United States Government Accountability Office

GAO

Report to Congressional Requesters

November 2009

PROGRAM EVALUATION

A Variety of Rigorous Methods Can Help Identify Effective Interventions

GAO-10-30

November 2009

PROGRAM EVALUATION

A Variety of Rigorous Methods Can Help Identify Effective Interventions

GAO Accountability • Integrity • Reliability

Highlights

Highlights of GAO-10-30, a report to congressional requesters

Why GAO Did This Study

Recent congressional initiatives seek to focus funds for certain federal social programs on interventions for which randomized experiments show sizable, sustained benefits to participants or society. The private, nonprofit Coalition for Evidence-Based Policy undertook the Top Tier Evidence initiative to help federal programs identify interventions that meet this standard.

GAO was asked to examine (1) the validity and transparency of the Coalition's process, (2) how its process compared to that of six federally supported efforts to identify effective interventions, (3) the types of interventions best suited for assessment with randomized experiments, and (4) alternative rigorous methods used to assess effectiveness. GAO reviewed documents, observed the Coalition's advisory panel deliberate on interventions meeting its top tier standard, and reviewed other documents describing the processes the federally supported efforts had used. GAO reviewed the literature on evaluation methods and consulted experts on the use of randomized experiments.

The Coalition generally agreed with the findings. The Departments of Education and Health and Human Services provided technical comments on a draft of this report. The Department of Justice provided no comments.

What GAO Recommends

GAO makes no recommendations.

View GAO-10-30 or key components.
For more information, contact Nancy Kingsbury at (202) 512-2700 or kingsburyn@gao.gov.

What GAO Found

The Coalition's Top Tier Evidence initiative criteria for assessing evaluation quality conform to general social science research standards, but other features of its overall process differ from common practice for drawing conclusions about intervention effectiveness. The Top Tier initiative clearly describes how it identifies candidate interventions but is not as transparent about how it determines whether an intervention meets the top tier criteria. In the absence of detailed guidance, the panel defined sizable and sustained effects through case discussion. Over time, it increasingly obtained agreement on whether an intervention met the top tier criteria.

The major difference in rating study quality between the Top Tier and the six other initiatives examined is a product of the Top Tier standard as set out in certain legislative provisions: the other efforts accept well-designed, well-conducted, nonrandomized studies as credible evidence. The Top Tier initiative's choice of broad topics (such as early childhood interventions), emphasis on long-term effects, and use of narrow evidence criteria combine to provide limited information on what is effective in achieving specific outcomes. The panel recommended only 6 of 63 interventions reviewed as providing "sizeable, sustained effects on important outcomes." The other initiatives acknowledge a continuum of evidence credibility by reporting an intervention's effectiveness on a scale of high to low confidence.

The program evaluation literature generally agrees that well-conducted randomized experiments are best suited for assessing effectiveness when multiple causal influences create uncertainty about what caused results. However, they are often difficult, and sometimes impossible, to carry out. An evaluation must be able to control exposure to the intervention and ensure that treatment and control groups' experiences remain separate and distinct throughout the study.

Several rigorous alternatives to randomized experiments are considered appropriate for other situations: quasi-experimental comparison group studies, statistical analyses of observational data, and—in some circumstances—in-depth case studies. The credibility of their estimates of program effects relies on how well the studies' designs rule out competing causal explanations. Collecting additional data and targeting comparisons can help rule out other explanations.

GAO concludes that

- requiring evidence from randomized studies as sole proof of effectiveness will likely exclude many potentially effective and worthwhile practices;
- reliable assessments of evaluation results require research expertise but can be improved with detailed protocols and training;
- deciding to adopt an intervention involves other considerations in addition to effectiveness, such as cost and suitability to the local community; and
- improved evaluation quality would also help identify effective interventions.

Contents

Abbreviations

AHRQ	Agency for Healthcare Research and Quality
CDC	Centers for Disease Control and Prevention
EPC	Evidence-based Practice Centers
GPRA	Government Performance and Results Act of 1993
HHS	Department of Health and Human Services
MPG	Model Programs Guide
NREPP	National Registry of Evidence-based Programs and Practices
OMB	Office of Management and Budget
PART	Program Assessment Rating Tool
PRS	HIV/AIDS Prevention Research Synthesis
SAMHSA	Substance Abuse and Mental Health Administration
SCHIP	State Children's Health Insurance Program
WWC	What Works Clearinghouse

GAO
Accountability * Integrity * Reliability

United States Government Accountability Office
Washington, DC 20548

November 23, 2009

The Honorable Joseph I. Lieberman
Chairman
The Honorable Susan M. Collins
Ranking Member
Committee on Homeland Security and Governmental Affairs
United States Senate

The Honorable Mary L. Landrieu
Chairman
Subcommittee on Disaster Recovery
Committee on Homeland Security and Governmental Affairs
United States Senate

Several recent congressional initiatives seek to focus funds in certain federal social programs on activities for which the evidence of effectiveness is rigorous—specifically, well-designed randomized controlled trials showing sizable, sustained benefits to program participants or society. To help agencies, grantees, and others implement the relevant legislative provisions effectively, the private, nonprofit Coalition for Evidence-Based Policy launched the Top Tier Evidence initiative in 2008 to identify and validate social interventions meeting the standard of evidence set out in these provisions. In requesting this report, you expressed interest in knowing whether limiting the search for effective interventions to those that had been tested against these particular criteria might exclude from consideration other important interventions. To learn whether the Coalition's approach could be valuable in helping federal agencies implement such funding requirements, you asked GAO to independently assess the Coalition's approach. GAO's review focused on the following questions.

1. How valid and transparent is the process the Coalition used—searching, selecting, reviewing, and synthesizing procedures and criteria—to identify social interventions that meet the standard of "well-designed randomized controlled trials showing sizable, sustained effects on important outcomes"?

2. How do the Coalition's choices of procedures and criteria compare to (a) generally accepted design and analysis techniques for identifying effective interventions and (b) similar standards and processes other federal agencies use to evaluate similar efforts?

3. What types of interventions do randomized controlled experiments appear to be best suited to assessing effectiveness?

4. For intervention types for which randomized controlled experiments appear not to be well suited, what alternative forms of evaluation are used to successfully assess effectiveness?

To assess the Coalition's Top Tier initiative, we reviewed documents, conducted interviews, and observed the deliberations of its advisory panel, who determined which interventions met the "top tier" evidence standard—well-designed, randomized controlled trials showing sizable, sustained benefits to program participants or society. We evaluated the transparency of the initiative's process against its own publicly stated procedures and criteria, including the top tier evidence standard. To assess the validity of the Coalition's approach, we compared its procedures and criteria to those recommended in program evaluation textbooks and related publications, as well as to the processes actually used by six federally supported initiatives with a similar purpose to the Coalition. Through interviews and database searches, we identified six initiatives supported by the U.S. Department of Education, Department of Health and Human Services (HHS), and Department of Justice that also conduct systematic reviews of evaluation evidence to identify effective interventions.[1] We ascertained the procedures and criteria these federally supported efforts used from interviews and document reviews.

We identified the types of interventions for which randomized controlled experiments—the Coalition's primary evidence criterion—are best suited and alternative methods for assessing effectiveness by reviewing the program evaluation methodology literature and by having our summaries of that literature reviewed by a diverse set of experts in the field. We obtained reviews from seven experts who had published on evaluation methodology, held leadership positions in the field, and had experience in diverse subject areas and methodologies.

We conducted this performance audit from May 2008 through November 2009 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to

[1]In addition, the federal Interagency Working Group on Youth Programs Web site www.findyouthinfo.gov provides interactive tools and other resources to help youth-serving organizations assess community assets, identify local and federal resources, and search for evidence-based youth programs.

obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Background

Over the past two decades, several efforts have been launched to improve federal government accountability and results, such as the strategic plans and annual performance reports required under the Government Performance and Results Act of 1993 (GPRA). The act was designed to provide executive and congressional decision makers with objective information on the relative effectiveness and efficiency of federal programs and spending. In 2002, the Office of Management and Budget (OMB) introduced the Program Assessment Rating Tool (PART) as a key element of the budget and performance integration initiative under President George W. Bush's governmentwide Management Agenda. PART is a standard set of questions meant to serve as a diagnostic tool, drawing on available program performance and evaluation information to form conclusions about program benefits and recommend adjustments that may improve results.

The success of these efforts has been constrained by lack of access to credible evidence on program results. We previously reported that the PART review process has stimulated agencies to increase their evaluation capacity and available information on program results.[2] After 4 years of PART reviews, however, OMB rated 17 percent of 1,015 programs "results not demonstrated"—that is, did not have acceptable performance goals or performance data. Many federal programs, while tending to have limited evaluation resources, require program evaluation studies, rather than performance measures, in order to distinguish a program's effects from those of other influences on outcomes.

Program evaluations are systematic studies that assess how well a program is working, and they are individually tailored to address the client's research question. Process (or implementation) evaluations assess the extent to which a program is operating as intended. Outcome evaluations assess the extent to which a program is achieving its outcome-

[2]GAO, *Program Evaluation: OMB's PART Reviews Increased Agencies' Attention to Improving Evidence of Program Results*, GAO-06-67 (Washington, D.C.: October 28, 2005), p. 28.

oriented objectives but may also examine program processes to understand how outcomes are produced. When external factors such as economic or environmental conditions are known to influence a program's outcomes, an impact evaluation may be used in an attempt to measure a program's net effect by comparing outcomes with an estimate of what would have occurred in the absence of the program intervention. A number of methodologies are available to estimate program impact, including experimental and nonexperimental designs.

Concern about the quality of social program evaluation has led to calls for greater use of randomized experiments—a method used more widely in evaluations of medical than social science interventions. Randomized controlled trials (or randomized experiments) compare the outcomes for groups that were randomly assigned either to the treatment or to a nonparticipating control group before the intervention, in an effort to control for any systematic difference between the groups that could account for a difference in their outcomes. A difference in these groups' outcomes is believed to represent the program's impact. While random assignment is considered a highly rigorous approach in assessing program effectiveness, it is not the only rigorous research design available and is not always feasible.

The Coalition for Evidence-Based Policy is a private, nonprofit organization that was sponsored by the Council for Excellence in Government from 2001 until the Council closed in 2009. The Coalition aims to improve the effectiveness of social programs by encouraging federal agencies to fund rigorous studies—particularly randomized controlled trials—to identify effective interventions and to provide strong incentives and assistance for federal funding recipients to adopt such interventions.[3] Coalition staff have advised OMB and federal agencies on how to identify rigorous evaluations of program effectiveness, and they manage a Web site called "Social Programs That Work" that provides examples of evidence-based programs to "provide policymakers and practitioners with clear, actionable information on what works, as demonstrated in scientifically-valid studies. . . ."[4]

[3]See Coalition for Evidence-Based Policy, www.coalition4evidence.org.

