Order Code RL33301

CRS Report for Congress
Received through the CRS Web

Congress and Program Evaluation: An Overview of Randomized Controlled Trials (RCTs) and Related Issues

March 7, 2006

Clinton T. Brass
Analyst in American National Government
Government and Finance Division

Blas Nunez-Neto
Analyst in Social Legislation
Domestic Social Policy Division

Erin D. Williams
Specialist in Bioethical Policy
Domestic Social Policy Division

Congress and Program Evaluation: An Overview of Randomized Controlled Trials (RCTs) and Related Issues

Summary

Program evaluations can play an important role in public policy debates and in oversight of government programs, potentially affecting decisions about program design, operation, and funding. One technique that has received significant recent attention is the randomized controlled trial (RCT). There are also many other types of evaluation, including observational and qualitative designs.

An RCT attempts to estimate a program's impact upon an outcome of interest (e.g., crime rate). An RCT randomly assigns subjects to treatment and control groups, administers an intervention to the treatment group, and afterward measures the average difference between the groups. The quality of an RCT is typically assessed by its internal, external, and construct validity. At the federal level, RCTs have been a subject of interest and some controversy in education policy and the George W. Bush Administration's effort to integrate budgeting and performance using the Program Assessment Rating Tool (PART). In addition, in the 109th Congress, pending legislation provides for RCTs (e.g., Sections 3 and 15 of S. 1934; Section 114 of S. 667 (Senate committee-reported bill); and Section 5 of S. 1129).

Views about the practical capabilities and limitations of RCTs, compared to other evaluation designs, have sometimes been contentious. There is wide consensus that, under certain conditions, well-designed and implemented RCTs provide the most valid estimate of an intervention's impact, and can therefore provide useful information on whether, and the extent to which, an intervention causes favorable impacts for a large group of subjects, on average. However, RCTs are also seen as difficult to design and implement well. There also appears to be less consensus about what proportion of evaluations that are intended to estimate impacts should be RCTs and about the conditions under which RCTs are appropriate. Many observers argue that other types of evaluations are necessary complements to RCTs, or sometimes necessary substitutes for them, and can be used to establish causation, help bolster or undermine an RCT's findings, or in some situations validly estimate impacts. There is increasing consensus that a single study of any type is rarely sufficient to reliably support decision making. Many researchers have therefore embraced systematic reviews, which synthesize many similar or disparate studies.

A number of issues regarding RCTs might arise when Congress considers making program evaluation policy or when actors in the policy process present program evaluations to influence Congress. Should Congress focus on RCTs in these situations, a number of issues might be considered, including an RCT's parameters, capabilities, and limitations. In addition, Congress might examine the types of program evaluations that are necessary, question an evaluation's definitions or assumptions, consider how to appropriately use evaluation information in its learning and decision making, evaluate how much confidence to have in a study, and investigate whether agencies have capacity to properly conduct, interpret, and objectively present evaluations. This report will be updated in the 110th Congress.

Contents

Congress and Program Evaluation: An Overview of Randomized Controlled Trials (RCTs) and Related Issues

Introduction

Program evaluations can play an important role in public policy debates and in oversight of government programs, potentially affecting decisions about program design, operation, and funding. Many different techniques of program evaluation can be used and presented with an intention to inform and influence policy makers. One technique that has received significant recent attention in the federal government is the randomized controlled trial (RCT). This report discusses what RCTs are and identifies a number of issues regarding RCTs that might arise when Congress considers making program evaluation policy. For example, in the 109th Congress, Section 3 of S. 1934 (as introduced) would establish a priority for RCTs when evaluating offender reentry demonstration projects; Section 114 of S. 667 (Senate Finance Committee-reported bill) would require RCTs for demonstration projects for low-income families; and Section 5 of S. 1129 (as introduced) would call for RCTs for projects and policies of multilateral development banks. Issues regarding RCTs could also arise when actors in the policy process present specific program evaluations to Congress (e.g., in the President's budget proposals) to influence Congress's views and decision making. For many reasons, evaluations often merit scrutiny and care in interpretation.

Before discussing RCTs in detail, the report places them in context by discussing (1) questions that program evaluations are typically intended to address, (2) how RCTs relate to other program evaluation methods, and (3) two major roles that Congress often takes with regard to program evaluation. The report next describes the basic attributes of an RCT, major ways to judge an RCT's quality, and diverse views about the practical capabilities and limitations of RCTs as a form of program evaluation. In light of concerns about the reliability of individual studies to support decision making, the report also discusses how RCTs can fit into systematic reviews of many evaluations. The report next highlights two areas where RCTs have garnered recent attention — in education policy and the President's annual budget proposal to Congress. Finally, the report identifies potential issues for Congress that could apply to the highlighted cases, oversight of other policy areas, and pending legislation. Because the vocabulary of program evaluation can be confusing, an appendix provides a glossary with definitions of selected terms.

Congress, Program Evaluation, and Policy Making

Key Questions about Government Programs and Policies. Citizens, elected officials, civil servants, interest groups, and many other participants in governance of the United States have an interest in the performance and results of

government programs and policies. To that end, stakeholders might want answers to many questions about programs and policies. For example, how should public policy problem(s) be defined? Is a program addressing some or all of the problem(s)? How well are federal programs and policies managed? What are they achieving? How can they improve? How are stakeholders affected? What unintended consequences might result? In the future, what activities and policies should the federal government pursue in order to best serve the public? What resources should be devoted to a program or policy?

In addition, stakeholders might want answers to questions about the quality of evaluations that are brought to policy discussions, given that participants in the policy process will not always advertise weaknesses in studies that also happen to support their policy positions. What might those weaknesses be? Stakeholders might also ask how well federal agencies evaluate the programs they lead and administer. For example, what methods are appropriate to assess a given type of program or policy? Given the available quantity and quality of research, what degree of confidence should be placed in findings, to date? Do agencies have sufficient capacity to evaluate their programs? Are they performing the necessary types of evaluation? Do agencies have sufficient independence to credibly evaluate their programs and policies? What role should agencies play in evaluating programs?

At times, many or all of these questions might be of interest to Congress and program stakeholders. All of them will typically be of interest to agency program managers and leaders. Therefore, any of these questions might be potential subjects of congressional oversight or law making.

Program Evaluation and Informed Policy Making. In response to questions like those posed above, program evaluations can be introduced into policy discussions by actors in the policy-making process. These actors — who include organizations and individuals both inside and outside of government — might be interest groups, think tanks, academics, legislators, state or local governments, the President, federal agencies, or nonpartisan institutions. Many actors bring evaluations to policy discussions on their own initiative, oftentimes to emphasize the results or findings that they interpret to support their positions. Some actors (e.g., federal agencies) might bring evaluations in response to legislative or executive branch requirements. Depending on many circumstances, the evaluations that agencies bring might, or might not, support the policy views of the agency's head or the President.

When actors bring program evaluations into policy discussions, the studies will oftentimes use different approaches, because there are many possible ways to help answer the questions cited previously. The term *program evaluation*, therefore, has in practice been interpreted in several ways. For example, there is no consensus definition for the term *program*. In practice, the term has been used to refer to a government policy, activity, project, initiative, law, tax provision, function, or set thereof. Accordingly, this report uses the term *program* to refer to any of these things, as appropriate, that someone might wish to evaluate.[1] The term *evaluation*

---

[1] In program evaluation, the terms *intervention* and *treatment* are sometimes used as (continued...)

can seem similarly ambiguous. A recent reference work in the program evaluation literature defined *evaluation* as "an applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs, value, merit, worth, significance, or quality of a program, product, person, policy, proposal, or plan."[2] Perhaps with many of these considerations in mind, Congress defined *program evaluation*, for purposes of the Government Performance and Results Act of 1993 (GPRA), as "an assessment, through objective measurement and systematic analysis, of the manner and extent to which Federal programs achieve intended objectives."[3] GPRA requires most executive branch agencies to develop five-year strategic plans, annual performance plans (including goals and performance indicators, among other things), and annual program performance reports. When reporting GPRA to the Senate, the Senate Committee on Governmental Affairs contemplated that not all forms of evaluation and measurement would necessarily be quantifiable, because of the diversity of federal government activities.[4] In sum, program evaluation has been considered in practice, in the scholarly literature, and under GPRA as concerned with investigating both a program's operations and its results. Furthermore, program evaluation has been seen as (1) informing conclusions at particular points in time, and also (2) a cumulative process over time of forming conclusions, as more evaluation information is collected and interpreted.

Program evaluations might help inform policy makers, including Members and committees of Congress in their authorizations, appropriations, and oversight work. However, viewpoints about program evaluations can be contentious, both in policy debates and among expert evaluators. In their interactions with Congress, many actors cite program evaluations as part of the rationale for policy changes. In addition, observers and practitioners sometimes disagree about the practical capabilities and limitations of various program evaluation methods, the quality of an individual evaluation, or how a study's findings should be interpreted and used. Therefore, many observers believe it is important for policy makers, including Members and committees of Congress, to be informed consumers of evaluation information when weighing these considerations and making policy decisions.

Types of Program Evaluation

Practitioners and theorists categorize different types of program evaluation (sometimes referred to as different designs or methods) in several ways.

---

[1] (...continued)
synonyms for *program*.

[2] Deborah M. Fournier, "Evaluation," in Sandra Mathison, ed., *Encyclopedia of Evaluation* (Thousand Oaks, CA: SAGE Publications, 2005), p. 139.

[3] P.L. 103-62; 107 Stat. 285, at 288. The Senate Committee on Governmental Affairs report that accompanied GPRA also clarified that the term should be read to include evaluations of "unintended" results, program implementation, and operating policies and practices, but not routine program monitoring. See U.S. Congress, Senate Committee on Governmental Affairs, *Government Performance and Results Act of 1993*, report to accompany S. 20, 103rd Cong., 1st sess., June 16, 1993, S.Rept. 103-58 (Washington: GPO, 1993), pp. 32-33.

[4] Ibid., p. 30. See also Section 4(b) of the act, which was codified as 31 U.S.C. § 1115(b).

Unfortunately, these categorizations are not always consistent with each other, and practitioners and theorists do not always use consistent terminology to describe program evaluations. They sometimes use different definitions for the same term, or use different terms as synonyms for the same definition. This report discusses some of these methods, but does not attempt to provide an overall taxonomy for program evaluation types.[5] The sections below provide basic descriptions of RCTs and other methods. A more detailed description and discussion of RCTs is located later in the report.

Randomized Controlled Trials (RCTs). One program evaluation method that has been a subject of recent interest at the federal level, as well as some controversy, is the RCT. As discussed later in this report, an RCT attempts to estimate a program's *impact* on an *outcome of interest*. An *outcome of interest* is something, oftentimes a public policy goal, that one or more stakeholders care about (e.g., unemployment rate, which many actors might like to be lower). An *impact* is an estimated measurement of how an intervention affected the outcome of interest, compared to what would have happened without the intervention.[6] A simple RCT *randomly assigns* some subjects to one or more *treatment groups* (also sometimes called *experimental* or *intervention groups*) and others to a *control group*. The treatment group participates in the program being evaluated and the control group does not. After the treatment group experiences the intervention, an RCT compares what happens to the two groups by measuring the difference between the two groups on the outcome of interest. This difference is considered an estimate of the program's impact. The terms *randomized field trial (RFT)*, *random assignment design*, *experimental design*, *random experiment*, and *social experiment* are sometimes used as synonyms for RCT, and vice versa. However, use of the word *field* in this context is often intended to imply that an evaluation is being conducted in a more naturalistic setting instead of a laboratory or other artificial environment.

---

[5] A large array of publications are available concerning different types of program evaluation. For perspectives on different types of evaluations and the purposes for which they can be used, see, for example: U.S. Government Accountability Office, *Performance Measurement and Evaluation: Definitions and Relationships*, GAO-05-739SP, May 2005; Joseph S. Wholey, Harry P. Hatry, and Kathryn E. Newcomer, eds., *Handbook of Practical Program Evaluation* (San Francisco: Jossey-Bass, 1994); Joseph S. Wholey, Harry P. Hatry, and Kathryn E. Newcomer, eds., *Handbook of Practical Program Evaluation*, 2nd ed. (San Francisco: Jossey-Bass, 2004); Peter H. Rossi, Mark W. Lipsey, and Howard E. Freeman, *Evaluation: A Systematic Approach*, 7th ed. (Thousand Oaks, CA: Sage Publications, 2004); Lawrence B. Mohr, *Impact Analysis for Program Evaluation*, 2nd ed. (Thousand Oaks, CA: Sage Publications, 1995); William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin, 2001); and Michael Quinn Patton, *Qualitative Research and Evaluation Methods*, 3rd ed. (Thousand Oaks, CA: Sage Publications, 2002).

[6] For example, if the unemployment rate in a geographic area would have been 6% without a program intervention, but was estimated to be 5% because of the intervention, the impact would be a 1% reduction in the unemployment rate (i.e., 6% minus 5% equals an impact of 1%), or, alternatively, as a 16.7% reduction in the unemployment rate, if one characterizes the impact as a proportion of the prior unemployment rate. Some theorists and practitioners use the term *effect* as a synonym for impact. This report uses only the term *impact* for this definition to avoid potential confusion sometimes associated with the term *effect*.

