader population of interest?
- Did the program or policy include special features or operate under unique circumstances that are not likely to be duplicated when operated on a larger scale?
- Are the findings generalizable to larger populations, different settings, and different social and economic conditions? How was this assessment made?
- If the findings are not generalizable, how does this limit their usefulness?

Replication. Perhaps the greatest shortcoming of many of the "successful" programs is that they have not been replicated. We do not know how robust the findings would be if the programs were applied to different populations or run by less committed and supervised staff. We do not know which programmatic elements bring about desired outcomes, nor do we know how much a program or policy is needed. Hence, even a study that seems to have been conducted in accordance with the highest professional standards should be replicated and evaluated to see whether the findings hold up when the project is operated by different individuals in different locations. Important questions include:

- Has the evaluation been replicated in another place?
- Have the findings been consistent across studies?
- Was another principal investigator involved?

Evaluator's description of findings. No matter how well analyzed numerically, numbers do not speak for themselves nor do they speak directly to policy issues. The findings of the statistical analysis must be interpreted in an even-handed manner with justifiable statements about the meaning of the findings. The evaluation report should disclose the limitations of the data analyses and present alternative interpretations. Important questions include:

- When alternative analysis strategies are possible, did the evaluation show how sensitive findings are to the use of such alternatives?
- Are alternative interpretations of the data discussed?
- Are important caveats regarding the findings stated?
- Are the findings placed in the proper policy or programmatic context?

Evaluator's independence. Many evaluations are conducted by those who developed the program, rather than by independent evaluators. This can create perceptions that the findings are somehow biased. Important questions include:

- Are the evaluators involved in the program's development or operations? Do they have a stake, even indirect, in the findings?
- Are sufficient data about the program and the evaluation available to outsiders? Are they easily analyzed by outsiders?
- Has the research undergone peer review?

Policy Significance

Statistical significance/confidence intervals. Tests of statistical significance can determine whether the changes captured by an evaluation are the result of chance or are caused by the program or policy being evaluated. Although the standards for statistical significance often vary, most evaluations consider findings statistically significant if the probability of their happening by chance is less than 5 percent. A statistically significant finding, however, does not mean that the size of the impact is known with precision, but it does allow the impact to be placed within a confidence interval that gives the lower and upper bound of the estimate. (Findings that are not statistically significant do not mean that the program or policy had zero

effect, only that, given the size of the sample, the effect was not strong enough to rule out a no effect conclusion.) Important questions include:

- Were statistical significance tests reported? What level of significance was used?
- Were confidence intervals provided for the main outcomes of interest?

Effect size. Statistical significance indicates whether the estimated effects of a program or policy are real, but it does not provide an indication of whether the findings are large or small and, thus, of policy significance. One way to assess the size of a program's impacts is to standardize the estimates by converting them into effect sizes.[50] Although the definition of what constitutes a large or small effect size is subjective, until very recently, most researchers used as a rule of thumb criteria proposed by Jacob Cohen: "small" (.20), "medium" (.50), and "large" (.80).[51] Some evaluators are now considering effect sizes as low as .10 as important. In some cases, even small effect sizes can be important, for example, if a program or policy reduces mortality. Hence, they should be assessed relative to other programs or policies that have similar goals. Important questions include:

- Were effect sizes calculated for all impact estimates?
- Were there some outcomes for which effect sizes could not be calculated or for which such estimates would not be appropriate?
- Were effect sizes discussed in the context of the program or policy being studied as well as other programs or policies that have similar goals?
- Were research design and methodological issues that could affect the magnitude of effect sizes discussed?

Sustained effects. Many programs show impacts while individuals are actively participating in a program or are affected by a policy. Past research in many fields, however, indicates that for many programs, these initial impacts "fade out" after participation in the program has been completed. (In some cases, they only appear after program completion.) It is, therefore, important to measure program effects beyond the program phase to determine whether the effects are sustained or not. Important questions include:

[50] An effect size is calculated by taking the difference in mean scores between the program group and the control (or comparison) group and dividing by the standard deviation of the control (or comparison) group. The absolute value of most effect sizes ranges from 0 to 1.

[51] *See* Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, NJ: Lawrence Erlbaum, 1988), p. 535. Cohen cautions that small effect sizes should not be dismissed because the meaning of any given effect size is "in the final analysis, a function of the context in which it is embedded."

- Were program impacts measured after the evaluation was completed? Was the length of the follow-up period sufficient to determine if the effects were sustained?

- Were program impacts measured over multiple periods to establish a trend?

Benefit-cost analysis. Even if a program or policy has statistically significant findings and meaningful effect sizes, it is important to determine whether the benefits of a program or policy outweigh its costs. If they do, the investment may be considered worthwhile. If not, the continuation of the program or policy (at least in its present form) should be reassessed, especially given other possible uses of the funds. (In some cases, a program or policy may be justifiable even if the costs exceed the benefits, perhaps because it produces something of value to society that is difficult to monetize.) A comprehensive benefit-cost analysis can demonstrate whether a program's effects justify the cost. Hence, an evaluation should describe the various benefits and costs associated with a program or policy. It should also identify the major groups affected, such as program participants, taxpayers, and society as a whole, and describe how the findings vary across them. Important questions include:

- Was a comprehensive benefit-cost analysis conducted?

- Were the major benefits and costs been identified?

- Were benefits and costs identified for all affected parties, such as program participants, taxpayers, and society as a whole? Was a benefit-cost analysis conducted for each of the affected groups?

- Were all benefits and costs expressed in monetary terms? What assumptions were made to monetize the benefits and costs of effects typically expressed in nonmonetary terms, such as reductions in pain and suffering stemming from less crime?

- Were any nonmonetary benefits not considered?

- Were dollar values adjusted to reflect the time value of money?

- Were future benefits and costs projected? Were the assumptions for any projections described and based on sound theory/data?

- Were confidence intervals, which show the margin of error around an estimate, presented?

Cost-effectiveness analysis. Sometimes it is not possible to convert all of the benefits of a program into dollar terms—for example, the value of a life saved due to the reduction in criminal activity achieved by a youth employment program. As a result, a comprehensive benefit-cost analysis may not be possible. Programs can still be assessed, however, in terms of

the cost of achieving various outcomes. Such a cost-effectiveness analysis measures the efficacy of a program in achieving an outcome in relation to its costs. For example, the cost-effectiveness of a youth employment program could be measured in terms of the dollar cost of placing a young person in a job. In this way, programs with similar goals can be compared based on the costs of achieving specific outcomes. Hence, an evaluation should describe the various costs and key outcomes associated with a program or policy. Important questions include:

- Was a cost-effectiveness analysis conducted?
- Were the major costs and impacts identified?
- Were dollar values adjusted to reflect the time value of money?
- Were confidence intervals, which show the margin of error around an estimate, presented?

Assessing meta-analyses. The causal validity of a meta-analysis depends on the quality of the evaluation included in the review and the statistical techniques used to derive the findings. This can be judged by reviewing the process by which the evaluations were selected and the statistical procedures used. Important questions include:

- Was there a description of the criteria used to identify relevant evaluations, including a description of program goals, target groups, outcome variables, and research designs?
- How were the evaluations for the meta-analysis selected? Were all evaluations (including unpublished ones) on a particular topic included? Were any excluded on methodological or other grounds? Were the evaluations limited to well-designed and well-executed experimental evaluations?
- Were the features of programs that were included sufficiently similar so that the findings could appropriately be combined?
- Was there a description of the statistical procedures used to compute the findings?
- Did the analysis include sensitivity tests, such as the effect of excluding unpublished studies or studies with weak evaluations?