[4]See Coalition for Evidence-Based Policy, Social Programs That Work, www.evidencebasedprograms.org.

In 2008, the Coalition launched a similar but more formal effort, the Top Tier Evidence initiative, to identify only interventions that have been shown in "well-designed and implemented randomized controlled trials, preferably conducted in typical community settings, to produce sizeable, sustained benefits to participants and/or society." [5] At the same time, it introduced an advisory panel of evaluation researchers and former government officials to make the final determination. The Coalition has promoted the adoption of this criterion in legislation to direct federal funds toward strategies supported by rigorous evidence. By identifying interventions meeting this criterion, the Top Tier Evidence initiative aims to assist agencies, grantees, and others in implementing such provisions effectively.

Federally Supported Initiatives to Identify Effective Interventions

Because of the flexibility provided to recipients of many federal grants, achieving these federal programs' goals relies heavily on agencies' ability to influence their state and local program partners' choice of activities. In the past decade, several public and private efforts have been patterned after the evidence-based practice model in medicine to summarize available effectiveness research on social interventions to help managers and policymakers identify and adopt effective practices. The Department of Education, HHS, and Department of Justice support six initiatives similar to the Coalition's to identify effective social interventions. These initiatives conduct systematic searches for and review the quality of evaluations of intervention effectiveness in a given field and have been operating for several years.

We examined the processes used by these six ongoing federally supported efforts to identify effective interventions in order to provide insight into the choices of procedures and criteria that other independent organizations made in attempting to achieve a similar outcome as the Top Tier initiative: to identify interventions with rigorous evidence of effectiveness. The Top Tier initiative, however, aims to identify not all effective interventions but only those supported by the most definitive evidence of effectiveness. The processes each of these initiatives (including Top Tier) takes to identify effective interventions are summarized in appendix I.

[5]See Coalition for Evidence-Based Policy, Top Tier Evidence, http://toptierevidence.org. The criterion is also sometimes phrased more simply as interventions that have been shown in well-designed randomized controlled trials to produce sizable, sustained effects on important outcomes.

Evidence-Based Practice Centers

In 1997, the Agency for Healthcare Research and Quality (AHRQ) established the Evidence-based Practice Centers (EPC) (there are currently 14) to provide evidence on the relative benefits and risks of a wide variety of health care interventions to inform health care decisions.[6] EPCs perform comprehensive reviews and synthesize scientific evidence to compare health treatments, including pharmaceuticals, devices, and other types of interventions. The reviews, with a priority on topics that impose high costs on the Medicare, Medicaid, or State Children's Health Insurance (SCHIP) programs, provide evidence about effectiveness and harms and point out gaps in research. The reviews are intended to help clinicians and patients choose the best tests and treatments and to help policy makers make informed decisions about health care services and quality improvement.[7]

The Guide to Community Preventive Services

HHS established the Guide to Community Preventive Services (the Community Guide) in 1996 to provide evidence-based recommendations and findings about public health interventions and policies to improve health and promote safety. With the support of the Centers for Disease Control and Prevention (CDC), the Community Guide synthesizes the scientific literature to identify the effectiveness, economic efficiency, and feasibility of program and policy interventions to promote community health and prevent disease. The Task Force on Community Preventive Services, an independent, nonfederal, volunteer body of public health and prevention experts, guides the selection of review topics and uses the evidence gathered to develop recommendations to change risk behaviors, address environmental and ecosystem challenges, and reduce disease, injury, and impairment. Intended users include public health professionals, legislators and policy makers, community-based organizations, health care service providers, researchers, employers, and others who purchase health care services.[8]

HIV/AIDS Prevention Research Synthesis

CDC established the HIV/AIDS Prevention Research Synthesis (PRS) in 1996 to review and summarize HIV behavioral prevention research literature. PRS conducts systematic reviews to identify evidence-based HIV behavioral interventions with proven efficacy in preventing the

[6]AHRQ was formerly called the Agency for Health Care Policy and Research.

[7]See Agency for Healthcare Research and Quality, Effective Health Care, www.effectivehealthcare.ahrq.gov.

[8]See Guide to Community Preventive Services, www.thecommunityguide.org/index.html.

acquisition or transmission of HIV infection (reducing HIV-related risk behaviors, sexually transmitted diseases, HIV incidence, or promoting protective behaviors). These reviews are intended to translate scientific research into practice by providing a compendium of evidence-based interventions to HIV prevention planners and providers and state and local health departments for help with selecting interventions best suited to the needs of the community.[9]

Model Programs Guide

The Office of Juvenile Justice and Delinquency Prevention established the Model Programs Guide (MPG) in 2000 to identify effective programs to prevent and reduce juvenile delinquency and related risk factors such as substance abuse. MPG conducts reviews to identify effective intervention and prevention programs on the following topics: delinquency; violence; youth gang involvement; alcohol, tobacco, and drug use; academic difficulties; family functioning; trauma exposure or sexual activity and exploitation; and accompanying mental health issues. MPG produces a database of intervention and prevention programs intended for juvenile justice practitioners, program administrators, and researchers.[10]

National Registry of Evidence-Based Programs and Practices

The Substance Abuse and Mental Health Services Administration (SAMHSA) established the National Registry of Evidence-based Programs and Practices (NREPP) in 1997 and provides the public with information about the scientific basis and practicality of interventions that prevent or treat mental health and substance abuse disorders.[11] NREPP reviews interventions to identify those that promote mental health and prevent or treat mental illness, substance use, or co-occurring disorders among individuals, communities, or populations. NREPP produces a database of interventions that can help practitioners and community-based organizations identify and select interventions that may address their particular needs and match their specific capacities and resources.[12]

[9]See Centers for Disease Control and Prevention, HIV/AIDS Prevention Research Synthesis Project, www.cdc.gov/hiv/topics/research/prs.

[10]See Office of Juvenile Justice and Delinquency Prevention Programs, OJJDP Model Programs Guide, www2.dsgonline.com/mpg.

[11]It was established as the National Registry of Effective Prevention Programs; it was expanded in 2004 to include mental health and renamed the National Registry of Evidence-based Programs and Practices.

[12]See NREPP, SAMHSA's National Registry of Evidence-based Programs and Practices, www.nrepp.samhsa.gov.

What Works Clearinghouse

The Institute of Education Sciences established the What Works Clearinghouse (WWC) in 2002 to provide educators, policymakers, researchers, and the public with a central source of scientific evidence on what improves student outcomes. WWC reviews research on the effectiveness of replicable educational interventions (programs, products, practices, and policies) to improve student achievement in areas such as mathematics, reading, early childhood education, English language, and dropout prevention. The WWC Web site reports information on the effectiveness of interventions through a searchable database and summary reports on the scientific evidence.[13]

Top Tier Initiative's Process Is Mostly Transparent

The Coalition provides a clear public description on its Web site of the first two phases of its process—search and selection to identify candidate interventions. It primarily searches other evidence-based practice Web sites and solicits nominations from experts and the public. Staff post their selection criteria and a list of the interventions and studies reviewed on their Web site. However, their public materials have not been as transparent about the criteria and process used in the second two phases of its process—review and synthesize study results to determine whether an intervention met the Top Tier criteria. Although the Coalition provides brief examples of the panel's reasoning in making Top Tier selections, it has not fully reported the panel's discussion of how to define sizable and sustained effects in the absence of detailed guidance or the variation in members' overall assessments of the interventions.

The Top Tier Initiative Clearly Described Its Process for Identifying Interventions

Through its Web site and e-mailed announcements, the Coalition has clearly described how it identified interventions by searching the strongest evidence category of 15 federal, state, and private Web sites profiling evidence-based practices and by soliciting nominations from federal agencies, researchers, and the general public. Its Web site posting clearly indicated the initiative's search and selection criteria: (1) early childhood interventions (for ages 0–6) in the first phase of the initiative and interventions for children and youths (ages 7–18) in the second phase (starting in February 2009) and (2) interventions showing positive results in well-designed and implemented randomized experiments. Coalition staff then searched electronic databases and consulted with researchers to identify any additional randomized studies of the interventions selected

[13]See IES What Works Clearinghouse, http://ies.ed.gov/ncee/wwc.

for review. The July 2008 announcement of the initiative included its August 2007 "Checklist for Reviewing a Randomized Controlled Trial of a Social Program or Project, to Assess Whether It Produced Valid Evidence." The Checklist describes the defining features of a well-designed and implemented randomized experiment: equivalence of treatment and control groups throughout the study, valid measurement and analysis, and full reporting of outcomes. It also defines a strong body of evidence as consisting of two or more randomized experiments or one large multisite study.

In the initial phase (July 2008 through February 2009), Coalition staff screened studies of 46 early childhood interventions for design or implementation flaws and provided the advisory panel with brief summaries of the interventions and their results and reasons why they screened out candidates they believed clearly did not meet the Top Tier standard. Reasons for exclusion included small sample sizes, high sample attrition (both during and after the intervention), follow-up periods of less than 1 year, questionable outcome measures (for example, teachers' reports of their students' behavior), and positive effects that faded in later follow-up. Staff also excluded interventions that lacked confirmation of effects in a well-implemented randomized study. Coalition staff recommended three candidate interventions from their screening review; advisory panel members added two more for consideration after reviewing the staff summaries (neither of which was accepted as top tier by the full panel). While the Top Tier Initiative explains each of its screening decisions to program developers privately, on its Web site it simply posts a list of the interventions and studies reviewed, along with full descriptions of interventions accepted as top tier and a brief discussion of a few examples of the panel's reasoning.[14]

Reviewers Defined the Top Tier Criteria through Case Discussion

The Top Tier initiative's public materials are less transparent about the process and criteria used to determine whether an intervention met the Top Tier standard than about candidate selection. One panel member, the lead reviewer, explicitly rates the quality of the evidence on each candidate intervention using the Checklist and rating form. Coalition staff members also use the Checklist to review the available evidence and prepare detailed study reviews that identify any significant limitations. The full advisory panel then discusses the available evidence on the

[14]See http://toptierevidence.org.

recommended candidates and holds a secret ballot on whether an intervention meets the Top Tier standard, drawing on the published research articles, the staff review, and the lead reviewer's quality rating and Top Tier recommendation.