Many Other Methods. Many other types of program evaluation that are not RCTs can also be conducted in order to address one or more of the questions posed at the beginning of this report. Some of these "other" methods have been called, as a group, *observational* designs. The term *observational design* has been used in different ways, but often refers to empirical and qualitative studies of many types that are intended to help explain cause-and-effect relationships, but that do not attempt to approximate an experimental design. *Quasi-experimental designs* refer to studies that attempt to estimate a treatment's impact on a group of subjects, but, in contrast with RCTs, do not have random assignment to treatment and control groups. Some quasi-experiments are controlled studies (i.e, with a control group and at least one treatment group), but others lack a control group. Some quasi-experiments do not measure the outcome of interest before the treatment takes place. Many observers and practitioners consider quasi-experiments to be a form of observational design, but others put them in their own category. Methods that attempt to estimate an impact are sometimes called *impact analysis* designs. *Qualitative evaluation* often refers to judging the effectiveness of a program (e.g., whether it accomplishes its goals) by conducting open-ended interviews, directly observing program implementation and outcomes, reviewing documents, and constructing case studies. As used by different researchers, the term *nonexperiment* has been used at times to refer specifically to quasi-experiments and other times to anything that is not an RCT. *Systematic reviews* synthesize the results of many studies, as discussed later in this report. Many other program evaluation methods, including surveys and cost-benefit analyses, are also used to assess programs.[7] Because this report focuses on RCTs and related issues, only some of these "other" methods are discussed further in this report.

Possible Congressional Roles Concerning Program Evaluation

Congress can assume at least two major roles regarding program evaluation. These roles might be called (1) "making program evaluation policy" and, (2) when presented with one or more program evaluations, "scrutinizing and learning from program evaluations." Each of these broad roles can raise a number of issues for Congress regarding program evaluations generally, as well as RCTs specifically.

Making Program Evaluation Policy. First, Congress might make policy regarding how, when, and the extent to which agencies are to conduct, fund, or use program evaluations.[8] For example, Congress might, among other things, establish

---

[7] The term *performance measurement* can mean many things, but is usually considered different from *program evaluation*. Frequently, performance measurement refers to ongoing and periodic monitoring and reporting of program operations or accomplishments (e.g., progress toward quantitative goals), and sometimes also statistical information related to, but not necessarily influenceable by, a program.

[8] Congress sets program evaluation policy both for the executive branch generally (e.g., Government Performance and Results Act of 1993, 107 Stat. 285) and in specific policy areas (e.g., Education Sciences Reform Act of 2002, 116 Stat. 1940). Congress has also set policy to enhance Congress's own institutional capacity, and the capacity of its supporting

(continued...)

agencies or offices that have missions to evaluate programs, require that a percentage of a program's funding be devoted to program evaluation activities, appropriate funds for specific evaluations, articulate what questions should be studied, or specify what methods should or must be used. When policy makers consider and make these decisions, two considerations that many observers would likely consider important are for policy makers to be aware of both the practical capabilities and limitations of various program evaluation methods, and also how those capabilities and limitations might be balanced in light of multiple evaluation objectives.

Scrutinizing and Learning From Program Evaluations. Second, Members of Congress might use specific program evaluations to help inform their thinking, policy making, and oversight of federal policies. In the course of Congress's lawmaking and oversight work, actors inside and outside of government frequently cite program evaluations to justify their policy proposals and recommendations. In these situations, consumers of evaluation information, including Congress, can face challenges of assessing (1) quality and depth of evaluation information, which can be uneven, and (2) the relevance of evaluation information to a policy problem, which can vary. Therefore, should Congress wish to critically assess or scrutinize program evaluations, having insight into how to assess the quality, depth, and relevance of evaluation methods might be helpful.

Program evaluations themselves can help inform policy in several ways. Among other things, they can provide deeper understanding of a policy problem, suggest possible ways to modify or improve a program or policy, provide perspectives on whether goals are being accomplished, reveal consequences that might not have been intended, and inform deliberations regarding the allocation of scarce resources. Nonetheless, observers and stakeholders frequently disagree on the appropriate goals of government activities, which can make evaluations controversial. Furthermore, because "[p]rogram evaluation is a site for the resolution of ethical and democratic dilemmas,"[9] any assessment of a program's *merit* or *worth* is arguably always made in part through the lens of an observer's priorities, beliefs, values, and ethics. *Merit* and *worth* are program evaluation terms that often are defined as the overall intrinsic and extrinsic value, respectively, of a program to individuals and society.[10] Even when there is some consensus on goals, it has sometimes been difficult or impossible to specify with a single number, program evaluation, performance measure, or even group of evaluations and performance measures, how to comprehensively judge an organization's or program's success in accomplishing its mission. Thus, as experience has shown, the concepts of merit and worth are often in the eye of the beholder. Still, evaluations might help clarify what

---

[8] (...continued)
agencies, to evaluate policy and consider program evaluations (e.g., Legislative Reorganization Act of 1970, 84 Stat. 1140; Congressional Budget Act of 1974, 88 Stat. 297).

[9] Saville Kushner, "Program Evaluation," in Sandra Mathison, ed., *Encyclopedia of Evaluation*, p. 337.

[10] See Michael Scriven, *Evaluation Thesaurus*, 4th ed. (Newbury Park, CA: SAGE Publications, 1991), pp. 227-228, 382-383.

programs are accomplishing, and, when evaluations are done well, help policy makers make informed judgments when reconciling diverse views about policy problems and values.

Randomized Controlled Trials (RCTs)

What are RCTs?

RCT Defined. As briefly noted earlier, an RCT is a type of program evaluation that seeks to assess whether a program had an impact for one or more outcomes of interest (e.g., number of weeks a person remains unemployed) compared to what would have happened without the program. An *impact* is usually calculated for a large sample of subjects as the difference between (1) a measurement of the outcome of interest after an intervention takes place, averaged across subjects who received the treatment, and (2) a measurement of the outcome of interest after the intervention, averaged across subjects who did not receive the treatment. The study randomly assigns the subjects (also called *units of analysis*)[11] to one or more treatment groups and also a control group. A treatment group experiences the intervention, and the control group does not. After the intervention, measurement of the outcome of interest for the treatment group provides information on how the intervention might have affected these subjects. The control group, by contrast, is intended to simulate what would have happened to the treatment group subjects if they had not received the intervention. Depending on the policy area studied, an intervention could be, for example, a training regimen for unemployed workers or a new policy to reduce crime. Because the estimated impact is an average across subjects, the impact reflects the weighted average of the subjects who experienced favorable impacts, subjects who did not experience a change, and others who experienced unfavorable impacts.[12]

In theory, random assignment helps ensure that all of the groups in the study are made statistically equivalent at the beginning of the study.[13] If the only important difference in the subsequent experience of each group is the intervention, then

---

[11] In an RCT, the units of analysis are typically individual persons, but sometimes units might be things or organizations like schools, hospitals, or police stations.

[12] The concept that different subjects might respond differently to a treatment and receive different impacts, as opposed to assuming that everyone receives the same impact, has sometimes been referred to as "heterogeneous treatment impacts."

[13] Randomly assigning subjects to an intervention group and a comparison group increases confidence that on average, the two groups are initially "comparable." Random assignment also allows the use of certain statistical techniques for validly estimating an impact. However, sometimes the randomization is "unlucky" and does not necessarily result in comparable groups. The statistical techniques attempt to account for this chance. The term *random assignment* is different from the term *random selection*. *Random assignment* refers to assigning subjects to different groups or treatments in a controlled study. *Random selection* refers to how one draws a sample from a larger population (e.g., to undertake a survey that is intended to be representative of a broader population).

differences in the outcome of interest that are observed at the end of the trial can be attributed with greater confidence to the intervention, rather than to initial differences between the groups. Various statistical tools can be used to estimate whether observed differences are likely due to the intervention (i.e., the difference is found to be *statistically significant* with a small chance of error) or to chance.[14]

The quality of an RCT is often assessed by two criteria: *internal validity* and *external validity*. A third criterion, *construct validity*, is not always discussed, but is also considered important for judging an evaluation's quality.

Internal Validity. *Internal validity* is typically defined as the confidence with which one can state that the impact found or implied by a study was caused by the intervention being studied. For example, if an RCT of an aftercare program for juveniles shows that the juveniles who attended a program re-commit crimes (recidivate) at a lower rate than the juveniles who did not attend the program (the control group), an assessment of the study's internal validity would suggest whether this result was due to the aftercare program or whether it might have been due to some other factor. Internal validity is predicated on the methodological rigor of the study and an absence of other factors, unrelated to the program, affecting the outcome of interest for either the treatment or control group differently from the other group. The term also reflects that the better designed and implemented a study is, the more reliable its conclusions about causation will tend to be.[15] From the perspective of internal validity, the methodological rigor of an RCT study can depend on a number of factors, including, but not limited to the following:

- how effectively the random assignment of units creates statistically equivalent groups;
- whether the group of subjects is sufficiently numerous to ensure that an impact large enough to be of interest to stakeholders, if it occurs, will be found statistically significant (the more numerous the units

---

[14] The term *statistical significance* has different meanings in different contexts, even though each meaning is based on the same statistical concepts. In the context of an RCT, a finding of statistical significance is typically interpreted as a level of confidence (usually expressed as a probability, e.g., 95%, which is also referred to as "significance at the .05 level") that an impact is not merely the result of random variation. Assuming the RCT suffered from no defects, this finding would indicate that at least some of the measured impact may with substantial confidence (e.g., 95% confidence) be attributed to the treatment as a cause. Stated another way, significance at the .05 level indicates that there a 1 in 20 chance that the observed difference could have occurred by chance, if a program actually had no impact. However, simply because an estimated impact is found to be statistically significant does not necessarily mean the impact is large or important. See Lawrence B. Mohr, *Impact Analysis for Program Evaluation*, pp. 124, 127-130. See also Thomas H. Wonnacott and Ronald J. Wonnacott, *Introductory Statistics for Business and Economics*, 4th ed. (New York: Wiley, 1990), ch. 9.

[15] Thomas D. Cook and Donald T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* (Chicago: Rand McNally, 1979); Marilynn B. Brewer, "Internal Validity," in Michael S. Lewis-Beck, Alan Bryman, and Tim Futing Liao, eds., *The SAGE Enclyclopedia of Social Science Research Methods* (Thousand Oaks, CA: SAGE, 2004), p. 502.

in the group being studied, the better chance there is of detecting a program's potential impact);
- whether attrition in the treatment and control groups (e.g., subjects dropping out of the study) is comparable with respect to the attributes of subjects;
- whether factors other than the treatment "contaminate" one group but not the other (e.g., due to problems with delivering the treatment or incomplete environmental controls);
- whether the behavior of researchers or subjects is affected because they know who is receiving a treatment or not receiving a treatment (ideally, RCTs are double-blind studies, in which neither the subjects nor the researchers know which group gets the treatment, but double-blind studies in social science are uncommon);
- whether the units being studied comply with the intervention being provided (e.g., did the patient take the medicine being studied?);
- whether the presence of a randomized evaluation influences the treatment group's experience (e.g., randomization altering the process of selection into the program);
- whether subjects in the control group can obtain close substitutes for the treatment outside the program; and
- whether data collection and analysis procedures are reliable.

External Validity. *External validity* is typically defined as the extent to which an intervention being studied can be applied to other settings, times, or groups of subjects and be expected to deliver a similar impact on an outcome of interest.[16] Thus, external validity relates to both (1) whether the intervention itself can be replicated with high confidence and (2) whether an intervention will most likely result in a similar impact in other situations or environments, or with other subjects. In practice, some users of the term emphasize the second aspect noted here. The terms *generalizability*, *replicability*, and *repeatability* are sometimes used as synonyms for external validity.

The external validity of a study can depend on a variety of factors. As noted above, one factor is the confidence a person has that an intervention itself can be replicated. For example, if a new curriculum is introduced in a school, it is possible that the school might deviate from the prescribed curriculum in order to accommodate events or student needs unanticipated by the designers of the new curriculum or the study researchers. Unless the deviation were clearly documented, it might be difficult or impossible to replicate the same intervention in other sites. Furthermore, if the deviation affects the study's outcome of interest either positively or negatively compared to what it would have been without the deviation, the study's findings might not be generalizable.

Another major factor relating to external validity, and one of the most frequently cited, is the way in which a study's subjects were selected. For example, the results of a study measuring the impact of a certain curriculum in schools in Boston, New York City, and Philadelphia might not be generalizable to classrooms in small towns

---

[16] See Lawrence B. Mohr, *Impact Analysis for Program Evaluation*, pp. 92-97.

in the Midwest due to potential differences among the underlying populations and environments. The results also might not be generalizable to classrooms in other large cities. Among RCTs, the most generalizable are often those that estimate the same intervention's impact in different settings (also known as multi-site RCTs), as well as those which feature samples of subjects with diverse socioeconomic and demographic characteristics. RCTs can also be more generalizable if subjects are randomly selected from a certain population to participate in the RCT, in order to make the subjects more representative of that population. This is not always possible, however, because sometimes subjects can be selected only in a nonrandom way (e.g., if subjects volunteer for the program). Attrition can also affect external validity. Even if attrition among the treatment and control groups is equivalent, if too many people with certain characteristics drop out of a study, a treatment group that was diverse enough to provide some generalizability at the beginning of the study may no longer have as much at the end, even if the sample remains large enough to produce statistically significant results concerning the intervention's impact.

In response to many of these considerations, researchers sometimes carefully describe the intervention, a study's subjects, and the local environment so that other researchers and stakeholders can attempt to assess a study's external validity.