The advisory panel discussions did not generally dispute the lead reviewer's study quality ratings (on quality of overall design, group equivalence, outcome measures, and analysis reporting) but, instead, focused on whether the body of evidence met the Top Tier standard (for sizable, sustained effects on important outcomes in typical community settings). The Checklist also includes two criteria or issues that were not explicit in the initial statement of the Top Tier standard—whether the body of evidence showed evidence of effects in more than one site (replication) and provided no strong countervailing evidence. Because neither the Checklist nor the rating form provides definitions of how large a sizable effect should be, how long a sustained effect should last, or what constituted an important outcome, the panel had to rely on its professional judgment in making these assessments.

Although a sizable effect was usually defined as one passing tests of statistical significance at the 0.05 level, panel members raised questions about whether particular effects were sufficiently large to have practical importance. The panel often turned to members with subject matter expertise for advice on these matters. One member cautioned against relying too heavily on the reported results of statistical tests, because some studies, by conducting a very large number of comparisons, appeared to violate the assumptions of those tests and, thus, probably identified some differences between experimental groups as statistically significant simply by chance.

The Checklist originally indicated a preference for data on long-term outcomes obtained a year after the intervention ended, preferably longer, noting that "longer-term effects . . . are of greatest policy and practical importance."[15] Panel members disagreed over whether effects measured no later than the end of the second grade—at the end of the intervention—were sufficiently sustained and important to qualify as top tier, especially in the context of other studies that tracked outcomes to age 15 or older.

[15]Coalition for Evidence-Based Policy, "Checklist for Reviewing a Randomized Controlled Trial of a Social Program or Project, to Assess Whether It Produced Valid Evidence," August 2007, p. 5. http://toptierevidence.org

One panel member questioned whether it was realistic to expect the effects of early childhood programs to persist through high school, especially for low-cost interventions; others noted that the study design did not meet the standard because it did not collect data a year after the intervention ended. In the end, a majority (but not all) of the panel accepted this intervention as top tier because the study found that effects persisted over all 3 program years, and they agreed to revise the language in the Checklist accordingly.

Panel members disagreed on what constituted an important outcome. Two noted a pattern of effects in one study on cognitive and academic tests across ages 3, 5, 8, and 18. Another member did not consider cognitive tests an important enough outcome and pointed out that the effects diminished over time and did not lead to effects on other school-related behavioral outcomes such as special education placement or school drop-out. Another member thought it was unreasonable to expect programs for very young children (ages 1–3) to show an effect on a child at age 18, given all their other experiences in the intervening years.

A concern related to judging importance was whether and how to incorporate the cost of the intervention into the intervention assessment. On one hand, there was no mention of cost in the Checklist or intervention rating form. On the other hand, panel members frequently raised the issue when considering whether they were comfortable recommending the intervention to others. One aspect of this was proportionality: they might accept an outcome of less policy importance if the intervention was relatively inexpensive but would not if it was expensive. Additionally, one panel member feared that an expensive intervention that required a lot of training and monitoring to produce results might be too difficult to successfully replicate in more ordinary settings. In the February 2009 meeting, it was decided that program cost should not be a criterion for Top Tier status but should be considered and reported with the recommendation, if deemed relevant.

The panel discussed whether a large multisite experiment should qualify as evidence meeting the replication standard. One classroom-based intervention was tested by randomly assigning 41 schools nationwide. Because the unit of analysis was the school, results at individual schools were not analyzed or reported separately but were aggregated to form one experimental–control group comparison per outcome measure. Some panel members considered this study a single randomized experiment; others accepted it as serving the purpose of a replication, because effects were observed over a large number of different settings. In this case,

limitations in the original study report added to their uncertainty. Some panel members stated that if they had learned that positive effects had been found in several schools rather than in only a few odd cases, they would have been more comfortable ruling this multisite experiment a replication.

Reviewers Initially Disagreed in Assessing Top Tier Status

Because detailed guidance was lacking, panel members, relying on individual judgment, arrived at split decisions (4–3 and 3–5) on two of the first four early childhood interventions reviewed, and only one intervention received a unanimous vote. Panel members expressed concern that because some criteria were not specifically defined, they had to use their professional judgment yet found that they interpreted the terms somewhat differently. This problem may have been aggravated by the fact that, as one member noted, they had not had a "perfect winner" that met all the top tier criteria. Indeed, a couple of members expressed their desire for a second category, like "promising," to allow them to communicate their belief in an intervention's high quality, despite the fact that its evidence did not meet all their criteria. In a discussion of their narrow (4–3) vote at their next meeting (February 2009), members suggested that they take more time to discuss their decisions, set a requirement for a two-thirds majority agreement, or ask for votes from members who did not attend the meeting. The latter suggestion was countered with concern that absent members would not be aware of their discussion, and the issue was deferred to see whether these differences might be resolved with time and discussion of other interventions. Disagreement over Top Tier status was less a problem with later reviews, held in February and July 2009, when none of the votes on Top Tier status were split decisions and three of seven votes were unanimous.

The Coalition reports that it plans to supplement guidance over time by accumulating case decisions rather than developing more detailed guidance on what constitutes sizable and sustained effects. The December 2008 and May 2009 public releases of the results of the Top Tier Evidence review of early childhood interventions provided brief discussion of examples of the panel's reasoning for accepting or not accepting specific interventions. In May 2009, the Coalition also published a revised version of the Checklist that removed the preference for outcomes measured a year after the intervention ended, replacing it with a less specific

reference: "over a long enough period to determine whether the intervention's effects lasted at least a year, hopefully longer."[16]

At the February 2009 meeting, Coalition staff stated that they had received a suggestion from external parties to consider introducing a second category of "promising" interventions that did not meet the top tier standard. Panel members agreed to discuss the idea further but noted the need to provide clear criteria for this category as well. For example, they said it was important to distinguish interventions that lacked good quality evaluations (and thus had unknown effectiveness) from those that simply lacked replication of sizable effects in a second randomized study. It was noted that broadening the criteria to include studies (and interventions) that the staff had previously screened out may require additional staff effort and, thus, resources beyond those of the current project.

Top Tier Follows Rigorous Standards but Is Limited for Identifying Effective Interventions

The Top Tier initiative's criteria for assessing evaluation quality conform to general social science research standards, but other features of the overall process differ from common practice for drawing conclusions about intervention effectiveness from a body of research. The initiative's choice of a broad topic fails to focus the review on how to achieve a specific outcome. Its narrow evidence criteria yield few recommendations and limited information on what works to inform policy and practice decisions.

Review Initiatives Share Criteria for Assessing Research Quality

The Top Tier and all six of the agency-supported review initiatives we examined assess evaluation quality on standard dimensions to determine whether a study provides credible evidence on effectiveness. These dimensions include the quality of research design and execution, the equivalence of treatment and comparison groups (as appropriate), adequacy of samples, the validity and reliability of outcome measures, and appropriateness of statistical analyses and reporting. Some initiatives included additional criteria or gave greater emphasis to some issues than others. The six agency-supported initiatives also employed several features to ensure the reliability of their quality assessments.

In general, assessing the quality of an impact evaluation's study design and execution involves considering how well the selected comparison protects

[16]Coalition, 2007, p. 5.

against the risk of bias in estimating the intervention's impact. For random assignment designs, this primarily consists of examining whether the assignment process was truly random, the experimental groups were equivalent before the intervention, and the groups remained separate and otherwise equivalent throughout the study. For other designs, the reviewer must examine the assignment process even more closely to detect whether a potential source of bias (such as higher motivation among volunteers) may have been introduced that could account for any differences observed in outcomes between the treatment and comparison groups. In addition to confirming the equivalence of the experimental groups at baseline, several review initiatives examine the extent of crossover or "contamination" between experimental groups throughout the study because this could blur the study's view of the intervention's true effects.

All seven review initiatives we examined assess whether a study's sample size was large enough to detect effects of a meaningful size. They also assess whether any sample attrition (or loss) over the course of the study was severe enough to question how well the remaining members represented the original sample or whether differential attrition may have created significant new differences between the experimental groups. Most review forms ask whether tests for statistical significance of group differences accounted for key study design features (for example, random assignment of groups rather than individuals), as well as for any deviations from initial group assignment (intention-to-treat analysis).[17]

The rating forms vary in structure and detail across the initiatives. For example, "appropriateness of statistical analyses" can be found under the category "reporting of the intervention's effects" on one form and in a category by itself on another form. In the Model Programs Guide rating form, "internal validity"—or the degree to which observed changes can be attributed to the intervention—is assessed through how well both the research design and the measurement of program activities and outcomes controlled for nine specific threats to validity.[18] The EPC rating form notes whether study participants were blind to the experimental groups they

[17]In intention-to-treat analysis, members of the treatment and control groups are retained in the group to which they were originally assigned, even if some treatment group members failed to participate in or complete the intervention or some control group members later gained access to the intervention. See Checklist, p. 4.

[18]These factors were initially outlined in the classic research design book by Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963).

belonged to—standard practice in studies for medical treatments but not as common in studies of social interventions, while the PRS form does not directly address study blinding in assessing extent of bias in forming study groups.

The major difference in rating study quality between the Top Tier initiative and the six other initiatives is a product of the top tier standard as set out in certain legislative provisions: the other initiatives accept well-designed, well-conducted quasi-experimental studies as credible evidence. Most of the federally supported initiatives recognize well-conducted randomized experiments as providing the most credible evidence of effectiveness by assigning them their highest rating for quality of research design, but three do not require them for interventions to receive their highest evidence rating: EPC, the Community Guide, and National Registry of Evidence-based Programs and Practices (NREPP). The Coalition has, since its inception, promoted randomized experiments as the highest-quality, unbiased method for assessing an intervention's true impact. Federal officials provided a number of reasons for including well-conducted quasi-experimental studies: (1) random assignment is not feasible for many of the interventions they studied, (2) study credibility is determined not by a particular research design but by its execution, (3) evidence from carefully controlled experimental settings may not reflect the benefits and harms observed in everyday practice, and (4) too few high-quality, relevant random assignment studies were available.

The Top Tier initiative states a preference for studies that test interventions in typical community settings over those run under ideal conditions but does not explicitly assess the quality (or fidelity) of program implementation. The requirement that results be shown in two or more randomized studies is an effort to demonstrate the applicability of intervention effects to other settings. However, four other review initiatives do explicitly assess intervention fidelity—the Community Guide, MPG, NREPP, and PRS—through either describing in detail the intervention's components or measuring participants' level of exposure. Poor implementation fidelity can weaken a study's ability to detect an intervention's potential effect and thus lessen confidence in the study as a true test of the intervention model. EPC and the Community Guide assess how well a study's selection of population and setting matched those in which it is likely to be applied; any notable differences in conditions would undermine the relevance or generalizability of study results to what can be expected in future applications.

All seven initiatives have experienced researchers with methodological and subject matter expertise rate the studies and use written guidance or codebooks to help ensure ratings consistency. Codebooks varied but most were more detailed than the Top Tier Checklist. Most of the initiatives also provided training to ensure consistency of ratings across reviewers. In each initiative, two or more reviewers rate the studies independently and then reach consensus on their ratings in consultation with other experts (such as consultants to or supervisors of the review). After the Top Tier initiative's staff screening review, staff and one advisory panel member independently review the quality of experimental evidence available on an intervention, before the panel as a group discussed and voted on whether it met the top tier standard. However, because the panel members did not independently rate study quality or the body of evidence, it is unknown how much of the variation in their overall assessment of the interventions reflected differences in their application of the criteria making up the Top Tier standard.

Broad Scope Fails to Focus on Effectiveness in Achieving Specific Outcomes

The Top Tier initiative's topic selection, emphasis on long-term effects, and narrow evidence criteria combine to provide limited information on the effectiveness of approaches for achieving specific outcomes. It is standard practice in research and evaluation syntheses to pose a clearly defined research question—such as, Which interventions have been found effective in achieving specific outcomes of interest for a specific population?—and then assemble and summarize the credible, relevant studies available to answer that question.[19] A well-specified research question clarifies the objective of the research and guides the selection of eligibility criteria for including studies in a systematic evidence review. In addition, some critics of systematic reviews in health care recommend using the intervention's theoretical framework or logic model to guide analyses toward answering questions about how and why an intervention works when it does.[20] Evaluators often construct a logic model—a diagram

[19]GAO, *The Evaluation Synthesis*, GAO/PEMD-10.1.2 (Washington, D.C.: March 1992); Institute of Medicine, *Knowing What Works in Health Care* (Washington, D.C.: National Academies Press, 2008); Iain Chalmers, "Trying to Do More Good Than Harm in Policy and Practice: The Role of Rigorous, Transparent, Up-to-Date Evaluations," *The Annals of the American Academy of Political and Social Science* (Thousand Oaks, Calif.: Sage, 2003); Agency for Healthcare Research and Quality, *Systems to Rate the Strength of Scientific Evidence* (Rockville, Md.: 2002).

[20]Institute of Medicine, *Knowing What Works*; N. Jackson and E. Waters, "Criteria for the Systematic Review of Health Promotion and Public Health Interventions," *Health Promotion International* (2005): 367–74.

showing the links between key intervention components and desired results—to explain the strategy or logic by which it is expected to achieve its goals.[21] The Top Tier initiative's approach focuses on critically appraising and summarizing the evidence without having first formulated a precise, unambiguous research question and the chain of logic underlying the interventions' hypothesized effects on the outcomes of interest.

Neither of the Top Tier initiative's topic selections—interventions for children ages 0–6 or youths ages 7–18—identify either a particular type of intervention, such as preschool or parent education, or a desired outcome, such as healthy cognitive and social development or prevention of substance abuse, that can frame and focus a review as in the other effectiveness reviews. The other initiatives have a clear purpose and focus: learning what has been effective in achieving a specific outcome or set of outcomes (for example, reducing youth involvement in criminal activity). Moreover, recognizing that an intervention might be successful on one outcome but not another, EPC, NREPP, and WWC rate the effectiveness of an intervention by each outcome. Even EPC, whose scope is the broadest of the initiatives we reviewed, focuses individual reviews by selecting a specific healthcare topic through a formal process of soliciting and reviewing nominations from key stakeholders, program partners, and the public. Their criteria for selecting review topics include disease burden for the general population or a priority population (such as children), controversy or uncertainty over the topic, costs associated with the condition, potential impact for improving health outcomes or reducing costs, relevance to federal health care programs, and availability of evidence and reasonably well-defined patient populations, interventions, and outcome measures.

The Top Tier initiative's emphasis on identifying interventions with long-term effects—up to 15 years later for some early childhood interventions—also leads away from focusing on how to achieve a specific outcome and could lead to capitalizing on chance results. A search for interventions with "sustained effects on important life outcomes," regardless of the content area, means assembling results on whatever outcomes—special education placement, high school graduation, teenage pregnancy, employment, or criminal arrest—the studies happen to have measured. This is of concern because it is often not clear why some long-

[21]GAO, *Program Evaluation: Strategies for Assessing How Information Dissemination Contributes to Agency Goals*, GAO-02-923 (Washington, D.C.: Sept. 30, 2002).

term outcomes were studied for some interventions and not others. Moreover, focusing on the achievement of long-term outcomes, without regard to the achievement of logically related short-term outcomes, raises questions about the meaning and reliability of those purported long-term program effects. For example, without a logic model or hypothesis linking preschool activities to improving children's self-control or some other intermediate outcome, it is unclear why one would expect to see effects on their delinquent behavior as adolescents. Indeed, one advisory panel member raised questions about the mechanism behind long-term effects measured on involvement in crime when effects on more conventional (for example, academic) outcomes disappeared after a few years. Later, he suggested that the panel should consider only outcomes the researcher identified as primary. Coalition staff said that reporting chance results is unlikely because the Top Tier criteria require the replication of results in multiple (or multi-site) studies, and they report any nonreplicated findings as needing confirmation in another study.

Unlike efforts to synthesize evaluation results in some systematic evidence reviews, the Top Tier initiative examines evidence on each intervention independently, without reference to similar interventions or, alternatively, to different interventions aimed at the same goal. Indeed, of the initiatives we reviewed, only EPC and the Community Guide directly compare the results of several similar interventions to gain insight into the conditions under which an approach may be successful. (WWC topic reports display effectiveness ratings by outcome for all interventions they reviewed in a given content area, such as early reading, but do not directly compare their approaches.) These two initiatives explicitly aim to build knowledge about what works in an area by developing logic models in advance to structure their evaluation review by defining the specific populations and outcome measures of interest. A third, MPG, considers the availability of a logic model and the quality of an intervention's research base in rating the quality of its evidence. Where appropriate evidence is available, EPCs conduct comparative effectiveness studies that directly compare the effectiveness, appropriateness, and safety of alternative approaches (such as drugs or medical procedures) to achieving the same health outcome. Officials at the other initiatives explained that they did not compare or combine results from different interventions because they did not find them similar enough to treat as replications of the same approach. However, most initiatives post the results of their reviews on their Web sites by key characteristics of the intervention (for example, activities or setting), outcomes measured, and population, so that viewers can search for particular types of interventions or compare their results.

Narrow Evidence Criteria Yield Limited Guidance for Practitioners

The Top Tier initiative's narrow primary criterion for study design quality—randomized experiments only—diverges from the other initiatives and limits the types of interventions they considered. In addition, the exclusivity of its top tier standard also diverges from the more common approach of rating the credibility of study findings along a continuum and resulted in the panel's recommending only 6 of 63 interventions for ages 0–18 reviewed as providing "sizable, sustained effects on important life outcomes." Thus, although they are not their primary audience, the Top Tier initiative provides practitioners with limited guidance on what works.

Two basic dimensions are assessed in effectiveness reviews: (1) the credibility of the evidence on program impact provided by an individual study or body of evidence, based on research quality and risk of bias in the individual studies, and (2) the size and consistency of effects observed in those studies. The six other evidence reviews report the credibility of the evidence on the interventions' effectiveness in terms of their level of confidence in the findings—either with a numerical score (0 to 4, NREPP) or on a scale (high, moderate, low, or insufficient, EPC). Scales permit an initiative to communicate intermediate levels of confidence in an intervention's results and to distinguish approaches with "promising" evidence from those with clearly inadequate evidence. Federal officials from initiatives using this more inclusive approach indicated that they believed that it provides more useful information and a broader range of choices for practitioners and policy makers who must decide which intervention is most appropriate and feasible for their local setting and available resources. To provide additional guidance to practitioners looking for an intervention to adopt, NREPP explicitly rates the interventions' readiness for dissemination by assessing the quality and availability of implementation materials, resources for training and ongoing support, and the quality assurance procedures the program developer provides.

Some initiatives, like Top Tier, provide a single rating of the effectiveness of an intervention by combining ratings of the credibility and size (and consistency, if available) of intervention effects. However, combining scores creates ambiguity in an intermediate strength of evidence rating—it could mean that reviewers found strong evidence of modest effects or weak evidence of strong effects. Other initiatives report on the credibility of results and the effect sizes separately. For example, WWC reports three summary ratings for an intervention's result on each outcome measured: an improvement index, providing a measure of the size of the intervention's effect; a rating of effectiveness, summarizing both study quality and the size and consistency of effects; and an extent of evidence

rating, reflecting the number and size of effectiveness studies reviewed. Thus, the viewer can scan and compare ratings on all three indexes in a list of interventions rank-ordered by the improvement index before examining more detailed information about each intervention and its evidence of effectiveness.

Randomized Experiments Can Provide the Most Credible Evidence of Effectiveness under Certain Conditions

In our review of the literature on program evaluation methods, we found general agreement that well-conducted randomized experiments are best suited for assessing intervention effectiveness where multiple causal influences lead to uncertainty about what has caused observed results but, also, that they are often difficult to carry out. Randomized experiments are considered best suited for interventions in which exposure to the intervention can be controlled and the treatment and control groups' experiences remain separate, intact, and distinct throughout the study. The evaluation methods literature also describes a variety of issues to consider in planning an evaluation of a program or of an intervention's effectiveness, including the expected use of the evaluation, the nature and implementation of program activities, and the resources available for the evaluation. Selecting a methodology follows, first, a determination that an effectiveness evaluation is warranted. It then requires balancing the need for sufficient rigor to draw firm conclusions with practical considerations of resources and the cooperation and protection of participants. Several other research designs are generally considered good alternatives to randomized experiments, especially when accompanied by specific features that help strengthen conclusions by ruling out plausible alternative explanations.