Construct Validity. A third type of validity that is considered to be important in any type of evaluation, but is not always explicitly discussed, is *construct validity*. As one reference work explains,

> [s]implistically, construct validity is about naming something (a program, an attribute) accurately. For example, when an evaluator measures student achievement, the issue of construct validity entails a judgment about whether the measures or the operationalization of achievement are really measuring student achievement, as opposed to, for example, social capital.[17]

The "construct" in this example is a specific way of measuring student achievement. Thus, one definition of construct validity concerns the extent to which a study actually evaluates the question it is being represented as evaluating. Perhaps unsurprisingly, actors in the policy process will sometimes have different views on appropriate ways to measure student achievement, or more generally, appropriate ways to measure "success" in achieving a program's mission and goals.

Alternatively, when the term construct validity is used in relation to a program rather than a measurement method, it often refers to the extent to which an actual program reflects one's ideas and theories of (1) how the program is supposed to operate, and (2) the causal mechanism through which it is supposed to achieve outcomes.[18] This view of construct validity can be important when attempting to improve a program (e.g., modifying it to better achieve goals) or to understand the circumstances that are necessary for the program to achieve similar results at another

---

[17] "Construct Validity," in Sandra Mathison, ed., *Encyclopedia of Evaluation*, p. 81.

[18] For discussion, see William M.K. Trochim, *The Research Methods Knowledge Base*, 2nd ed. (Cincinnati, OH: Atomicdogpublishing.com, 2001), p. 69.

time or place, or with different subjects. With insight into the mechanism of causation, it might also be possible to mitigate any unintended consequences.

Evaluation Quality. Each type of validity is considered important to the overall quality of an RCT. High internal validity helps to ensure that estimated impacts were due to the intervention being studied and not to other factors such as contamination of the experiment (e.g., improper treatment delivery or incomplete or improper environmental controls, when the treatment and control groups experience different events aside from the treatment). High external validity helps to ensure that an intervention could achieve similar results for other subjects, at another time, or in a different setting. High construct validity helps to increase confidence that (1) the outcome of interest actually measures what it is being represented as measuring and (2) the program actually caused an impact in the way that was theorized or intended.

However, it is not necessarily always possible to enjoy the best of all of these worlds in an evaluation. For example, many scholars and practitioners have viewed RCTs as an evaluation design that, although potentially having high internal validity, in certain cases can lack external validity (e.g., if random assignment makes it more difficult to use typical subjects and natural or representative settings).[19] Some have seen RCTs as trading off external validity in order to achieve high internal validity, but others disagree that there is an implied tradeoff. RCTs with low internal validity cannot be used to confidently state that an intervention caused an observed impact, because the evidence they provide may be dubious. RCTs with high internal validity but low external validity may indicate that an intervention somehow made an impact for one population, but not whether the intervention can be replicated and would make an impact in a different population or setting. With regard to construct validity, there is not always consensus that a particular outcome of interest represents the "best" way to evaluate a program. In addition, establishing the mechanisms of causation (e.g., to ensure the intervention caused an impact in the theorized way) can be difficult. Complementary evaluation methods, in addition to an RCT, might be required to do so. Commentators have also suggested other criteria for assessing quality. For example, some have concluded that "[e]ven within [RCTs], quality is an elusive metric," and that in addition to internal validity, "a complete definition of quality also should take into account the trial's external validity and its statistical analysis, as well as, perhaps, its ethical aspects."[20]

Practical Capabilities and Limitations of RCTs

Claims about RCTs' practical capabilities and limitations, both in comparison with other research designs and in isolation, have at times been controversial. Although there is much regarding RCTs about which many observers agree, certain issues have at times sparked controversy. Sorting out these arguments can be challenging, however, because the terms that observers use to describe RCTs can be difficult to interpret. For example, some observers and practitioners view RCTs as

---

[19] Lawrence B. Mohr, *Impact Analysis for Program Evaluation*, p. 97.

[20] Jesse A. Berlin and Drummond Rennie, "Measuring the Quality of Trials: The Quality of Quality Scales," *JAMA*, vol. 282, Sept. 15, 1999, pp. 1083-1085.

the best way to determine a program's "effectiveness." However, it is not always clear whether these observers are using the term *effectiveness* as a synonym for impact, merit and worth, or accomplishment of specific, intended goals,[21] as illustrated later in this report in connection with education policy and recent uses of program evaluation during the federal budget process.

Apart from complications stemming from terminology, some observers appear to have advocated that government social and economic activities should be funded only if they can be "proven effective" by RCTs and certain quasi-experiments.[22] Others might dispute such emphasis on RCTs or have a different threshold for what "proven" means.

In addition, some observers appear to see RCTs and "non-RCTs" (i.e., any evaluation design other than an RCT) primarily as competitors or substitutes for each other when judging "effectiveness." For example, they might see value in RCTs and quasi-experiments for their ability to estimate an impact, but less in other evaluation methods that are not designed to estimate an impact. Observers might also see less value in other methods that do attempt to estimate an impact but that are judged to be so unreliable as to have little or no value. As noted in the next section, other observers argue observational and qualitative methods can be appropriate for estimating impacts in various circumstances.[23] At the same time, many observers have seen RCT and non-RCT designs as complements rather than substitutes in some

---

[21] A frequent definition for *effectiveness* in program evaluation usually concerns achievement of a desired and intended outcome, but does not necessarily incorporate costs, values, or (sometimes detrimental) unintended outcomes. See Jane Davidson, "Effectivness," in Sandra Mathison, ed., *Encyclopedia of Evaluation*, p. 122. Some use *effectiveness* as a synonym for *merit* and *worth*, or less concretely, "success." Many researchers use the terms *effect*, *effective*, and *effectiveness* to refer or relate to *impact*. The word *effect* is also sometimes used in the sense of something that inevitably follows an antecedent (as a cause or agent). See *Merriam-Webster's Collegiate Dictionary*, 11th ed. (Springfield, MA: Merriam-Webster, 2003), p. 397. For example, dropping a pen (a cause) results in a noise when the pen hits the floor (an effect). In health care, *effectiveness* has been defined as "the benefit (e.g., to health outcomes) of using a technology for a particular problem under general or routine conditions," whereas a related term, *efficacy*, has been defined as "the benefit of using a technology for a particular problem under ideal conditions, for example, in a laboratory setting." See National Institutes of Health, National Library of Medicine, "Glossary," available at [http://www.nlm.nih.gov/nichsr/hta101/ta101014.html].

[22] See Coalition for Evidence-Based Policy, *Bringing Evidence-Driven Progress to Education: A Recommended Strategy for the U.S. Department of Education*, Nov. 2002, p. 29, which called for "rigorous study designs" to be such a prerequisite and characterized only RCTs and certain quasi-experimental designs as rigorous. The report is available at [http://coexgov.securesites.net/admin/FormManager/filesuploading/coalitionFinRpt.pdf].

[23] For an example of what appears to be a qualitative method used to estimate an impact prospectively that appeared to influence decision making, see Michael Moss, "Pentagon Study Links Fatalities to Body Armor," *New York Times*, Jan. 7, 2006, p. A1. In response to disclosure of the study, the Senate and House Committees on Armed Services reportedly said they would hold hearings on the matter. See Michael Moss, "Pentagon Acts on Body Armor," *New York Times*, Jan. 21, 2006, p. A6; and Michael Moss, "Military Says It is Speeding Efforts to Add Side Armor," *New York Times*, Feb. 2, 2006, p. A18.

situations.[24] For example, many observers have argued that non-RCT studies, such as in-depth case studies and other observational or qualitative methods, are, among other things, (1) capable of casting doubt on an RCT's findings or causal claims by showing the RCT was contaminated or had poor design or implementation (i.e., revealing poor internal validity); (2) capable of showing a study's inferences are flawed or questionable (e.g., if the measured outcome of interest is judged to not fully reflect the program's goal(s), raising questions of construct validity); (3) essential for establishing the theory and conditions under which an intervention would be expected to make a favorable impact (increasing external validity); and (4) capable of establishing or strongly suggesting causation in certain circumstances (increasing internal validity), even if a study was not intended to estimate an impact. Nonetheless, different observers have at times seen either quantitative or qualitative methods as misguided, and the other methods as preferable or more legitimate.[25]

In light of the issues noted above, among others, several considerations about the practical capabilities and limitations of RCTs are summarized below.

RCT Capabilities. There is wide consensus that, under certain conditions, well-designed and implemented RCTs provide the most valid estimate of an intervention's average impact for a large sample of subjects, as measured on an outcome of interest.[26] This is the reason for the often stated claim by some observers, particularly in the medical field, that well-designed and implemented RCTs that are also double-blind are the "gold standard" for making a causal inference about an intervention's impact on an outcome of interest. (Usage of the term "gold standard" in fields other than medicine to describe the value of RCTs, however, has been considerably more contentious.) RCTs have been extensively used for decades in the medical arena as, usually, the third of four phases of the process for helping the Food and Drug Administration (FDA) evaluate drugs, devices, and biological products for approval, and for helping the medical community identify procedures that yield the most favorable health outcomes on selected outcomes of interest.[27] Rigorously estimating an impact is highly valued, because it provides a measurement of the

---

[24] For example, see National Research Council, Richard J. Shavelson and Lisa Towne, eds., *Scientific Research in Education* (Washington: National Academy Press, 2002), pp. 108-109; and Cesar G. Victora, Jean-Pierre Habicht, and Jennifer Bryce, "Evidence-Based Public Health: Moving Beyond Randomized Trials," *American Journal of Public Health*, vol. 94, Mar. 2004, pp. 400-405. When combined, these are often called "multiple" or "mixed methods" evaluations.

[25] See portions of Jennifer C. Greene and Gary T. Henry, "Qualitative-Quantitative Debate in Evaluation," in Sandra Mathison, ed., *Encyclopedia of Evaluation*, pp. 345-350.

[26] Because the estimated impact is an *average* across many subjects, it is possible the intervention may have affected different subjects in very different ways (e.g., some positively, some not at all, and some negatively). Some scholars have therefore advocated using methods to look at more than just an average impact.

[27] For a description of the four phases, see [http://www.clinicaltrials.gov/ct/info/glossary] and 21 C.F.R. § 312.21 (for phases 1 through 3) and § 312.85 (for phase 4).

magnitude of a program's impact on one or more outcomes that a stakeholder values and permits comparison among alternative treatments.[28]

In contrast, studies that rely on non-random assignment between treatment and control groups (e.g., quasi-experiments), or no control group at all, can be subject to certain threats to internal validity that undermine one's ability to make a causal inference when estimating an average impact for a large number of subjects.[29] For example, if subjects are not randomly assigned to treatment and control groups, there is greater risk that a prior existing difference between the two groups might be responsible for observed differences on an outcome of interest after an intervention. This threat to validity is often called *selection bias*.[30] In view of these strengths and advantages, among others, there seems to be some consensus among program evaluation theorists and practitioners in a variety of disciplines (public policy analysis, evaluation, economics, statistics, etc.) that, generally speaking, more RCTs should be performed in social science-related areas, when appropriate as part of a broad portfolio of evaluation strategies and methods.

Nonetheless, there appears to be less consensus, in a variety of disciplines, about what proportion of evaluations intended to estimate impacts should be RCTs (e.g., as opposed to quasi-experiments and other designs); what proportion of evaluations overall should be RCTs, in light of diverse evaluation needs; and the conditions under which RCTs would be most valuable, appropriate, and likely to result in valid findings. For example, many economists and other observers argue that RCTs have an important and often preferred role to play in estimating impacts, compared to quasi-experiments, due to their high potential internal validity. However, they have also argued that certain types of quasi-experimental methods (i.e., those often called *econometric* methods) are, under certain conditions for each method, capable of validly estimating impacts in ways that come reasonably close to the estimates from experimental methods. They furthermore have argued that quasi-experimental methods can provide useful information if an RCT is judged inappropriate due to factors like research needs, program circumstances, expense, timing requirements,

---

[28] Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, spring 1995, vol. 9, no. 2, pp. 63-84.

[29] See Steven Glazerman, Dan M. Levy, and David Myers, "Nonexperimental versus Experimental Estimates of Earnings Impacts," *Annals of the American Academy of Political and Social Science*, 589, Sept. 2003, pp. 63-93; Howard S. Bloom, et al., "Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?," MDRC Working Paper on Research Methodology, June 2002, available at [http://www.mdrc.org/publications/66/abstract.html]; and Thomas D. Cook, William R. Shadish Jr., and Vivian C. Wong, "Within Study Comparisons of Experiments and Non-Experiments: Can They Help Decide on Evaluation Policy?" paper presented at the French Econometric Society Meeting on Program Evaluation, Paris, France, Dec. 2005, available at [http://www.crest.fr/conference/Program_bis.htm].