Conditions Necessary for Conducting Effectiveness Evaluations

In reviewing the literature on evaluation research methods, we found that randomized experiments are considered appropriate for assessing intervention effectiveness only *after* an intervention has met minimal requirements for an effectiveness evaluation—that the intervention is important, clearly defined, and well-implemented and the evaluation itself is adequately resourced. Conducting an impact evaluation of a social intervention often requires the expenditure of significant resources to both collect and analyze data on program results and estimate what would have happened in the absence of the program. Thus, impact evaluations need not be conducted for all interventions but reserved for when the effort and cost appear warranted. There may be more interest in an impact evaluation when the intervention addresses an important problem, there is interest in adopting the intervention elsewhere, and preliminary evidence suggests its effects may be positive, if uncertain. Of course, if the

intervention's effectiveness were known, then there would be no need for an evaluation. And if the intervention was known or believed to be ineffective or harmful, then it would seem wasteful as well as perhaps unethical to subject people to such a test. In addition to federal regulations concerning the protection of human research subjects, the ethical principles of relevant professional organizations require evaluators to try to avoid subjecting study participants to unreasonable risk, harm, or burden. This includes obtaining their fully informed consent.[22]

An impact evaluation is more likely to provide useful information about what works when the intervention consists of clearly defined activities and goals and has been well implemented. Having clarity about the nature of intended activities and evidence that critical intervention components were delivered to the intended targets helps strengthen confidence that those activities caused the observed results; it also improves the ability to replicate the results in another study. Confirming that the intervention was carried out as designed helps rule out a common explanation for why programs do not achieve their goals; when done before collecting expensive outcome data, it can also avoid wasting resources. Obtaining agreement with stakeholders on which outcomes to consider in defining success also helps ensure that the evaluation's results will be credible and useful to its intended audience. While not required, having a well-articulated logic model can help ensure shared expectations among stakeholders and define measures of a program's progress toward its ultimate goals.

Regardless of the evaluation approach, an impact evaluation may not be worth the effort unless the study is adequately staffed and funded to ensure the study is carried out rigorously. If, for example, an intervention's desired outcome consists of participants' actions back on the job after receiving training, then it is critical that all reasonable efforts are made to ensure that high-quality data on those actions are collected from as many participants as possible. Significant amounts of missing data raises the possibility that the persons reached are different from those who were not reached (perhaps more cooperative) and thus weakens confidence that the observed results reflect the true effect of the intervention. Similarly, it is important to invest in valid and reliable measures of desired outcomes

[22]See 45 C.F.R. Part 46 (2005) and, for example, the American Evaluation Association's *Guiding Principles for Evaluators*, revised in 2004. www.eval.org/Publications/GuidingPrinciples.asp

to avoid introducing error and imprecision that could blur the view of the intervention's effect.

Interventions Where Random Assignment Is Well Suited

We found in our review of the literature on evaluation research methods that randomized experiments are considered best suited for assessing intervention effectiveness where multiple causal influences lead to uncertainty about program effects and it is possible, ethical, and practical to conduct and maintain random assignment to minimize the effect of those influences.

When Random Assignment Is Needed

As noted earlier, when factors other than the intervention are expected to influence change in the desired outcome, the evaluator cannot be certain how much of any observed change reflects the effect of the intervention, as opposed to what would have occurred anyway without it. In contrast, controlled experiments are usually not needed to assess the effects of simple, comparatively self-contained processes like processing income tax returns. The volume and accuracy of tax returns processed simply reflect the characteristics of the returns filed and the agency's application of its rules and procedures. Thus, any change in the accuracy of processed returns is likely to result from change in the characteristics of either the returns or the agency's processes. In contrast, an evaluation assessing the impact of job training on participants' employment and earnings would need to control for other major influences on those outcomes—features of the local job market and the applicant pool. In this case, randomly assigning job training applicants (within a local job market) to either participate in the program (forming the treatment group) or not participate (forming the control group) helps ensure that the treatment and control groups will be equally affected.

When Random Assignment Is Possible, Ethical, and Practical

Random assignment is, of course, suited only to interventions in which the evaluator or program manager can control whether a person, group, or other entity is enrolled in or exposed to the intervention. Control over program exposure rules out the possibility that the process by which experimental groups are formed (especially, self-selection) may reflect preexisting differences between them that might also affect the outcome variable and, thus, obscure the treatment effect. For example, tobacco smokers who volunteer for a program to quit smoking are likely to be more highly motivated than tobacco smokers who do not volunteer. Thus, smoking cessation programs should randomly assign volunteers to receive services and compare them to other volunteers who do not receive services to avoid confounding the effects of the services with the effects of volunteers' greater motivation.

Random assignment is well suited for programs that are not universally available to the entire eligible population, so that some people will be denied access to the intervention in any case. This addresses one concern about whether a control group experiment is ethical. In fact, in many field settings, assignment by lottery has often been considered the most equitable way to assign individuals to participate in programs with limits on enrollment. Randomized experiments are especially well suited to demonstration programs for which a new approach is tested in a limited way before committing to apply it more broadly. Another ethical concern is that the control group should not be harmed by withholding needed services, but this can be averted by providing the control group with whatever services are considered standard practice. In this case, however, the evaluation will no longer be testing whether a new approach is effective *at all*; it will test whether it is more effective than standard practice.

Random assignment is also best suited for interventions in which the treatment and control groups' experiences remain separate, intact, and distinct throughout the life of the study so that any differences in outcomes can be confidently attributed to the intervention. It is important that control group participants not access comparable treatment in the community on their own (referred to as contamination). Their doing so could blur the distinction between the two groups' experiences. It is also preferred that control group and treatment group members not communicate, because knowing that they are being treated differently might influence their perceptions of their experience and, thus, their behavior. Sometimes people selected for an experimental treatment are motivated by the extra attention they receive; sometimes those not selected are motivated to work harder to compete with their peers. Thus, random assignment works best when participants have no strong beliefs about the advantage of the intervention being tested and information about their experimental status is not publicly known. For example, in comparing alternative reading curriculums in kindergarten classrooms, an evaluator needs to ensure that the teachers are equally well trained and do not have preexisting conceptions about the "better" curriculum. Sometimes this is best achieved by assigning whole schools—rather than individuals or classes—to the treatment and control groups, but this can become very expensive, since appropriate statistical analyses now require about as many schools to participate in a study as the number of classes participating in the simpler design.

Interventions are well suited for random assignment if the desired outcomes occur often enough to be observed with a reasonable sample

size or study length. Studies of infrequent but not rare outcomes—for example, those occurring about 5 percent of the time—may require moderately large samples (several hundred) to allow the detection of a difference between the experimental and control groups. Because of the practical difficulties of maintaining intact experimental groups over time, randomized experiments are also best suited for assessing outcomes that occur within 1 to 2 years after the intervention, depending on the circumstances. Although an intervention's key desired outcome may be a social, health, or environmental benefit that takes 10 or more years to fully develop, it may be prohibitively costly to follow a large enough proportion of both experimental groups over that time to ensure reliable results. Evaluators may then rely on intermediate outcomes, such as high-school graduation, as an adequate outcome measure rather than accepting the costs of directly measuring long-term effects on adult employment and earnings.

Interventions for Which Random Assignment Is Not Well Suited

Random assignment is not appropriate for a range of programs in which one cannot meet the requirements that make this strategy effective. They include entitlement programs or policies that apply to everyone, interventions that involve exposure to negative events, or interventions for which the evaluator cannot be sure about the nature of differences between the treatment and control groups' experiences.

Random Assignment Is Not Possible

For a few types of programs, random assignment to the intervention is not possible. One is when all eligible individuals are exposed to the intervention and legal restrictions do not permit excluding some people in order to form a comparison group. This includes entitlement programs such as veterans' benefits, Social Security, and Medicare, as well as programs operating under laws and regulations that explicitly prohibit (or require) a particular practice.

A second type of intervention for which random assignment is precluded is broadcast media communication where the individual—rather than the researcher—controls his or her exposure (consciously or not). This is true of radio, television, billboard, and Internet programming, in which the individual chooses whether and how long to hear or view a message or communication. To evaluate the effect of advertising or public service announcements in broadcast media, the evaluator is often limited to simply measuring the audience's exposure to it. However, sometimes it is possible to randomly assign advertisements to distinct local media markets and then compare their effects to other similar but distinct local markets.

A third type of program for which random assignment is generally not possible is comprehensive social reforms consisting of collective, coordinated actions by various parties in a community—whether school, organization, or neighborhood. In these highly interactive initiatives, it can be difficult to distinguish the activities and changes from the settings in which they take place. For example, some community development partnerships rely on increasing citizen involvement or changing the relationships between public and private organizations in order to foster conditions that are expected to improve services. Although one might randomly assign communities to receive community development support or not, the evaluator does not control who becomes involved or what activities take place, so it is difficult to trace the process that led to any observed effects.

Random assignment is often not accepted for testing interventions that prevent or mitigate harm because it is considered unethical to impose negative events or elevated risks of harm to test a remedy's effectiveness. Thus, one must wait for a hurricane or flood, for example, to learn if efforts to strengthen buildings prevented serious damage. Whether the evaluator is able to randomly apply different approaches to strengthening buildings may depend on whether the approaches appear to be equally likely to be successful in advance of a test. In some cases, the possibility that the intervention may fail may be considered an unacceptable risk. When evaluating alternative treatments for criminal offenders, local law enforcement officers may be unwilling to assign the offenders they consider to be the most dangerous to the less restrictive treatments.

As implied by the previous discussion of when random assignment is well suited, it may simply not be practical in a variety of circumstances. It may not be possible to convince program staff to form control groups by simple random assignment if it would deny services to some of the neediest individuals while providing service to some of the less needy. For example, individual tutoring in reading would usually be provided only to students with the lowest reading scores. In other cases, the desired outcome may be so rare or take so long to develop that the required sample sizes or prospective tracking of cases over time would be prohibitively expensive.

Finally, the evaluation literature cautions that as social interventions become more complex, representing a diverse set of local applications of a broad policy rather than a common set of activities, randomized experiments may become less informative. When how much of the intervention is actually delivered, or how it is expected to work, is

influenced by characteristics of the population or setting, one cannot be sure about the nature of the difference between the treatment and control group experiences or which factors influenced their outcomes. Diversity in the nature of the intervention can occur at the individual level, as when counselors draw on their experience to select the approach they believe is most appropriate for each patient. Or it can occur at a group level, as when grantees of federal flexible grant programs focus on different subpopulations as they address the needs of their local communities. In these cases, aggregating results over substantial variability in what the intervention entails may end up providing little guidance on what, exactly, works.

Rigorous Alternatives to Random Assignment Are Available

In our review of the literature on evaluation research methods, we identified several alternative methods for assessing intervention effectiveness when random assignment is not considered appropriate—quasi-experimental comparison group studies, statistical analyses of observational data, and in-depth case studies. Although experts differed in their opinion of how useful case studies are for estimating program impacts, several other research designs are generally considered good alternatives to randomized experiments, especially when accompanied by specific features that help strengthen conclusions by ruling out plausible alternative explanations.

Quasi-Experimental Comparison Groups

Quasi-experimental comparison group designs resemble randomized experiments in comparing the outcomes for treatment and control groups, except that individuals are not assigned to those groups randomly. Instead, unserved members of the targeted population are selected to serve as a control group that resembles the treatment group as much as possible on variables related to the desired outcome. This evaluation design is used with partial coverage programs for which random assignment is not possible, ethical, or practical. It is most successful in providing credible estimates of program effectiveness when the groups are formed in parallel ways and not based on self-selection—for example, by having been turned away from an oversubscribed service or living in a similar neighborhood where the intervention is not available. This approach requires statistical analyses to establish groups' equivalence at baseline.

Regression discontinuity analysis compares outcomes for a treatment and control group that are formed by having scores above or below a cut-point on a quantitative selection variable rather than through random assignment. When experimental groups are formed strictly on a cut-point

and group outcomes are analyzed for individuals close to the cut-point, the groups are left otherwise comparable except for the intervention. This technique is used where those considered most "deserving" are assigned to treatment, in order to address ethical concerns about denying services to those in need—for example, when additional tutoring is provided only to children with the lowest reading scores. The technique requires a quantitative assignment variable that users believe is a credible selection criterion, careful control over assignment to ensure that a strict cut-point is achieved, large sample sizes, and sophisticated statistical analysis.

Statistical Analyses of Observational Data

Interrupted time-series analysis compares trends in repeated measures of an outcome for a group before and after an intervention or policy is introduced, to learn if the desired change in outcome has occurred. Long data series are used to smooth out the effects of random fluctuations over time. Statistical modeling of simultaneous changes in important external factors helps control for their influence on the outcome and, thus, helps isolate the impact of the intervention. This approach is used for full-coverage programs in which it may not be possible to form or find an untreated comparison group, such as for change in state laws defining alcohol impairment of motor vehicle drivers ("blood alcohol concentration" laws). But because the technique relies on the availability of comparable information about the past—before a policy changed—it may be limited to use near the time of the policy change. The need for lengthy data series means it is typically used where the evaluator has access to long-term, detailed government statistical series or institutional records.

Observational or cross-sectional studies first measure the target population's level of exposure to the intervention rather than controlling its exposure and then comparing the outcomes of individuals receiving different levels of the intervention. Statistical analysis is used to control for other plausible influences. Level of exposure to the intervention can be measured by whether one was enrolled or how often one participated or heard the program message. This approach is used with full-coverage programs, for which it is impossible to directly form treatment and control groups; nonuniform programs, in which individuals receive different levels of exposure (such as to broadcast media); and interventions in which outcomes are observed too infrequently to make a prospective study practical. For example, an individual's annual risk of being in a car crash is so low that it would be impractical to randomly assign (and monitor) thousands of individuals to use (or not use) their seat belts in order to assess belts' effectiveness in preventing injuries during car crashes.

Because there is no evaluator control over assignment to the intervention, this approach requires sophisticated statistical analyses to limit the influence of any concurrent events or preexisting differences that may be associated with why people had different exposure to the intervention.

In-depth Case Studies

Case studies have been recommended for assessing the effectiveness of complex interventions in limited circumstances when other designs are not available. In program evaluation, in-depth case studies are typically used to provide descriptive information on how an intervention operates and produces outcomes and, thus, may help generate hypotheses about program effects. Case studies may also be used to test a theory of change, as when the evaluator specifies in advance the expected processes and outcomes, based on the program theory or logic model, and then collects detailed observations carefully designed to confirm or refute that model. This approach has been recommended for assessing comprehensive reforms that are so deeply integrated with the context (for example, the community) that no truly adequate comparison case can be found.[23] To support credible conclusions about program effects, the evaluator must make specific, refutable predictions of program effects and introduce controls for, or provide strong arguments against, other plausible explanations for observed effects. However, because a single case study most likely cannot provide credible information on what would have happened in the absence of the program, our experts noted that the evaluator cannot use this design to reliably estimate the magnitude of a program's effect.

Features That Can Strengthen Any Effectiveness Evaluation

Reviewing the literature and consulting with evaluation experts, we identified additional measurement and design features that can help strengthen conclusions about an intervention's impact from both randomized and nonrandomized designs. In general, they involve collecting additional data and targeting comparisons to help rule out plausible alternative explanations of the observed results. Since all evaluation methods have limitations, our confidence in concluding that an

[23]See Karen Fulbright-Anderson, Anne S. Kubisch, and James P. Connell, eds., *New Approaches to Evaluating Community Initiatives*, vol. 2, *Theory, Measurement, and Analysis* (Washington, D.C.: Aspen Institute, 1998), and Patricia Auspos and Anne S. Kubisch, *Building Knowledge about Community Change: Moving Beyond Evaluations* (Washington, D.C.: Aspen Institute, 2004).

intervention is effective is strengthened when the conclusion is supported by multiple forms of evidence.

Collecting Additional Data

Although collecting baseline data is an integral component of the statistical approaches to assessing effectiveness discussed above, both experiments and quasi-experiments would benefit from including pretest measures on program outcomes as well as other key variables. First, by chance, random assignment may not produce groups that are equivalent on several important variables known to correlate with program outcomes, so their baseline equivalence should always be checked. Second, in the absence of random assignment, ensuring the equivalence of the treatment and control groups on measures related to the desired outcome is critical. The effects of potential self-selection bias or other preexisting differences between the treatment and control groups can be minimized through selection modeling or "propensity score analysis." Essentially, one first develops a statistical model of the baseline differences between the individuals in the treatment and comparison groups on a number of important variables and then adjusts the observed outcomes for the initial differences between the groups to identify the net effect of the intervention.

Extending data collection either before or after the intervention can help rule out the influence of unrelated historical trends on the outcomes of interest. This is in principle similar to interrupted time-series analysis, yielding more observations to allow analysis of trends in outcomes over time in relation to the timing of program activities. For example, one could examine whether the outcome measure began to change before the intervention could plausibly have affected it, in which case the change was probably influenced by some other factor.

Another way to attempt to rule out plausible alternative explanations for observed results is to measure additional outcomes that are or are not expected to be influenced by the treatment, based on program theory. If one can predict a relatively unique pattern of expected outcomes for the intervention, in contrast to an alternative explanation, and if the study confirms that pattern, then the alternative explanation becomes less plausible.

Targeting Comparisons

In comparison group studies, the nature of the effect one detects is defined by the nature of the differences between the experiences of the treatment and control groups. For example, if the comparison group receives no assistance at all in gaining employment, then the evaluation can detect the full effect of all the employment assistance (including child

care) the treatment group receives. But if the comparison group also receives child care, then the evaluation can detect only the effect, or value added, of employment assistance above and beyond the effect of child care. Thus, one can carefully design comparisons to target specific questions or hypotheses about what is responsible for the observed results and control for specific threats to validity. For example, in evaluating the effects of providing new parents of infants with health consultation and parent training at home, the evaluator might compare them to another group of parents receiving only routine health check-ups to control for the level of attention the first group received and test the value added by the parent training.

Sometimes the evaluator can capitalize on natural variations in exposure to the intervention and analyze the patterns of effects to learn more about what is producing change. For example, little or no change in outcomes for dropouts—participants who left the program—might reflect either the dropouts' lower levels of motivation compared to other participants or their reduced exposure to the intervention. But if differences in outcomes are associated with different levels of exposure for administrative reasons (such as scheduling difficulties at one site), then those differences may be more likely to result from the intervention itself.

Gathering a Diverse Body of Evidence

As reflected in all the review initiatives we identified for this report, conclusions drawn from findings across multiple studies are generally considered more convincing than those based on a single study. The two basic reasons for this are that (1) each study is just one example of many potential experiences with an intervention, which may or may not represent that broader experience, and (2) each study employs one particular set of methods to measure an intervention's effect, which may be more or less likely than other methods to detect an effect. Thus, an analysis that carefully considers the results of diverse studies of an intervention is more likely to accurately identify when and for whom an intervention is effective.

A recurring theme in the evaluation literature is the tradeoffs made in constructing studies to rigorously identify program impact by reducing the influence of external factors. Studies of interventions tested in carefully controlled settings, a homogenous group of volunteer participants, and a comparison group that receives no services at all may not accurately portray the results that can be expected in more typical operations. To obtain a comprehensive, realistic picture of intervention effectiveness, reviewing the results of several studies conducted in different settings and populations, or large multisite studies, may help ensure that the results

observed are likely to be found, or replicated, elsewhere. This is particularly important when the characteristics of settings, such as different state laws, are expected to influence the effectiveness of a policy or practice applied nationally. For example, states set limits on how much income a family may have while receiving financial assistance, and these limits—which vary considerably from state to state—strongly influence the proportion of a state's assistance recipients who are currently employed. Thus, any federal policy regarding the employment of recipients is likely to affect one state's caseload quite differently from that of another.

Because every research method has inherent limitations, it is often advantageous to combine multiple measures or two or more designs in a study or group of studies to obtain a more comprehensive picture of an intervention. In addition to choosing whether to measure intermediate or long-term outcomes, evaluators may choose to collect, for example, student self-reports of violent behavior, teacher ratings of student disruptive behavior, or records of school disciplinary actions or referrals to the criminal justice system, which might yield different results. While randomized experiments are considered best-suited for assessing intervention impact, blended study designs can provide supplemental information on other important considerations of policy makers. For example, an in-depth case study of an intervention could be added to develop a deeper understanding of its costs and implementation requirements or to track participants' experiences to better understand the intervention's logic model. Alternatively, a cross-sectional survey of an intervention's participants and activities can help in assessing the extent of its reach to important subpopulations.

Concluding Observations

The Coalition provides a valuable service in encouraging government adoption of interventions with evidence of effectiveness and in drawing attention to the importance of evaluation quality in assessing that evidence. Reliable assessments of the credibility of evaluation results require expertise in research design and measurement, but their reliability can be improved by providing detailed guidance and training. The Top Tier initiative provides another useful model in that it engages experienced evaluation experts to make these quality assessments.

Requiring evidence from randomized experiments as sole proof of an intervention's effectiveness is likely to exclude many potentially effective and worthwhile practices for which random assignment is not practical. The broad range of studies assessed by the six federally supported

initiatives we examined demonstrates that other research designs can provide rigorous evidence of effectiveness if designed well and implemented with a thorough understanding of their vulnerability to potential sources of bias.

Assessing the importance of an intervention's outcomes entails drawing a judgment from subject matter expertise—the evaluator must understand the nature of the intervention, its expected effects, and the context in which it operates. Defining the outcome measures of interest in advance, in consultation with program stakeholders and other interested audiences, may help ensure the credibility and usefulness of a review's results. Deciding to adopt an intervention involves additional considerations—cost, ease of use, suitability to the local community, and available resources. Thus, practitioners will probably want information on these factors and on effectiveness when choosing an approach.