[30] Nonetheless, even in an experiment there is a chance that an "unlucky" randomization of subjects could result in treatment and control groups that are not comparable.

ethical considerations, or other factors.[31] It should also be noted, however, that some academics and practitioners have perceived arguments for RCTs and other quantitatively oriented evaluation methods as attempts to exclude some kinds of qualitative research from being considered "scientific" or being funded or considered in policy-oriented research or debates.[32]

RCT Limitations. Scholars and practitioners have also qualified what they view as practical capabilities of RCTs, particularly in areas of public policy that closely relate to the social sciences. RCTs are seen as very strong in making cause-effect inferences about impacts for large samples of subjects, if they are designed and implemented well. However, RCTs are often seen as difficult to design and implement well.[33] A number of observers argue that "[t]here is a sizable divergence between the theoretical capabilities of evaluations based on random assignment and the practical results of such evaluations."[34] Some policy areas have been seen as more difficult compared to others for successfully implementing RCTs, leading some observers to disavow the "gold standard" title for RCTs, while still supporting

---

[31] For discussion of some of these issues, see Jeffrey Smith, "Evaluating Local Economic Development Policies: Theory and Practice," in Alistair Nolan and Ging Wong, *Evaluating Local Economic and Employment Development: How to Assess What Works Among Programmes and Policies* (Paris: OECD, 2004), pp. 287-332; and Robert A. Moffitt, "The Role of Randomized Field Trial in Social Science Research: A Perspective from Evaluations of Reforms of Social Welfare Programs," National Bureau of Economic Research working paper no. T0295, Oct. 2003.

[32] One example of such a critique, taken from the literature about education research, might be Elizabeth Adams St. Pierre, "'Science' Rejects Postmodernism," *Educational Researcher*, vol. 31, Nov. 2002, pp. 25-27. The term *experiment* is often associated with science. However, science is not only experimental. Science has been defined as "the observation, identification, description, experimental investigation, and theoretical explanation of phenomena" (*American Heritage Dictionary of the English Language*, 3rd ed. (Boston: Houghtn Mifflin, 1992), p. 1616). For discussion about the definition of science, see CRS Report RL32992, *The Endangered Species Act and "Sound Science"*, by Eugene H. Buck, M. Lynne Corn, and Pamela Baldwin. Some observers make a distinction — often a controversial one among scientists and scholars — between "hard science" and "soft science." The natural sciences, including physics, chemistry, and many fields of biology, have sometimes been called "hard," and the social sciences, including fields such as psychology, sociology, economics, and political science, have sometimes been called (frequently pejoratively) "soft." Some scholars have made influential critiques of this distinction. (See Thomas S. Kuhn, *The Structure of Scientific Revolutions*, 2nd ed (Chicago: University of Chicago Press, 1970).) In general, phenomena in the natural sciences have tended to be seen as easier to observe, quantify, and experiment within controlled settings, while phenomena in the social sciences have tended to be seen as more difficult to observe, quantify, and experiment within controlled settings.

[33] The question if an experiment will be implemented well has been considered a "big *if*" according to some scholars; see William M.K. Trochim, *The Research Methods Knowledge Base*, p. 191.

[34] James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, spring 1995, pp. 85-110.

increased use of RCTs as the most credible way to estimate an impact.[35] Some researchers in the medical field have argued that certain types of well-designed and implemented observational studies can yield similar results to RCTs.[36] As with all forms of study, poorly designed or implemented RCTs can yield inaccurate results. In addition, an RCT's capability to support causal inferences does not necessarily hold for determining causality in small numbers of subjects or individual cases, for which other methods are often judged more appropriate.[37] Nor do RCTs necessarily provide an advantage, compared to other evaluation research designs, in generalizing a specific intervention's ability to make an impact to a broader or different population. RCTs have therefore been seen by some observers as often having external validity limitations.[38] Frequently, RCTs rely on support from other evaluation methods for making inferences about external validity.

As discussed in more detail later in this report, RCTs can sometimes be seen as impractical, unethical, requiring too much time, or being too costly compared to other designs that also seek to assess whether a program causes favorable impacts and outcomes. There is wide consensus that RCTs are particularly well suited for answering certain types of questions, but not necessarily other questions, compared to other evaluation research designs. For example, RCTs typically do not assess how and why impacts occur, how a program might be modified to improve program results, or a program's cost-effectiveness. RCTs also typically do not provide a full picture of whether unintended consequences may have resulted from a program or indicate whether a study is using valid measures or concepts for judging a program's success (e.g., assessing a study's or a measure's construct validity). Many of these kinds of questions have been considered to be more appropriately addressed with observational or qualitative designs.

---

[35] See Thomas D. Cook and Monique R. Payne, "Objecting to the Objections to Using Random Assignment in Educational Research," in Frederick Mosteller and Robert Boruch, eds., *Evidence Matters* (Washington, DC: Brookings Institution Press, 2002), p. 174. For example, some scholars reject calling RCTs "the gold standard," because, among other things, "shortfalls often occur when implementing experiments in the field and we do not yet know enough about the robustness of designs to withstand these reality bruises." See Thomas D. Cook, William R. Shadish Jr., and Vivian C. Wong, "Within Study Comparisons of Experiments and Non-Experiments: Can They Help Decide on Evaluation Policy?," p. 31.

[36] See John Concato, Nirav Shah, and Ralph I. Horwitz, "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs," *The New England Journal of Medicine*, vol. 342, June 22, 2000, pp. 1887-1892.

[37] Lawrence B. Mohr, *The Causes of Human Behavior: Implications for Theory and Method in the Social Sciences* (Ann Arbor, MI: University of Michigan Press, 1996), pp. 9-10.

[38] See Lawrence B. Mohr, *Impact Analysis for Program Evaluation*, pp. 92-97; and National Research Council, *Scientific Research in Education*, p. 125.

RCTs in Context: Program Evaluation and Systematic Review

Concerns About Single Studies and Study Quality. In a variety of research fields, there appears to be consensus that a single study, no matter how well designed or implemented, is rarely sufficient to reliably support decision making. For example, in health care, where RCTs are often seen as the "gold standard" (for making causal inferences about impacts) and are more widely used than in any other field of study, there is strong reluctance to rely on a single RCT study. According to the Cochrane Collaboration, an international non-profit group founded in 1993 that supports the production and dissemination of RCT information about health care interventions, most single RCTs are seen as "not sufficiently robust against the effects of chance" and often having limited external validity.[39] Moreover, there have been widespread concerns about the quality of individual studies of any design. The Cochrane Collaboration has said that the amount of information about health care, including from individual RCTs, is overwhelming, but that "much of what [information] is available is of poor quality."

In seeking to address questions of how to use single studies and how to judge study quality, two major strategies that researchers have used include (1) classifying study types within "hierarchies of evidence" and (2) conducting systematic reviews.

Study Quality: A Hierarchy of Evidence? In recent years, there has been considerable debate on how to define what "quality" should mean when describing evaluations. Some observers in medicine and the social sciences hold the view that RCTs should be placed at the top of a hierarchy, in view of an RCT's potential for high internal validity in estimating an impact. Due to concerns about the varying quality of individual studies, however, some participants in health care evaluation have reoriented their approach.[40] The U.S. Preventive Services Task Force (USPSTF), an independent panel of experts in primary care and prevention that systematically reviews evidence regarding clinical preventive services, offered this assessment:

> For some years, the standard approach to evaluating the quality of individual studies was based on a hierarchical grading system of research design in which RCTs received the highest score.... The maturation of critical appraisal techniques has drawn attention to the limitations of this approach, which gives inadequate consideration to how well the study was conducted, a dimension

---

[39] Mike Clark, The Cochrane Collaboration, "Systematic Reviews and the Cochrane Collaboration," Apr. 22, 2004, available at [http://www.cochrane.org/docs/whycc.htm].

[40] Some observers have argued against a "rigid hierarchy," based on studies that found well-designed observational studies to yield similar results to RCTs, and that found RCTs to sometimes offer conflicting results. See John Concato, Nirav Shah, and Ralph I. Horwitz, "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs," *The New England Journal of Medicine*, June 22, 2000.

> known as internal validity. A well-designed cohort study may be more compelling than an inadequately powered or poorly conducted RCT.[41]

In response to these conclusions, the USPSTF listed a hierarchy of research designs, but as only one among several inputs to evaluating the quality of individual studies. That hierarchy placed designs in the following rankings, largely on the basis of a design's *potential* internal validity:

- RCTs;
- controlled trials without randomization (also called quasi-experiments, with a treatment group and a control group whose subjects were not assigned randomly);
- cohort or case-control analytic studies (observational studies in which similar groups serve as control and treatment groups);
- multiple time series with or without the intervention (uncontrolled experiments which look at the effects of an intervention on units over a significant amount of time); and
- opinions of respected authorities (based on clinical experience, descriptive studies and case reports, or reports of expert committees).[42]

To determine a study's overall quality, however, the task force also included *realized* internal validity (including design and implementation aspects) and, in addition, external validity, which the task force considered "on par" in importance with internal validity. Thus, an RCT might rank high in terms of *potential* internal validity, but it might have experienced implementation problems leading to poorly *realized* internal validity, or it might have limited external validity. In either case, or to provide assurance against chance, researchers and decision makers often wish to consider other studies, including studies other than RCTs, to inform their thinking and potential decision making. Two additional USPSTF criteria for judging quality included assessments of internal and external validity of all relevant studies for a given research question, and also the extent to which relevant studies or groups of studies linked interventions directly or indirectly to outcomes of interest. These latter efforts relate closely to systematic review.

Systematic Review in Health Care. In response to concerns about reliance on single studies and study quality, the health care field has broadly embraced *systematic review* as a method for identifying gaps in knowledge, drawing whatever

---

[41] Agency for Healthcare Research and Quality, U.S. Preventive Services Task Force, "Current Methods of the U.S. Preventive Services Task Force: A Review of the Process," p. M-21 [2001], available at [http://www.ahrq.gov/clinic/ajpmsuppl/review.pdf] (hereafter USPSTF Report). See also Earl P. Steinberg and Bryan R. Luce, "Evidence Based? Caveat Emptor!" *Health Affairs,* Jan./Feb. 2005, pp. 82-84. A cohort study has been defined as "an observational study in which outcomes in a group of patients that received an intervention are compared with outcomes in a similar group i.e., the cohort, either contemporary or historical, of patients that did not receive the intervention" (see [http://www.nlm.nih.gov/nichsr/hta101/ta101014.html].

[42] USPSTF Report, p. M-21.

conclusions are possible about interventions based on available evidence (including impact analyses), and thereby helping to inform decision making about research priorities and provision of health care to patients. *Systematic review* has been defined as "a form of structure[d] literature review that addresses a question that is formulated to be answered by analysis of evidence, and involves objective means of searching the literature, applying predetermined inclusion and exclusion criteria to this literature, critically appraising the relevant literature, and extraction and synthesis of data from evidence base to formulate findings."[43] This leads some researchers to place a systematic review as an additional category above RCTs at the top of an evidence hierarchy focused solely on potential internal validity. Some researchers also place a *meta-analysis* (of RCTs or perhaps other intervention studies) at the top of an evidence hierarchy. A meta-analysis is a type of systematic review that uses statistical methods to derive quantitative results from the analysis of multiple sources of quantitative evidence.[44]

However, for various reasons, both conducting and interpreting a systematic review can be challenging and can require caution. A systematic review, like an RCT or other evaluation, might not be comprehensive for all stakeholders, or even necessarily for a single stakeholder, in assessing their evaluation needs.[45] Systematic reviews typically focus on a specific question by looking at a specific outcome of interest, as described in relevant studies or chosen by an evaluator (e.g., in health care, outcomes like mortality, quality of life, clinical events). However, they would not necessarily focus on all important outcomes of interest. This might raise issues in the context of evaluating public policies, because a program's mission might encompass a variety of potential outcomes of interest, or be difficult to represent with one or more outcome measures. Furthermore, different stakeholders might not agree about the relative importance of varying outcome measures or might be interested in different ones. Finally, although systematic reviews typically focus much attention on concerns about internal validity of various studies, judgments about external validity, or generalizability of findings, are often left to readers to assess, based on their implicit or explicit decision, "how applicable the [systematic review's] evidence is to their particular circumstances."[46] Unfortunately, these judgments are often

---

[43] National Institutes of Health, National Library of Medicine, "Glossary," available at [http://www.nlm.nih.gov/nichsr/hta101/ta101014.html].

[44] Meta-analyses also typically incorporate some kind of qualitative decision concerning the validity of the RCTs being studied. This sometimes causes situations in which different meta-analyses of the same RCTs come to opposing conclusions about an intervention. For example, see Peter Jüni, et al., "The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis," *JAMA*, vol. 282, Sept. 15, 1999, pp. 1054-1060.

[45] The ways in which evidence is defined, identified, compiled, scrutinized, and aggregated will often affect conclusions. While there is no universally embraced "best" process for evidence review, there is considerable interest in two processes, *risk analysis*, which is increasingly applied to the regulation of environmental hazards, and *systematic review*, which is increasingly applied in the health care field.

[46] P. Alderson, S. Green, and J.P.T. Higgins, eds., *Cochrane Reviewers' Handbook 4.2.3*, Section 9.2, (updated May 2005), available at [http://www.cochrane.dk/cochrane/handbook/hbook.htm].

hindered, because many studies provide little information that might assist in assessing external validity.

Systematic Review in Social Science-Related Areas. Usage of systematic review in health care raises the question of how systematic review might be used in other contexts (e.g., when evaluating government programs of various types, which are often assessed with social science research methods). In program evaluation, systematic reviews have been performed under various names (e.g., evaluation synthesis, integrative review, research synthesis) and in different ways.[47] However, they have been much less common in social science-related areas than in health care. This might be the case, in part, because RCTs and other, non-RCT evaluations have been relatively more scarce in policy areas related to the social sciences, as compared to medicine. For example, over 250,000 RCT studies had reportedly been published in the medical literature as of 2002, but about 11,000 were known in all of the social sciences combined.[48] Other possible historical reasons for a relative lack of systematic review in the social sciences, compared with health care, might include the following: comparatively less funding devoted to evaluation; more technically challenging research settings and problems (e.g., absence of laboratory controls that can make experimental evaluations more difficult to successfully design and implement, increasing the risk that studies might result in evaluation funding being wasted); resistance to using RCTs; disagreements about appropriate ways to evaluate programs; and less interest from policy makers and institutions.