A comprehensive understanding of which practices or interventions are most effective for achieving specific outcomes requires a synthesis of credible evaluations that compares the costs and benefits of alternative practices across populations and settings. The ability to identify effective interventions would benefit from (1) better designed and implemented evaluations, (2) more detailed reporting on both the interventions and their evaluations, and (3) more evaluations that directly compare alternative interventions.

Agency and Third-Party Comments

The Coalition for Evidence-Based Policy provided written comments on a draft of this report, reprinted in appendix II. The Coalition stated it was pleased with the report's key findings on the transparency of its process and its adherence to rigorous standards in assessing research quality. While acknowledging the complementary value of well-conducted nonrandomized studies as part of a research agenda, the Coalition believes the report somewhat overstates the confidence one can place in such studies alone. The Coalition and the Departments of Education and Health and Human Services provided technical comments that were incorporated as appropriate throughout the text. The Department of Justice had no comments.

We are sending copies of this report to the Secretaries of Education, Justice, and Health and Human Services; the Director of the Office of Management and Budget; and appropriate congressional committees. The

report is also available at no charge on the GAO Web site at http://www.gao.gov.

If you have questions about this report, please contact me at (202) 512-2700 or kingsburyn@gao.gov. Contacts for our offices of Congressional Relations and Public Affairs are on the last page. Key contributors are listed in appendix III.

Nancy Kingsbury, Ph.D.
Managing Director
Applied Research and Methods

Appendix I: Steps Seven Evidence-Based Initiatives Take to Identify Effective Interventions

Search topic	Select studies	Review studies' quality	Synthesize evidence
1. Evidence-Based Practice Centers at the Agency for Healthcare Research and Quality			
Search for selected topics in health care services, pharmaceuticals, and medical devices through • Electronic databases • Major journals • Conference proceedings • Consultation with experts	Select • Randomized and quasi-experimental studies • Observational studies (e.g., cohort, case control)	A technical panel of expert physicians, content and methods experts, and other partners rates studies by outcome on • Study design and execution • Validity and reliability of outcome measures • Data analysis and reporting • Equivalence of comparison groups • Assessment of harm	Body of evidence on each outcome is scored on four domains: risk of bias, consistency, directness, and precision of effects. Strength of evidence for each outcome is classified as • High • Moderate • Low • Insufficient
2. Guide to Community Preventive Services at the Centers for Disease Control and Prevention			
Search for selected population-based policies, programs, and health care system interventions to improve health and promote safety through • Electronic databases • Major journals • Conference proceedings • Consultation with experts	Select • Randomized and quasi-experimental studies • Observational studies (e.g., time series, case control)	In consultation with method and subject matter experts, two trained reviewers independently rate studies using standardized forms on • Study design and execution • Validity and reliability of outcome measures • Data analysis and reporting • Intervention fidelity • Selection of population and setting	Body of evidence is assessed on number of studies, study quality, and size and consistency of effects to classify evidence of effectiveness as • Strong • Sufficient • Insufficient
3. HIV Prevention Research Synthesis at the Centers for Disease Control and Prevention			
Search for interventions that prevent new HIV/AIDS infections or behaviors that increase the risk of infection through • Electronic databases • Major journals • Conference proceedings • Consultation with experts • Nominations solicited from the public	Select randomized and quasi-experimental studies with one or more positive outcomes	Pairs of trained reviewers—Ph.D.s or M.A.s in behavioral science and health related areas—independently rate studies using standardized forms and codebook on • Study design and execution • Validity and reliability of outcome measures • Data analysis and reporting • Equivalence of comparison groups • Assessment of harm	Ratings of study quality and strength of findings are combined to classify interventions as • Best evidence • Promising evidence

Search topic	Select studies	Review studies' quality	Synthesize evidence
4. Model Programs Guide at the Office of Juvenile Justice and Delinquency Prevention			
Search for prevention and intervention programs to reduce problem behaviors (juvenile delinquency, violence, substance abuse) in at-risk juvenile population through • Electronic databases • Nominations solicited from the public	Select randomized and quasi-experimental studies with one or more positive outcomes and documentation of program implementation (fidelity)	A 3-person panel with 2 external Ph.D. content area experts—with a codebook and consensual agreement—independently rate studies on • Study design and execution • Validity and reliability of outcome measures • Data analysis and reporting • Equivalence of comparison groups • Intervention fidelity • Conceptual framework (logic and research base)	Ratings are combined across review criteria— including consistency of evidence—to classify interventions as • Exemplary • Effective • Promising
5. National Registry of Evidence-Based Programs and Practices at the Substance Abuse and Mental Health Services Administration			
Search for • Mental health promotion • Mental health treatment • Substance abuse prevention • Substance abuse treatment • Co-occurring disorders through • Electronic databases • Major journals • Nominations solicited from the public	Select randomized and quasi-experimental studies with one or more positive outcomes	Pairs of Ph.D. content specialists independently rate studies on • Study design and execution • Validity and reliability of outcome measures • Data analysis and reporting • Intervention fidelity Pairs of providers and implementation experts independently rate readiness for dissemination on • Implementation materials • Training and support resources • Quality assurance procedures	Summary research quality ratings (0–4) are provided for statistically significant outcomes. Interventions themselves are not rated. Scores on intervention readiness are averaged to provide a score of 0–4

Search topic	Select studies	Review studies' quality	Synthesize evidence
6. Top Tier Evidence Initiative at the Coalition for Evidence-Based Policy			
Search for early childhood (ages 0–6) and youth (ages 7–18) interventions through • Top evidence category of other evidence-based programs • Consultation with experts • Nominations solicited from the public	Select randomized studies with one or more positive outcomes	Team of M.A.s or Ph.D.s reviews studies and selects candidates for the advisory panel's review. Team reviews and one advisory panel member rates studies on • Study design and execution • Validity and reliability of outcome measures • Data analysis and reporting • Equivalence of comparison groups	The advisory panel reviews studies and quality ratings and assesses size and sustainability of effects in order to classify as Top Tier
7. What Works Clearinghouse at the Institute of Education Sciences			
Search for interventions that improve student achievement in • Early childhood education • Reading • Mathematics • Adolescent literacy • Dropout prevention • English language instruction through • Electronic databases • Major journals • Conference proceedings • Consultation with experts • Nominations solicited from the public	Select randomized and quasi-experimental studies	Two Ph.D. research analysts independently rate each study using codebook on • Study design and execution • Validity and reliability of outcome measures • Data analysis and reporting Ratings include • Meets evidence standards • Meets evidence standards with reservations	Across studies, ratings on quality of evidence and effect's direction, magnitude, and statistical significance for each outcome are combined and classified as • Positive • Potentially positive • Mixed • None discernible • Potentially negative • Negative Number and size of studies are rated separately as • Small • Medium to large

Appendix II: Comments from the Coalition for Evidence-Based Policy

Coalition for Evidence-Based Policy
A NON-PROFIT, NONPARTISAN ORGANIZATION

November 9, 2009

900 19th Street, NW
Suite 400
Washington, DC 20006
www.coalition4evidence.org

The Coalition for Evidence-Based Policy is pleased with GAO's confirmation of the Top Tier initiative's adherence to rigorous standards and overall transparency

The Coalition is pleased with the GAO report's key findings that the Top Tier initiative's criteria conform to general social science research standards (pp. 15-23), and that its process is mostly transparent (pp. 9-15). We also agree with its observation that the Top Tier initiative differs from common practice in its strong focus on randomized experiments, and would add that this was the initiative's goal from the start. Indeed, its stated purpose is to identify interventions meeting the top tier standard set out in recent Congressional legislation: "*well-designed randomized controlled trials [showing] sizeable, sustained effects on important ... outcomes*" (e.g., Public Laws 110-161 and 111-8).

Consistent with our initiative's unique focus on helping policymakers distinguish the relatively few interventions meeting this top evidentiary standard from the many that *claim* to, we have – as noted in the GAO report – identified 6 interventions as Top Tier out of the 63 reviewed thus far. The value of this process to policymakers is evidenced by the important impact these findings have already had on federal officials and legislation. For example, the initiative's findings for the Nurse-Family Partnership (NFP) have helped to spur the Administration and Congress' proposed national expansion of evidence-based home visitation. (The NFP study results are cited in the President's FY 2010 budget.) Similarly, the initiative's findings for the Carrera Adolescent Pregnancy Prevention program and Multidimensional Treatment Foster Care (MTFC) have helped inform the Administration and Congress' proposed evidence-based teen pregnancy prevention program. (The MTFC study results are cited in the Senate's FY10 Labor-HHS-Education Appropriations Committee report.[1])

In fact, OMB Director Peter Orszag recently posted on the OMB website a summary of the Administration's "two-tiered approach" to home visitation and teen pregnancy, which links to the Coalition's website.[2] The approach includes (i) funding for programs backed by strong evidence, which he identifies as "the top tier;" and (ii) additional funding for programs backed by "supportive evidence," with a requirement for rigorous evaluation that, if positive, could move them into the top tier.

Consistent with this Administration approach, we recognize (and agree with GAO) that nonrandomized studies provide important value – for example, in (i) informing policy decisions in areas where well-conducted randomized experiments are not feasible or not yet conducted; and (ii) identifying interventions that are particularly promising, and therefore ready to be evaluated in more definitive randomized experiments. We think the GAO report somewhat overstates the confidence one can place in nonrandomized findings alone, per (i) a recent National Academies recommendation[3] that evidence of effectiveness generally "cannot be considered definitive" without ultimate confirmation in well-conducted randomized experiments, "even if based on the next strongest designs;" and (ii) evidence that findings from nonrandomized studies are often overturned in definitive randomized experiments (see attachment). But the important and complementary value of well-conducted nonrandomized studies as part of an overall research agenda is a central theme of the Coalition's approach to evidence-based policy reform.

In conclusion, we appreciate GAO's thoughtful analysis, and will use its valuable observations to strengthen our initiative as it goes forward. Although the Congressionally-established top tier standard itself was not a main focus of the GAO report (as opposed to our process), we have attached some brief background on the standard and the reasons we support its use as an important element of appropriate policy initiatives.

Jon Baron, President

The Congressionally-established Top Tier evidence standard is based on a well-established concept in the scientific community, and strong evidence regarding the importance of random assignment.