In response to such comparisons, some efforts have been undertaken to increase production of systematic reviews in social science-related areas. For example, a group of social science researchers created the Campbell Collaboration, a non-profit organization that promotes the use of systematic reviews in the social sciences. In defining "evidence," the Campbell Collaboration has focused primarily on RCTs and secondarily on quasi-experiments in order to determine impacts.[49] However, the organization's guidelines also allow implementation studies and qualitative research to be included in a systematic review.[50]

---

[47] For discussion of multiple types of systematic reviews and synthesis in program evaluation, see David S. Cordray and Robert L. Fischer, "Synthesizing Evaluation Findings," in Joseph S. Wholey, Harry P. Hatry, and Kathryn E. Newcomer, eds., *Handbook of Practical Program Evaluation*, pp. 198-231. For discussion that primarily emphasizes systematic reviews of RCTs (from the second edition of the previous book), see Robert F. Boruch and Anthony Petrosino, "Meta-Analysis, Systematic Reviews, and Research Syntheses," in Joseph S. Wholey, Harry P. Hatry, and Kathryn E. Newcomer, eds., *Handbook of Practical Program Evaluation*, 2nd ed., pp. 176-203. GAO produced a guide to "evaluation synthesis," defined as "a systematic procedure for organizing findings from several disparate evaluation studies." See U.S. General Accounting Office, *The Evaluation Synthesis*, GAO/PEMD-10.1.2, Mar. 1992, p. 6.

[48] "Try It And See," *The Economist*, Mar. 2, 2002, pp. 73-74.

[49] Robert Boruch, Haluk Soydan, and Dorothy de Moya, "The Campbell Collaboration," *Brief Treatment and Crisis Intervention*, vol. 4, no. 3, autumn 2004, p. 227.

[50] Campbell Collaboration, *Campbell Systematic Reviews: Guidelines for the Preparation of Review Protocols*, ver. 1.0 (Jan. 1, 2001), pp. 4, 6, available at (continued...)

Recent Attention to Using RCTs in Program Evaluation

The following two subsections briefly illustrate how RCTs have been a subject of attention in two contexts: (1) setting program evaluation policy in one specific policy area (education) and (2) the citation and use of individual studies, or claimed lack thereof, to justify policy and budget proposals to Congress (in this case, as a component of the George W. Bush Administration's Program Assessment Rating Tool). Because this report's purpose is limited to providing an overview of RCTs and related issues, these cases are not analyzed in detail in the report.[51] However, many of the issues identified in this report could be applied to these and other cases.

Controversy in Education Policy: A Priority for RCTs?

In January 2005, the U.S. Department of Education (ED) published a "notice of final priority" in the *Federal Register*. The notice established a department-wide "priority" for the use of specific types of program evaluation, and especially RCTs, when evaluating certain education programs.[52] Under the priority, ED asserted that RCTs were "best for determining project effectiveness," and with some exceptions, would be preferred for funding compared to other evaluation methods. If ED determined an RCT to be infeasible, a quasi-experimental design would receive priority over other designs. The ED priority has provoked controversy in the education policy area and evaluation field generally.[53]

Authority Cited for the ED Priority: NCLB and "Scientifically Based Research". The ED priority was established at the discretion of the Secretary of Education and was not required by law. However, the priority appeared in a broader context of program evaluation-related statutory provisions enacted by Congress. Specifically, ED cited the Elementary and Secondary Education Act of 1965 (ESEA), as reauthorized by the No Child Left Behind Act of 2001 (NCLB; 115 Stat. 1425; P.L. 107-110), as the statutory authority for establishing the priority.[54] In its *Federal Register* notice, ED asserted

---

[50] (...continued)
[http://www.campbellcollaboration.org/Fraguidelines.html].

[51] For related discussion, however, see CRS Report RL33246, *Reading First: Implementation Issues and Controversies*, by Gail McCallion; and CRS Report RL32663, *The Bush Administration's Program Assessment Rating Tool (PART)*, by Clinton T. Brass.

[52] The priority took effect on Feb. 24, 2005. See U.S. Department of Education, "Scientifically Based Evaluation Methods," 70 *Federal Register* 3586, Jan. 25, 2005.

[53] For example, see Yudhijit Bhattacharjee, "Can Randomized Trials Answer The Question of What Works?" *Science*, vol. 307, Mar. 25, 2005, pp. 1861-1863.

[54] For an overview of NCLB, see CRS Report RL31284, *K-12 Education: Highlights of the No Child Left Behind Act of 2001 (P.L. 107-110)*, coordinated by Wayne Riddle. For an analysis of this and related legislative history, see Margaret Eisenhart and Lisa Towne, "Contestation and Change in National Policy on 'Scientifically Based' Education Research," *Educational Researcher*, vol. 32, Oct. 2003, pp. 31-38.

> [t]he ESEA as reauthorized by the NCLB uses the term *scientifically based research* more than 100 times in the context of evaluating programs to determine what works in education or ensuring that Federal funds are used to support activities and services that work. This final priority is intended to ensure that appropriate federally funded projects are evaluated using scientifically based research.[55]

Under ESEA as reauthorized by NCLB, *scientifically based research* is defined as "research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs."[56] The statutory definition also enumerates several kinds of research that are included within the term. The first enumerated item explicitly includes research that employs either observational or experimental methods.[57] In the fourth enumerated item, the definition also includes research that is evaluated using experimental or quasi-experimental designs. Among experimental and quasi-experimental designs, the definition expresses "a preference for random-assignment experiments ..." (Section 9101(37)(B)(iv)). Thus, the statutory definition of *scientifically based research* does not appear to give higher priority to experimental designs above designs that draw on observation, except when contrasting experimental and quasi-experimental designs versus one another. Apparently in light of these definitions, ED's notice of priority also said

> [t]he definition of scientifically based research in section 9201(37) [sic] of NCLB includes other research designs in addition to the random assignment and quasi-experimental designs that are the subject of this priority. However, the Secretary considers random assignment and quasi-experimental designs to be the most rigorous methods to address the question of project effectiveness.[58]

Additional statutory provisions related to program evaluation in education, located within the Education Sciences Reform Act of 2002 (ESRA; 116 Stat. 1940; P.L. 107-279), were enacted after NCLB and a year before the ED priority was proposed. The ED priority did not cite ESRA, but ESRA's provisions privilege RCTs in some ways that appear to be related to ED's subsequent actions. ESRA established ED's Institute of Education Sciences (IES) and set forth its functions.[59] ESRA's definition for *scientifically based research standards* holds that when IES-

---

[55] U.S. Department of Education, "Scientifically Based Evaluation Methods," 70 *Federal Register* 3586 (italics in original).

[56] 115 Stat. 1425, at 1964; Section 9101(37) of NCLB; 20 U.S.C. § 7801. Many, and perhaps most, of the references to scientifically based research in NCLB refer to the research upon which instructional techniques (that grantee states and local education agencies use in federally funded programs) are to be based.

[57] The provision says the term "includes research that ... employs systematic, empirical methods that draw on observation or experiment" (Section 9101(37)(B)(i) of NCLB).

[58] The cited section should have been Section 9101(37) of NCLB. U.S. Department of Education, "Scientifically Based Evaluation Methods," p. 3586.

[59] For more on IES, see U.S. Department of Education, Institute of Education Sciences, *Biennial Report To Congress*, [2005], available at [http://www.ed.gov/about/reports/annual/ies/biennialrpt05.pdf].

funded research is intended to "mak[e] claims of causal relationships," the IES research should include only "random assignment experiments" and "other designs (to the extent such designs substantially eliminate plausible competing explanations for the obtained results)."[60] ESRA does not specify what these "other" designs are or what it means to "make claims of causal relationships." Therefore, it appears that a claim of causal relationship need not be restricted to evaluations that seek to estimate an impact. ESRA's definition for *scientifically valid education evaluation* holds that when IES's National Center for Education Evaluation and Regional Assistance conducts evaluations to estimate the impact of programs, it should "employ experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible."[61] These ESRA definitions hold for funding controlled by IES, not for the entire Education Department.

Reactions to the Priority. Originally proposed in November 2003, the ED priority generated considerable debate in the evaluation field.[62] The priority did not explicitly define the word *effectiveness*. As noted earlier in this report, *effectiveness* is a program evaluation term that has been used in multiple ways (i.e., synonym for impact, goal achievement, or merit and worth). Upon close reading, the priority's usage of the term appeared to indicate the term was probably used in most cases as a synonym for *impact*. However, the priority's use of the term *effectiveness* and the phrase *what works* appeared to be interpreted by many observers to go beyond the definition of *impact*.[63] In addition, some text in the priority appeared to be

---

[60] See 116 Stat. 1943, Section 102(18).

[61] See 116 Stat. 1943-1944, Section 102(19); and 116 Stat. 1962 and 1964-1965, Sections 171(b)(2), 172, and 173.

[62] For the original notice of proposed priority, see U.S. Department of Education, "Scientifically Based Evaluation Methods," 68 *Federal Register* 62445, Nov. 4, 2003. For contrasting views on the subject, see (1) Thomas D. Cook, "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*, vol. 24, no. 3, fall 2002, pp. 175-199 [written before the ED priority, but advocating increased use of RCTs in education policy in spite of objections]; (2) Stewart I. Donaldson and Christina A. Christie, "The 2004 Claremont Debate: Lipsey vs. Scriven: Determining Causality in Program Evaluation and Applied Research: Should Experimental Evidence Be the Gold Standard?" pp. 4-8 [summary and video of a debate about the priority between two well-known evaluation experts], available at [http://www.cgu.edu/include/SBOS_2004_Debate.pdf], online video of the discussion, along with selected transcriptions, available at [http://www.cgu.edu/pages/2668.asp]; and (3) Michael Quinn Patton, "The Debate About Randomized Controls in Evaluation: The Gold Standard Question," lecture delivered at National Institutes of Health, National Cancer Institute, Sept. 14, 2004 [criticizing the priority and advocating a different approach], online video available at [http://videocast.nih.gov/ram/nci091404.ram], presentation slides available at [http://videocast.nih.gov/ppt/nci_patton091404.ppt].

[63] See, for example, the archived e-mail discussion list ("listserv") of the American Evaluation Association, EVALTALK, entries from Nov. 4, 2003, to Jan. 4, 2004, available at [http://bama.ua.edu/archives/evaltalk.html].

interpreted to claim RCTs were best for demonstrating causation, even apart from estimating an impact.

During the proposed priority's month-long comment period, nearly 300 parties sent comments to ED. These comments were summarized and analyzed in the ED January 25, 2005, notice of final priority, which also included statements from then-Secretary of Education Rod Paige regarding where he agreed or disagreed with comments that were submitted. From ED's summary and categorization of the comments, it appears many more comments were critical of the priority than supportive.[64] ED determined that while the comments it received were substantive, the comments did not warrant changes in the priority.

Both prior to and after publication of the ED priority, many observers who supported the priority (as well as some who opposed the ED priority) agreed that more RCTs were needed in education policy, in certain circumstances, in order to estimate impacts of different educational interventions. Many also agreed that RCTs had been unjustifiably de-emphasized in the past compared to other evaluation methods, to greater or lesser extents. Furthermore, many supporters of the ED priority argued the priority was an appropriate change in the education field, because they believed more information about the impacts of educational interventions was needed to help inform practitioners and policy makers, and because they believed ED's research agenda had been previously influenced by hostility to RCTs and similar types of studies. However, many critics of the ED priority argued that RCTs have been oversold in terms of their practical capabilities; that the priority unjustifiably de-emphasized other evaluation designs in terms of their practical capabilities to contribute to understanding of causes, effects, impacts, "effectiveness," and in some cases to making claims of causal relationships (even if some of the designs are not intended to calculate impacts); and that the ED priority would detrimentally affect overall priorities for evaluation.

Implications and Related Developments. To the extent that evaluations help frame future choices, it appears the way in which the ED priority will be implemented could affect the future course of education programs and policy. At a minimum, the priority has influenced the use of hundreds of millions of research dollars controlled by ED and arguably education policy, as implemented by ED. For example, ED's department-wide strategic planning documents and activities appear to reflect the priority. The ED strategic plan established a goal that 75% of "new research and evaluation projects funded by the Department that address causal questions ... employ randomized experimental designs."[65] The question of what types of research can make causal claims, or make causal claims while substantially eliminating plausible competing explanations, has often been contentious in the

---

[64] Of the nearly 300 commenters, ED said that 29 expressed support for the priority. ED did not tabulate how many commenters opposed the priority. However, in ED's analysis of the comments, several of ED's categories of comments appeared to reflect criticisms of the priority, and the number of commenters in some of those areas ranged from 168 to 242.

[65] See U.S. Department of Education, *Strategic Plan 2002-2007*, Mar. 2002, p. 53, available at [http://www.ed.gov/about/reports/strat/plan2002-07/index.html]. This document pre-dated proposal of the ED priority by over a year.

evaluation field and the philosophy of science.[66] The ED "performance budget" accompanying the department's FY2007 budget proposal contains a department-wide objective to "encourage the use of scientifically based methods within federal education programs."[67] One performance measure is listed for that objective: "[t]he proportion of school-adopted approaches that have strong evidence of effectiveness compared to programs and interventions without such evidence."

The ED strategic plan also states that "[t]he Department will seek funding for programs that work, and will seek to reform or eliminate programs that do not."[68] The strategic plan does not define what is meant by the word "work," but the word might mean "increases student achievement," in some cases. However, if ED determined that some programs "increase student achievement," but at undesirable cost or with unintended side effects, then presumably ED would not seek funding for those programs. Thus, the word "work" might instead be intended to indicate *merit* and *worth*, in some cases.

A month after the priority was originally proposed, ED issued a prominent guidance document "to provide educational practitioners with user-friendly tools to distinguish practices supported by rigorous evidence from those that are not."[69] The ED guidance asserted that evaluation methods other than RCTs and certain quasi-experiments (1) have "no meaningful evidence" to contribute to establishing whether an intervention was "effective" and (2) cannot be considered "scientifically-rigorous evidence" or "rigorous evidence" to support using an educational practice to "improve educational and life outcomes for [children]."[70] The document appears to define "evidence" that would support these decisions to include only estimations of impact. The document also cites NCLB as calling on "educational practitioners to use

---

[66] Many arguments about this subject concern definitions of "causation" and explore what types of knowledge different methods of research are capable of discovering. Some observers give RCTs privileged place among evaluation methods in making causal claims, arguing it is theoretically the best way to avoid threats to internal validity. Others dispute that contention, arguing that RCTs are not always appropriate, many methods can make scientifically valid causal claims (e.g., citing the effort to prove the causal relationship between smoking and cancer), and RCTs often depend on the support of other methods to justify internal validity claims. Sometimes, researchers in a variety of fields arguably answer this question of "What methods are capable of making causal claims?" according to their training and preferred research techniques. One scholar called this phenomenon, when it occurs, *the law of the instrument*, under which "some preferred set of techniques will come to be identified with scientific method as such." See Abraham Kaplan, *The Conduct of Inquiry: Methodology for Behavioral Science* (Scranton, PA: Chandler, 1964), pp. 28-30.

[67] U.S. Department of Education, *Fiscal Year 2007 Performance Budget*, Feb. 2006, p. 3, available at [http://www.ed.gov/about/reports/annual/2007plan/fy07perfplan.pdf].

[68] U.S. Department of Education, *Strategic Plan 2002-2007*, p. 80.

[69] See U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance [prepared by the Coalition for Evidence-Based Policy], *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide* (Washington: U.S. Department of Education, Dec. 2003), p. iii, available at [http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html].

[70] Ibid., pp. iii, v, 11, and 17.

'scientifically-based research' to guide their decisions about which interventions to implement," but does not discuss roles many observers argue that other evaluation methods can play in complementing, bolstering, or undermining an RCT's findings.

In addition, the department established a What Works Clearinghouse (WWC) to "evaluate the strength of the evidence of effectiveness of educational interventions ... to help educators and education policymakers incorporate scientifically based research into their educational decisions."[71] In this formulation, it appears the WWC might be using the word *effectiveness* and the phrase *what works* as synonyms for "showing favorable impact on an outcome of interest." According to the WWC website, the WWC "was established in 2002 by the U.S. Department of Education's Institute of Education Sciences (IES) to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education."[72] WWC study reports state that "neither the What Works Clearinghouse (WWC) nor the U.S. Department of Education endorses any interventions."[73]

The WWC states it includes only certain evaluation designs in its databases — which happen to be only the designs that were listed in the ED priority[74] — because they "provide the strongest evidence of effects."[75] It appears the term *effects* is used here as a synonym for *impacts*. The WWC excludes other types of evaluations, because the WWC asserts they are not "outcome evaluations."[76] However, in the latest Government Accountability Office (GAO) pamphlet on definitions of major types of program evaluation, GAO defined *outcome evaluation* to also include evaluations that focus on unintended effects and that "assess program process to understand how outcomes are produced."[77] These other types of evaluations would appear to be excluded from the WWC and, perhaps, from ED determinations of "what works."[78] In addition, the ED strategic plan states under the department's

---

[71] See [http://www.whatworks.ed.gov/reviewprocess/standards.html].

[72] See [http://www.whatworks.ed.gov/faq/what_is_wwc.html].

[73] See reports listed at [http://www.whatworks.ed.gov/Products/BrowseByLatestReportsResults.asp?EvidenceRptID=03&ReportType=All].

[74] See [http://www.whatworks.ed.gov/reviewprocess/notmeetscreens.html].

[75] See [http://www.whatworks.ed.gov/reviewprocess/standards.html].

[76] See [http://www.whatworks.ed.gov/reviewprocess/notmeetscreens.html].

[77] U.S. Government Accountability Office, *Performance Measurement and Evaluation: Definitions and Relationships*, GAO-05-739SP.

[78] Strictly speaking, the phrase "what works" might be at odds with what the WWC presents. Instead, the phrase "what may have worked" might be more appropriate for what the WWC presents. In common speech, the phrase "what works" appears to assert that if a program previously "worked" in one instance, the program can be expected to "work" in other instances; that is, that the program's results are generalizable to other contexts, times, or places. However, the WWC website is careful to say in its study reports "[n]o single Study Report should be used as a basis for making policy decisions because (1) few studies are designed and implemented flawlessly and (2) all studies are tested on a limited number of

(continued...)

Objective 1.4 that every ED program will develop a "'what works' guide," to be distributed to program grantees, that "whenever possible ... will be informed by the What Works Clearinghouse."[79] Therefore, it appears that ED and IES determinations of "what works" might be drawn primarily, or only, from studies that include RCTs, certain quasi-experiments, and the exceptions allowed for in the ED priority, and not other kinds of program evaluation.

At the time of this report's writing, the WWC website presents a definition for the term *scientifically based research* that is at variance with the term's definition under NCLB.[80] As noted previously, the NCLB definition for *scientifically based research* includes observational studies and experimental studies, expressing a preference for RCTs over quasi-experiments, but not expressing a preference for RCTs over other evaluation research designs, even when questions of causation are being examined. However, the WWC website presents a different definition for the term *scientifically based research*. Specifically, the WWC website instead presents what appears to be the ESRA definition for *scientifically based research standards*, which applies only to IES-funded research. The ESRA term's definition for *scientifically based research standards* holds, as noted earlier, that when IES-funded research is intended to "mak[e] claims of causal relationships," the IES research should include only "random assignment experiments" and "other designs (to the extent such designs substantially eliminate plausible competing explanations for the obtained results)." Although ESRA does not indicate what these "other designs" could be, the WWC appears to include only the designs mentioned in the ED priority in that category. Nevertheless, the ED priority was established for the entire department. At the same time, the WWC website appears to tie WWC to NCLB's term *scientifically based research*.[81]

In effect, as demonstrated through ED policy and strategic planning documents, the ED priority appears in some respects to be extending ED's interpretation of ESRA definitions and preferences for certain types of evaluations (especially RCTs) beyond IES to the entire Education Department, notwithstanding the apparently differing definitions and preferences expressed in NCLB.

Assessing Programs in the Budget Process: The PART

The Bush Administration's Program Assessment Rating Tool (PART). In 2004, the Bush Administration elevated RCTs as the preferred way to evaluate federal executive branch "programs" under portions of its Program

---

[78] (...continued)
participants, using a limited number of outcomes, at a limited number of times, so generalizing from one study to any context is very difficult." See, for example, [http://www.whatworks.ed.gov/PDF/Ridgway_2002_Brief_Study_Report.pdf].

[79] U.S. Department of Education, *Strategic Plan 2002-2007*, p. 16.

[80] See [http://www.whatworks.ed.gov/faq/what_research.html]. The website attributed the definition to the IES.

[81] See [http://www.whatworks.ed.gov/faq/wwc_nclb_readfirst.html].

Assessment Rating Tool (PART) initiative.[82] The PART is a set of questionnaires that the Office of Management and Budget (OMB) developed in 2002 and annually revised thereafter to determine the "overall effectiveness" of programs included in the President's annual budget proposal. Although OMB did not provide an explicit definition for the term *overall effectiveness*, OMB and the Bush Administration appeared to use the term as at least a partial synonym for *merit* and *worth*.[83] The Administration has used the PART, with some controversy, to justify budget proposals and the proposed elimination or reduction of many programs.

OMB first presented the PART to Congress for the FY2004 budget cycle, assessing programs that represented approximately 20% of the federal budget. For succeeding budget cycles, OMB said that cumulative 20% increments of federal programs would be assessed with the PART, in addition to some reassessments of programs previously "PARTed." The Administration subsequently released PART ratings for selected programs along with the President's FY2004, FY2005, FY2006, and FY2007 budget proposals. An additional round of ratings is planned to be released with the President's FY2008 budget proposal, with the final year assessing all remaining executive branch spending and programs. Thereafter, all programs would presumably be assessed or reassessed each year.

Depending on how a PART questionnaire is filled in and evaluated for a program, it will produce a single numerical score in percentage terms between 0% and 100%. This figure determines the program's *overall effectiveness* rating. Four ratings are possible based on the score: "effective" (a score of 85% to 100%), "moderately effective" (70% to 84%), "adequate" (50% to 69%), and "ineffective" (0% to 49%). OMB characterizes these ratings as qualitative rather than quantitative. A different designation was created, regardless of PART score, for programs that OMB decided "do not have acceptable performance measures or have not yet

---

[82] For an overview and analysis of the PART, see CRS Report RL32663, *The Bush Administration's Program Assessment Rating Tool (PART)*, by Clinton T. Brass. The definitions that OMB used for the term *program* when conducting PART assessments have been criticized by some observers. OMB's definitions of programs to be assessed, which were typically determined by a budgetary perspective, sometimes aggregated several activities into a single "PART program" or disaggregated a group of activities into several "PART programs."

[83] The Administration's usage appears to go beyond a typical program evaluation definition of *effectiveness* (which refers to achievement of a goal), because under the PART, OMB makes judgments about a program's mission and performance goals (e.g., whether they "address a specific and existing problem, interest, or need"; whether they are "unclear" because of "multiple and overlapping objectives"; and whether they are "effectively targeted," "meaningful," and "ambitious") and a program's cost-effectiveness. The Administration has also, for example, referred to the PART as proving a program's "worth" (U.S. Office of Management and Budget, *Budget of the United States Government, Fiscal Year 2004*, p. 51). Nevertheless, the Administration appears to distinguish between the term *overall effectiveness* and its views about "priorities" for a budget proposal (e.g., according to its FY2006 listing, these priorities were defense, homeland security, "economic opportunity," and "fostering compassion") and its views about which programs' missions have an "appropriate federal role." See U.S. Office of Management and Budget, *Major Savings and Reforms in the President's 2006 Budget* (Washington: GPO, 2005), p. 4.

collected performance data."[84] The designation was called "results not demonstrated." GAO has said "[i]t is important for users of the PART information to interpret the 'results not demonstrated' designation as 'unknown effectiveness' rather than as meaning the program is 'ineffective.'"[85] GAO also found that disagreements between OMB and agencies on appropriate performance measures for certain programs helped lead to the 'results not demonstrated' designation being given to these programs for purposes of the PART.[86]

Use of the PART. While the Administration has called the PART a management tool,[87] it has also said the PART's overall purpose is to "lay the groundwork" for funding decisions.[88] Furthermore, the Administration also provided guidance to agencies that a program's PART questionnaire should include "good" performance goals that "provide information that helps make budget decisions," but need not include "performance goals to improve the management of the program."[89] In the context of the President's FY2006 budget proposal, which called for terminating or substantially reducing 154 discretionary programs,[90] an OMB official reportedly said "we have to focus more resources on what works, and the PART is the primary tool to make that judgment."[91]

Of 99 discretionary programs proposed by the Administration for termination for FY2006, 48, or nearly half of all proposed terminations, were in the Department of Education.[92] In the Administration's justification document, RCT evaluations

---

[84] U.S. Office of Management and Budget, *Budget of the United States Government, Fiscal Year 2005, Analytical Perspectives* (Washington: GPO, 2004), p. 10.

[85] See U.S. General Accounting Office, *Performance Budgeting: Observations on the Use of OMB's Program Assessment Rating Tool for the Fiscal Year 2004 Budget*, p. 25.

[86] Ibid.

[87] U.S. Office of Management and Budget, *Budget of the United States Government, Fiscal Year 2005, Analytical Perspectives* (Washington: GPO, 2004), p. 9.

[88] U.S. Office of Management and Budget, *Budget of the United States Government, Fiscal Year 2004, Analytical Perspectives* (Washington: GPO, 2003), p 9.

[89] U.S. Office of Management and Budget, "Performance Measurement Challenges and Strategies," June 18, 2003, p. 4, available at [http://www.whitehouse.gov/omb/part/challenges_strategies.pdf].

[90] Some, but not all, of these programs had been assessed with the PART. See U.S. Office of Management and Budget, *Major Savings and Reforms in the President's 2006 Budget*.

[91] Amelia Gruber, "The Big Squeeze," *Government Executive*, Feb. 2005, p. 48.