Congress' Top Tier standard is based on a concept well-established in the scientific community – that when results of multiple (or multisite) well-conducted randomized experiments, carried out in real-world community settings, are available for a particular intervention, they generally comprise the most definitive evidence regarding that intervention's effectiveness. The standard further recognizes a key concept articulated in a recent National Academies recommendation: although many research methods can help identify effective interventions, evidence of effectiveness generally "cannot be considered definitive" without ultimate confirmation in well-conducted randomized experiments, "even if based on the next strongest designs."[3]

Although promising findings in nonrandomized quasi-experimental studies are valuable for decisionmaking in the absence of stronger evidence, too often such findings are overturned in subsequent, more definitive randomized experiments. Reviews in medicine, for example, have found that 50-80% of promising results from phase II (mostly quasi-experimental) studies are overturned in subsequent phase III randomized trials.[4] Similarly, in education, eight of the nine major randomized experiments sponsored by the Institute of Education Sciences since its creation in 2002 have found weak or no positive effects for the interventions being evaluated – interventions which, in many cases, were based on promising, mostly quasi-experimental evidence (e.g., the LETRS teacher professional development program for reading instruction).[5] Systematic "design replication" studies comparing well-conducted randomized experiments with quasi-experiments in welfare, employment, and education policy have also found that many widely-used and accepted quasi-experimental methods produce unreliable estimates of program impact.[6]

Thus, we support use of the Top Tier standard as a key element of policy initiatives seeking to scale up interventions backed by the most definitive evidence of sizeable, sustained effects, in areas where such proven interventions already exist. The standard has a strong basis in scientific authority and evidence, as reflected, for example, in the recent National Academies recommendation.

References

[1] Sen. Rept. 111-66.

[2] Peter Orszag's summary of the Administration's two-tiered approach is posted at http://www.whitehouse.gov/omb/blog/09/06/08/BuildingRigorousEvidencetoDrivePolicy/.

[3] National Research Council and Institute of Medicine. (2009). *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities*. Committee on Prevention of Mental Disorders and Substance Abuse Among Children, Youth and Young Adults: Research Advances and Promising Interventions. Mary Ellen O'Connell, Thomas Boat, and Kenneth E. Warner, Editors. Board on Children, Youth, and Families, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. Recommendation 12-4, page 371.

[4] John P. A. Ioannidis, "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association*, vol. 294, no. 2, July 13, 2005, pp. 218-228. Mohammad I. Zia, Lillian L. Siu, Greg R. Pond, and Eric X. Chen, "Comparison of Outcomes of Phase II Studies and Subsequent Randomized Control Studies Using Identical Chemotherapeutic Regimens," *Journal of Clinical Oncology*, vol. 23, no. 28, October 1, 2005, pp. 6982-6991. John K. Chan et. al., "Analysis of Phase II Studies on Targeted Agents and Subsequent Phase III Trials: What Are the Predictors for Success," *Journal of Clinical Oncology*, vol. 26, no. 9, March 20, 2008.

[5] *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement*, Institute of Education Sciences, NCEE 2008-4031, September 2008, http://ies.ed.gov/ncee/pubs/20084030/.

[6] Howard S. Bloom, Charles Michalopoulos, and Carolyn J. Hill, "Using Experiments to Assess Nonexperimental Comparison-Groups Methods for Measuring Program Effects," in *Learning More From Social Experiments: Evolving Analytic Approaches*, Russell Sage Foundation, 2005, pp. 173-235. Thomas D. Cook, William R. Shadish, and Vivian C. Wong, "Three Conditions Under Which Experiments and Observational Studies Often Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons," *Journal of Policy Analysis and Management,* vol. 27, no. 4, pp. 724-50. Steve Glazerman, Dan M. Levy, and David Myers, "Nonexperimental versus Experimental Estimates of Earnings Impact," *The American Annals of Political and Social Science,* vol. 589, September 2003, pp. 63-93.

Appendix III: GAO Contact and Staff Acknowledgments

GAO Contact

Nancy Kingsbury, (202) 512-2700 or kingsburyn@gao.gov

Staff Acknowledgments

In addition to the person named above, Stephanie Shipman, Assistant Director, and Valerie Caracelli made significant contributions to this report.

Bibliography

Agency for Healthcare Research and Quality. *Systems to Rate the Strength of Scientific Evidence: Summary.* Evidence Report/Technology Assessment No. 47. Rockville, Md.: U.S. Department of Health and Human Services, March 2002. www.ahrq.gov/clinic/epcsums/strengthsum.htm

Auspos, Patricia, and Anne C. Kubisch. *Building Knowledge about Community Change: Moving beyond Evaluations*. New York: The Aspen Institute, 2004.

Berk, Richard A. *Randomized Experiments as the Bronze Standard*. California Center for Population Research On-Line Working Paper Series CCPR-030-05. Los Angeles: August 2005. http://repositories.cdlib.org/uclastat/papers/2005080201

Boruch, Robert. "Encouraging the Flight of Error: Ethical Standards, Evidence Standards, and Randomized Trials." *New Directions for Evaluation* no. 113 (spring 2007): 55–73.

Boruch, Robert F., and Ellen Foley. "The Honestly Experimental Society: Sites and Other Entities as the Units of Allocation and Analysis in Randomized Trials." In *Validity and Social Experimentation: Donald Campbell's Legacy*, Leonard Bickman, ed. Thousand Oaks, Calif.: Sage, 2000.

Campbell, Donald T., and Julian C. Stanley. *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally, 1966.

Chalmers, Iain. "Trying to Do More Good Than Harm in Policy and Practice: The Role of Rigorous, Transparent, Up-to-date Evaluations." *The Annals of the American Academy of Political and Social Science* 589 (2003): 22–40.

Cook, Thomas D. "Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community Has Offered for Not Doing Them." *Educational Evaluation and Policy Analysis* 24:3 (2002): 175–99.

European Evaluation Society. *EES Statement: The Importance of a Methodologically Diverse Approach to Impact Evaluation—Specifically with Respect to Development Aid and Development Interventions*. Nijkerk, The Netherlands: December 2007. www.europeanevaluation.org

Flay, Brian R., and others. "Standards of Evidence: Criteria for Efficacy, Effectiveness, and Dissemination." *Prevention Science* 6:3 (2005): 151–75.

Fulbright-Anderson, Anne C. Kibisch, and James P. Connell, eds. *New Approaches to Evaluating Community Initiatives.* Vol. 2. *Theory, Measurement, and Analysis.* Washington, D.C.: The Aspen Institute, 1998.

Glazerman, Steven, Dan M. Levy, and David Myers. "Nonexperimental versus Experimental Estimates of Earnings Impacts." *The Annals of the American Academy of Political and Social Science* 589 (2003): 63–93.

Institute of Medicine. *Knowing What Works in Health Care: A Roadmap for the Nation.* Washington, D.C.: The National Academies Press, 2008.

Jackson, N., and E. Waters. "Criteria for the Systematic Review of Health Promotion and Public Health Interventions." For the Guidelines for Systematic Reviews in Health Promotion and Public Health Task Force. *Health Promotion International* 20:4 (2005): 367–74.

Julnes, George, and Debra J. Rog. "Pragmatic Support for Policies on Methodology." *New Directions for Evaluation* no. 113 (spring 2007): 129–47.

Mark, Melvin M., and Charles S. Reichardt. "Quasi-experimental and Correlational Designs: Methods for the Real World When Random Assignment Isn't Feasible." In *The Sage Handbook of Methods in Social Psychology*, Carol Sansone, Carolyn C. Morf, and A. T. Panter, eds. Thousand Oaks, Calif.: Sage, 2004.

Moffitt, Robert A. "The Role of Randomized Field Trials in Social Science Research: A Perspective from Evaluations of Reforms of Social Welfare Programs." *The American Behavioral Scientist* 47:5 (2004): 506–40.

National Research Council, Center for Education. *Scientific Research in Education*. Washington, D.C.: National Academies Press, 2002.

National Research Council, Committee on Law and Justice. *Improving Evaluation of Anticrime Programs*. Washington, D.C.: National Academies Press, 2005.

Orr, Larry L. *Social Experiments: Evaluating Public Programs with Experimental Methods.* Thousand Oaks, Calif.: Sage, 1999.

Posavac, Emil J., and Raymond G. Carey. *Program Evaluation: Methods and Case Studies*, 6th ed. Upper Saddle River, N.J.: Prentice Hall, 2003.

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman. *Evaluation: A Systematic Approach*, 7th ed. Thousand Oaks, Calif.: Sage, 2004.

Trochim, William M. K. President, American Evaluation Association, Chair, AEA Evaluation Policy Task Force. Letter to Robert Shea, Associate Director for Administration and Government Performance, Office of Management and Budget, Washington, D.C., March 7, 2008, and attachment, "Comments on 'What Constitutes Strong Evidence of a Program's Effectiveness?'" www.eval.org/EPTF.asp

Victora, Cesar G., Jean-Pierre Habicht, and Jennifer Bryce. "Evidence-Based Public Health: Moving beyond Randomized Trials." *American Journal of Public Health* 94:3 (March 2004): 400–05.

West, Stephen, and others. "Alternatives to the Randomized Controlled Trial." *American Journal of Public Health* 98:8 (August 2008): 1359–66.

Related GAO Products

Juvenile Justice: Technical Assistance and Better Defined Evaluation Plans Will Help to Improve Girls' Delinquency Programs. GAO-09-721R. Washington, D.C.: July 24, 2009.

Health-Care-Associated Infections in Hospitals: Leadership Needed from HHS to Prioritize Prevention Practices and Improve Data on These Infections. GAO-08-283. Washington, D.C.: March 31, 2008.

School Mental Health: Role of the Substance Abuse and Mental Health Services Administration and Factors Affecting Service Provision. GAO-08-19R. Washington, D.C.: October 5, 2007.

Abstinence Education: Efforts to Assess the Accuracy and Effectiveness of Federally Funded Programs. GAO-07-87. Washington, D.C.: October 3, 2006.

Program Evaluation: OMB's PART Reviews Increased Agencies' Attention to Improving Evidence of Program Results. GAO-06-67. Washington, D.C.: October 28, 2005.

Program Evaluation: Strategies for Assessing How Information Dissemination Contributes to Agency Goals. GAO-02-923. Washington, D.C.: September 30, 2002.

The Evaluation Synthesis. GAO/PEMD-10.1.2. Washington, D.C.: March 1992.

Designing Evaluations. GAO/PEMD-10.1.4. Washington, D.C.: March 1991.

Case Study Evaluations. GAO/PEMD-10.1.9. Washington, D.C.: November 1990.

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's Web site (www.gao.gov). Each weekday afternoon, GAO posts on its Web site newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to www.gao.gov and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's Web site, http://www.gao.gov/ordering.htm.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact:

Web site: www.gao.gov/fraudnet/fraudnet.htm
E-mail: fraudnet@gao.gov
Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Ralph Dawn, Managing Director, dawnr@gao.gov, (202) 512-4400
U.S. Government Accountability Office, 441 G Street NW, Room 7125
Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, youngc1@gao.gov, (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149
Washington, DC 20548