[92] Of the 48, ED explicitly cited 7 in the department's summary budget justification as receiving the PART rating "results not demonstrated" (i.e., unknown results) and 3 as receiving the PART rating "ineffective." See U.S. Department of Education, Fiscal Year 2006 Budget Summary and Background Information, Feb. 7, 2005, pp. 72-79, available at [http://www.ed.gov/about/overview/budget/budget06/summary/06summary.pdf], or at [http://www.ed.gov/about/overview/budget/budget06/summary/edlite-section3.html].

were explicitly cited as supporting termination of the ED Even Start program,[93] and the Administration further cited lack of performance information or lack of "rigorous evaluations" to support termination of many of the other ED programs. On December 23, 2005, the White House reportedly sent to "reporters and surrogates" a listing of "terminations accepted in whole or in part" by Congress, which some media posted on their websites.[94] Neither the White House nor OMB posted the document on their websites. According to the OMB document, 17 of the 49 proposed ED terminations were accepted by Congress in whole or in part, including a 56% cut in Even Start. For the President's FY2007 budget proposal, the Administration proposed "141 programs that should be terminated or significantly reduced in size," government-wide.[95] Of these, the Administration proposed to terminate 42 programs within the Department of Education (nearly 30% of the government-wide total proposed for termination or major reduction), "including many that the PART has shown to be ineffective or unable to demonstrate results."[96] The department said in a press release the 42 programs were "proven ineffective."[97] ED also stated in budget justification documents that a requested termination was "consistent with the Department's goal to eliminate support for programs that show limited or no evidence of effectiveness."[98]

RCTs and the PART. For purposes of the PART, in 2004, OMB elevated RCTs as the preferred way to evaluate a program's "effectiveness" with release of a document entitled *What Constitutes Strong Evidence of a Program's Effectiveness?*[99]

---

[93] U.S. Office of Management and Budget, *Major Savings and Reforms in the President's 2006 Budget*, p. 25. For analysis of the justification, see CRS Report RL33071, *Even Start: Funding Controversy*, by Gail McCallion.

[94] See U.S. Office of Management and Budget, *Major Savings in the 2006 Budget* (Washington: Dec. 22, 2005), available at [http://www.cq.com/budgettracker.do], newsletter for Jan. 9, 2006, and at [http://hotlineblog.nationaljournal.com/archives/2005/12/index.html], listed under web log entry entitled "WH Touts Budget Successes." See also White House Office of Communications, "Fiscal Year 2006: Keeping the Commitment to Restrain Spending," press release, Dec. 22, 2005, available at [http://www.whitehouse.gov/news/releases/2005/12/20051222-15.html].

[95] U.S. Office of Management and Budget, *Fiscal Year 2007 Budget of the U.S. Government* (Washington: GPO, 2005), p. 2.

[96] Ibid., p. 83.

[97] See U.S. Department of Education, "Fiscal Year 2007 Budget Request Advances NCLB Implementation and Pinpoints Competitiveness: President's Budget Supports New Math and Science Instruction and High School Reform; Targets Resources and Eliminates 42 Programs Proven Ineffective, Saving $3.5 Billion," press release, Feb. 6, 2006, available at [http://www.ed.gov/news/pressreleases/2006/02/02062006.html]; and U.S. Department of Education, *Fiscal Year 2007 Budget Summary and Background Information*, Feb. 6, 2006, available at [http://www.ed.gov/about/overview/budget/budget07/summary/index.html].

[98] For example, see U.S. Department of Education, *Fiscal Year 2007 Justifications of Appropriation Estimates to the Congress, Volume II* (Washington: Feb. 2006), p. M-30.

[99] U.S. Office of Management and Budget, "What Constitutes Strong Evidence of a Program's Effectiveness?" undated white paper [2004], p. 1 (hereafter cited as the OMB

(continued...)

This document, intended to provide guidance to agencies on appropriate evaluations, particularly highlighted RCTs as "best" for evaluating "effectiveness." The document was apparently written with the assistance of the Coalition for Evidence-Based Policy (CEBP), an organization sponsored by the Council for Excellence in Government.[100] According to CEBP's website, the PART and OMB's annual guidance to agencies for the PART (in this case, for the FY2006 budget) were "revised," as a result of collaboration between CEBP and OMB, "to endorse randomized controlled trials as the preferred method for measuring program effectiveness, and well-matched quasi-experimental studies as a possible alternative when randomized trials are not feasible."[101] Nevertheless, the OMB document also stated that

> RCTs are not suitable for every program and generally can be employed under very specific circumstances. Therefore, agencies often will need to consider alternative evaluation methodologies. In addition, even where it is not possible to demonstrate impact, use of evaluation to assist in the management of programs is extremely important.[102]

In OMB's *What Constitutes* document, OMB used the term *effectiveness* in at least two senses: (1) arguably as a partial synonym for *merit* or *worth* (consistent with the PART's usage) or (2) referring to demonstrating *impact*, which OMB defined as "the outcome of a program, which otherwise would not have occurred without the program intervention." These concepts do not necessarily represent the same thing, because merit or worth can be judged by many other factors in addition to impact on

---

[99] (...continued)
What Constitutes document), available at [http://www.whitehouse.gov/omb/part/2004_program_eval.pdf].

[100] CEBP's mission is "to promote government policymaking based on rigorous evidence of program effectiveness." CEBP has particularly emphasized the advantages of RCTs for that purpose and the disadvantages of using other evaluation research designs. For more on the group's purpose and agenda, see [http://www.excelgov.org/index.php?keyword=a432fbc34d71c7].

[101] See [http://coexgov.securesites.net/index.php?keyword=a432fbc34d71c7] for CEBP's description of its role in working with OMB. See also Amelia Gruber, "The Big Squeeze," *Government Executive*, p. 53. For the most recent iteration of OMB's guidance for the PART, see U.S. Office of Management and Budget, "Guidance for Completing the Program Assessment Rating Tool (PART)," Mar. 2005, available at [http://www.whitehouse.gov/omb/part/fy2005/2005_guidance.doc]. The guidance was intended to be used for the President's FY2007 budget proposal, but appeared to be labeled FY2005 to indicate the fiscal year in which it was released. OMB's PART guidance from previous years is available in electronic form or hard copy from this report's first-listed author.

[102] OMB, *What Constitutes* document, p. 1. In cases when "it is not possible to use RCTs to evaluate program impact," the document directs agencies to "consult with internal or external program evaluation experts, as appropriate, and OMB to identify other suitable evaluation methodologies to demonstrate a program's impact," but provides little explicit guidance in that regard (Ibid., p. 3).

a specific outcome of interest.[103] In some cases, the *What Constitutes* document is clear which definition of *effectiveness* is being used (e.g., when using the term *impact*). However, other instances are less clear or could potentially be interpreted by agencies as treating the definitions *impact* and *merit* and *worth* equivalently. A sampling is reprinted below.

- "The [PART] was developed to assess the effectiveness of federal programs and help inform management actions, budget requests, and legislative proposals directed at achieving results" (p. 1).
- "The revised PART guidance this year underscores the need for agencies to think about the most appropriate type of evaluation to demonstrate the effectiveness of their programs. As such, the guidance points to the [RCT] as an example of the best type of evaluation to demonstrate actual program impact" (p. 1).
- "Few evaluation methods can be used to measure a program's effectiveness, where effectiveness is understood to mean the impact of the program" (p. 1).
- "The most significant aspect of program effectiveness is *impact* — the outcome of the program, which otherwise would not have occurred without the program intervention" (p. 2, italics in original).
- "[Non-experimental direct analysis studies] often lack rigor and may lead to false conclusions if used to measure program effectiveness, and therefore, should be used in limited situations and only when necessary. Such methods may have use for examining *how* or *why* a program is effective, or for providing information that is useful for program management" (p. 3, italics in original).
- "Well-designed and implemented RCTs are considered the gold standard for evaluating an intervention's effectiveness across many diverse fields of human inquiry, such as medicine, welfare and employment, psychology, and education" (p. 4).[104]

---

[103] For example, in the view of stakeholders, additional determinants of merit or worth could, in addition to impact on a particular outcome of interest, include timeliness; quality; cost; efficiency; replicability of program implementation or results at other times or in other contexts; presence of unintended outcomes; impact on outcomes of interest that may not have been chosen by OMB or an agency to assess the program; achievement of mission-related goals that are not necessarily encompassed by quantitative measures of impact; program or environmental changes that make past estimations of impact obsolete; the comparative merit and worth of policy alternatives; and values (e.g., a stakeholder's normative views on the program's importance, goals, or means of implementation).

[104] This "gold standard" claim has been contentious in some fields (including medicine), and especially in education. In OMB's footnote supporting the "gold standard" assertion, OMB included literature supporting the assertion (e.g., an article from the *Journal of Economic Perspectives* calling for increased use of RCTs), but omitted or ignored literature arguing against the assertion (e.g., the adjacent article from the same issue of the *Journal of Economic Perspectives*, which argued the case for RCTs was overstated compared to alternative designs, except for some aspects of internal validity). See, respectively, Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, Spring 1995, pp. 63-84; and James J. Heckman and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, spring 1995, pp. 85-110.

OMB's annual guidance to agencies for the FY2006 budget's PART similarly elevated RCTs as the preferred way to assess *effectiveness* and *impact*. The guidance called *impact* "the most significant aspect of program effectiveness," called RCTs "generally the highest quality, unbiased evaluation to demonstrate the actual impact of the program," and further asserted that "[t]he most definitive data supporting a program's overall effectiveness would be from [an RCT], when appropriate and feasible."[105] The guidance also stated that RCTs are not suitable or feasible for every program, because "[f]ederal programs vary so dramatically." In such situations, the guidance suggested well-designed quasi-experimental studies as another way to assess impact, and other types of evaluations to "help address *how* or *why* a program is effective (or ineffective)" (italics in original).

For the FY2007 PART, however, the tenor of OMB's statements in its guidance about RCTs may have changed to some degree. OMB's discussion in the revised guidance largely mirrored that of the FY2006 guidance, but eliminated the description of RCTs as "the highest quality, unbiased evaluation to demonstrate the actual impact" and replaced it with language calling RCTs "particularly well suited to measuring impacts."[106] When RCTs are judged by OMB and agencies to not be feasible or suitable, the guidance exhorts agencies and OMB to consult with in-house or external evaluation experts, and directs them to "supplemental guidance" in the *What Constitutes* document.

Judging "Success". In practice, it appears that OMB's judgments regarding quality and suitability of evaluation designs have sometimes trumped agency judgments and therefore determined what evaluation methods are to be used for the PART, in spite of disagreements between OMB and agencies.[107] Disputes about the proper ways to judge program success have also emerged. For example, in the context of controversy over the Administration's "ineffective" PART rating of the Community Development Block Grant (CDBG) program and FY2006 budget proposal to significantly cut and consolidate 18 community and economic development programs, an article quoting OMB's Deputy Director for Management Clay Johnson III suggested that political views, or at least different views about program goals, might play a role:

> Johnson acknowledges that CDBG fails the [PART] test in part because the administration is applying a new definition of success. "We believe the goal of housing programs is not just to build houses, but the economic development that comes with them. So those are the results we want to focus on," Johnson said.

---

[105] U.S. Office of Management and Budget, "Instructions for the Program Assessment Rating Tool," [Mar. 22, 2004], pp. 24, 47, available at [http://www.whitehouse.gov/omb/part/2006_part_guidance.pdf].

[106] U.S. Office of Management and Budget, "Guidance for Completing the Program Assessment Rating Tool (PART)," Mar. 2005, p. 28.

[107] U.S. Government Accountability Office, *Program Evaluation: OMB's PART Reviews Increased Agencies' Attention to Improving Evidence of Program Results*, GAO-06-67, Oct. 2005, pp. 22-25; and *Performance Budgeting: PART Focuses Attention on Program Performance, but More Can Be Done to Engage Congress*, GAO-06-28, Oct. 2005, pp. 26-27.

> "You can say we are imposing our political views on people, or our favored views of the housing world or the CDBG world on people. Well, guilty as charged. It's important to focus on outcomes, not outputs."[108]

Another example might be what OMB has called for purposes of the PART the Vocational Education State Grants program, within ED, which the Administration proposed for termination for the FY2007 budget.[109] This program is the largest budgetary component relating to the Carl D. Perkins Vocational and Technical Education Act of 1998, commonly called Perkins III. For the FY2007 budget's PART, the Administration deemed the program "ineffective," the lowest PART rating.[110] The Administration justified termination by citing a particular study, the National Assessment of Vocational Education (NAVE),[111] stating that the NAVE "found no evidence that high school vocational courses themselves contribute to academic achievement or college enrollment."[112] This perspective appears to consider academic achievement and college enrollment to be the goals of federally supported vocational education. However, the June 2004 NAVE also found that "[t]he short- and medium-term benefits of vocational education are most clear when it comes to its longstanding measure of success — earnings," citing research findings that "students earned almost 2 percent more for each extra high school vocational course they took," extending to varying degrees "to the large group of high school graduates who enroll in postsecondary education and training, to both economically and

---

[108] Paul Singer, "By the Horns," *National Journal*, Mar. 26, 2005, p. 904. Observers, scholars, and even statutes often have different definitions of terms like outcome, output, impact, and a variety of other program evaluation terms, and different views about the importance or applicability of those terms. Under GPRA, *outcome measure* "refers to an assessment of the results of a program activity compared to its intended purpose," and *output measure* "refers to the tabulation, calculation, or recording of activity or effort and can be expressed in a quantitative or qualitative manner" (31 U.S.C. § 1115).

[109] U.S. Office of Management and Budget, *Major Savings and Reforms in the President's 2007 Budget* (Washington: Feb. 2006), p. 27, available at [http://www.whitehouse.gov/omb/budget/fy2007/pdf/savings.pdf].

[110] See [http://www.whitehouse.gov/omb/expectmore/summary.10000212.2005.html].

[111] Marsha Silverberg, et al., U.S. Department of Education, Office of the Under Secretary, Policy and Program Studies Service, *National Assessment of Vocational Education: Final Report to Congress*, June 2004, available at [http://www.ed.gov/rschstat/eval/sectech/nave/index.html]. The report presented "a synthesis of evidence on the implementation and outcomes of vocational education and the 1998 Perkins Act" (see p. 16). According to the study's authors, the NAVE was conducted on an independent basis, as called for by law, and did not necessarily reflect official views or policies of ED (p. xvi). For discussion of Perkins III and also the NAVE study, see CRS Report RL31747, *The Carl D. Perkins Vocational and Technical Education Act of 1998: Background and Implementation*, by Rebecca R. Skinner and Richard N. Apling.

[112] U.S. Office of Management and Budget, *Major Savings and Reforms in the President's 2007 Budget*, p. 27. See also U.S. Department of Education, *Fiscal Year 2007 Budget Summary and Background Information*, p. 51.

educationally disadvantaged students, to those with disabilities, and to both men and women."[113] The study authors further observed:

> Perkins III and its legislative predecessors have largely focused on improving the prospects for students who take vocational education in high school, a group that has historically been considered low achieving and noncollege-bound. However, students who participate most intensively in vocational progams ... are actually quite diverse... . ... The vocational courses most high school students take improve their later earnings but have no effect on other outcomes that have become central to the mission of secondary education — such as improving academic achievement or college transitions... . Whether the program as currently supported by federal legislation is judged successful depends on which outcomes are most important to policymakers.[114]

The PART assessment of the program released with the FY2007 budget was originally released with the FY2004 budget in February 2003, reflecting data available in 2002, and had not been updated to reflect the June 2004 NAVE. The PART assessment's worksheet provided only brief reference to evidence regarding the impact of vocational education on earnings.[115] Neither the Administration's justification document for terminating the program nor ED's *FY2007 Budget Summary* mentioned earnings benefits of federally supported vocational education.

Disputes about an existing program's proper goals can raise questions about the construct validity of studies that purport to evaluate the program, regardless of whether the study is an RCT, the PART, or another type of evaluation. How should one measure "success"? What outcome of interest — or outcomes — are the most important ones? Proponents of the PART have viewed favorably the initiative's effort to raise program performance to a more salient place in budget deliberations. Many observers have also seen favorably the PART's transparency, with detailed justifications for the Administration's views available on the Web for consideration by Congress and the public. However, critics and other observers have said the Administration's criteria for evaluating programs sometimes deviated from the programs' purposes, as determined by Congress, or that the Administration and OMB substituted their views about appropriate program goals and measures over those developed by agencies under the statutory framework established by GPRA, which explicitly provides for stakeholder views, including those of Congress.[116]

---

[113] Marsha Silverberg, et al., U.S. Department of Education, Office of the Under Secretary, Policy and Program Studies Service, *National Assessment of Vocational Education: Final Report to Congress*, pp. xix-xx, 18, and 266.

[114] Ibid., pp. xix, 265.

[115] See [http://www.whitehouse.gov/omb/budget/fy2004/pma/vocationaleducation.xls], question 4.5, for the FY2004 budget's PART assessment. The worksheet is available in Microsoft Excel format. For the FY2007 budget PART assessment, see [http://www.whitehouse.gov/omb/expectmore/detail.10000212.2005.html].

[116] For example, see Aimee Curl, "Supporters Call Even Start a Case Study in Faulty Program Assessments," *Federal Times*, Jan. 9, 2006, p. 4; and U.S. General Accounting Office, *Performance Budgeting: Observations on the Use of OMB's Program Assessment*

(continued...)

Potential Issues for Congress

The previous section of this report illustrated how RCTs have been subjects of prominent attention in two contexts: (1) setting program evaluation policy and (2) citation and use of individual studies, or lack thereof, to justify policy and budget proposals to Congress. In these and potentially other cases, a focus on RCTs might raise multiple issues for Congress.[117] Some relate specifically to RCT studies, including an RCT's structural requirements and constraints. Other issues relate to program evaluation generally, and therefore to RCTs. A number of these issues are identified and analyzed below.

Issues When Directing or Scrutinizing RCTs

If Congress wants to focus on RCTs in the context of program evaluation policy (e.g., developing legislation for, or conducting oversight over, a program, agency, or the entire government; or prospectively deciding whether to fund specific evaluations), Congress might consider a number of issues related to the parameters of these studies and prospective risks to their internal and external validity. In addition, issues could arise if Congress wants to interpret or scrutinize individual RCT studies (e.g., when they are presented to Congress in the budget or authorization processes).

Considering Study Parameters. When making program evaluation policy, Congress might opt to focus on some key parameters of studies, including random assignment, the cost of an RCT, the length of time that an RCT would take before producing findings, and privacy or ethical considerations.

Random Assignment. As discussed earlier, the central attribute of an RCT is the random assignment of subjects to treatment and control groups, which helps a researcher to make inferences that a particular intervention was responsible for an impact and to estimate that impact with reliable statistical tools.[118] In some programs, however, it may not be feasible or cost-effective to randomly assign units to an intervention group and a control group. For example, it is not possible to conduct an RCT on whether a policy regulating the release of chlorofluorocarbons into the environment contributes to overall global warming, because there is only one planet earth to study. Thus, if Congress is considering whether to require certain

---

[116] (...continued)
Rating Tool for the Fiscal Year 2004 Budget, GAO-04-174, Jan. 2004, pp. 6-7.

[117] For example, Congress might consider legislation to set program evaluation policy in additional areas, or might conduct oversight over program evaluation policies that are already in statute. In addition, Congress might consider directing, funding, interpreting, or scrutinizing specific evaluations during its lawmaking and oversight work. There is little doubt that participants and stakeholders in the policy process will bring program evaluations to Congress to try to influence the thinking and decision making of Members and committees.

[118] For discussion regarding ethical dimensions to random assignment, see the section below entitled "Privacy, Ethics, and Study Oversight."

types of evaluation for a program, or if Congress is asked by an actor in the policy process to change funding for a program due to lack of experimental evidence of program impact, it might be important to question whether random assignment is possible or practical. On the other hand, if a program would appear to allow for an evaluation using random assignment but none is planned, or an alternative is planned (e.g., a quasi-experiment), it might be important to consider whether an RCT evaluation would be more appropriate. For example, with respect to the cases discussed in this report, what programs assessed by the PART are practically suitable for evaluation by an RCT? Under the ED priority, why should quasi-experiments also receive a priority for funding, in addition to RCTs?

Cost of RCTs. If Congress considers setting evaluation policy or directing specific studies for an agency or program, it might take the likely costs of an evaluation method or study into consideration, to weigh against the study's potential benefits.[119] Large scale multi-site RCTs are typically expensive, especially when the units being studied are not individual persons but rather organizations such as schools or jails. Large multi-site RCTs of K-12 education funded by ED have reportedly cost $10 million to $50 million.[120] In many cases, this level of funding is not available for the study of a federal program, due in part to tight budgetary constraints and the fact that program evaluations frequently must be paid for out of a program's budget. Smaller scale RCTs featuring the random assignment of 100-200 individuals might be relatively inexpensive, reportedly costing from $300,000 to $700,000.[121] Even this cost, however, can often be as much as a studied program's funding level. Quasi-experiments are frequently, but not always, less expensive, but also might bring a different set of potential benefits and risks compared to an RCT.

Even if considerable funding is available for evaluations, pursuing a particular evaluation will leave fewer resources for other evaluation needs that might be judged important. The opportunity cost of pursuing certain evaluations rather than others (i.e., cost associated with opportunities that are foregone by not putting resources to their highest value use) can be high. Nevertheless, because the potential benefits and costs of evaluations can vary widely depending on an agency's or program's portfolio of evaluation needs, weighing these considerations can be difficult. In addition, priorities might change depending on developments in an agency or policy environment. In light of these and other considerations, Congress might weigh the possible benefits of a study against the likely cost in view of the broader portfolio of evaluation needs, mindful of the risk that a study might be poorly designed or implemented or suffer from contamination that reduces confidence in the study's findings. Evaluation designs that do not offer the theoretical advantages of RCTs

---

[119] For example, potential benefits could be reallocation of funds if the program is determined to be a failure and not appropriate to be modified or "fixed," or improved accomplishment of the mission if ways are found to improve the program.

[120] OMB *What Constitutes* document, pp. 11, 18.

[121] Coalition for Evidence-Based Policy, *Bringing Evidence-Driven Progress To Crime and Substance-Abuse Policy: A Recommended Federal Strategy*, p. 13, available at [http://www.excelgov.org/usermedia/images/uploads/PDFs/Final_report_-_Evidence-based_crime_&_subs_abuse_policy.pdf].

regarding internal validity might, or might not, be worthwhile alternatives to RCTs, depending on Congress's evaluation objectives in specific situations.

Length of Time to Yield Findings. RCTs might take a long period of time to yield findings that can inform thinking and policy decisions. For example, an RCT that aims to estimate whether a certain aftercare program reduces the recidivism of juvenile offenders must follow the program's graduates, and the control group, over a multi-year time span. This elongated time window might be problematic if policy decisions need to be made expeditiously. However, the length of time necessary to conduct an RCT (and perhaps other complementary evaluations) might justify waiting to make a policy decision until more evidence becomes available. Furthermore, programs or the external environment might have changed by the time the "old" program was evaluated. How should the possible benefits of evaluations be weighed against risks of evaluation information becoming dated or obsolete? What are the implications for the types and extent of evaluation research to be conducted? Congress could be called upon to consider these issues if it chose to establish program evaluation policy or directing and funding specific evaluations.

Privacy, Ethics, and Study Oversight. When considering whether to direct the use of RCTs or other evaluations of public programs and policies, Congress might consider whether the agencies conducting them are charged, or should be charged, with ethical duties to protect RCT study participants' privacy, access to programs, and opportunity to give informed consent. In addition, if Congress determined that an agency should be charged with the ethical duties, Congress might consider requiring oversight — by law, regulation, or institutional action — to help ensure that the duties were fulfilled.[122]

Privacy issues may arise in an RCT if, for example, a program evaluation captures information about individual citizens or clients that could be used inside the government for a purpose other than for which it was collected, or released to the public in some form. What, if any, safeguards should be required of those collecting the information? Several privacy protections are currently legally required for agency-conducted program evaluations and RCTs by the Privacy Act (5 U.S.C. § 552a).[123] Among other things, the act sets conditions concerning the disclosure of personally identifiable information, prescribes requirements for the accounting of certain disclosures of the information, requires agencies to specify their authority and purposes for collecting personally identifiable information from an individual, and

---

[122] This section of the report discusses existing executive-branch-wide statutory and regulatory provisions that might provide certain kinds of protection to study participants. It does not examine provisions that might be program or agency specific. For discussion of the rights of human subjects in program evaluations and guidance for the program evaluation field, see Joint Committee on Standards for Educational Program Evaluation, *The Program Evaluation Standards*, 2nd ed. (Thousand Oaks, CA: Sage, 1994). See also the website of the Office of Human Subjects Research, National Institutes of Health, regarding regulations and ethical guidelines, available at [http://ohsr.od.nih.gov/guidelines/guidelines.html].

[123] The act does not specifically mention program evaluations, but regulates how executive branch agencies may maintain, collect, use, and disseminate information about individuals.

provides civil and criminal enforcement arrangements.[124] If a program evaluation is funded or directed by an agency but conducted by a non-federal entity (e.g., if a non-federal entity creates and maintains records about program evaluation participants), the Privacy Act's coverage is often stated in contracts.

The issue of access to government programs might arise in an RCT if, for example, the RCT were designed with a control group that was to be denied access to a program as a part of the evaluation. Although access in some cases might not be required by law, its denial raises the question of whether the benefits of testing a program outweigh the burden of denying access to certain prospective subjects. Would it be appropriate to design RCTs for entitlement programs, which guarantee services to clients, with such control groups?

The issue of informed consent might arise in an RCT if it were deemed appropriate to enroll only willing participants. Although informed consent would not always be required by law for many RCTs and other types of program evaluation,[125] its use would guarantee that persons who participate in RCTs fully understand and agree to their participation. Would the benefits of this outweigh burdens of the time and money that must be spent to achieve such a goal?

If Congress chose to ensure that some or all of the above individual protections were implemented in RCTs and other forms of program evaluation, it might consider requiring some form of protection-specific oversight. Although probably not required by law for RCTs, the institutional review of a proposed research trial's protections for human participants is a common requirement for a great deal of research.[126] These reviews generally require that researchers obtain the approval of an institutional board prior to beginning the research. The board checks to make

---

[124] See CRS Report RL30795, *General Management Laws: A Compendium*, coordinated by Clinton T. Brass, entry for "Privacy Act" in section I.F. of the report, by Harold C. Relyea. The Privacy Act's implications for federal program evaluation should be distinguished from the Privacy Rule (45 C.F.R. 164), which established a set of national standards for the protection of individually identifiable health information. The Privacy Rule restricts the actions of "covered entities" which are, generally speaking, health care plans and providers. For more information on the Privacy Rule, see CRS Report RL32909, *Federal Protection for Human Research Subjects: An Analysis of the Common Rule and Its Interactions with FDA Regulations and the HIPAA Privacy Rule*, by Erin D. Williams, **Appendix A**.

[125] The Common Rule (45 C.F.R. 46, Subpart A) generally requires the informed consent of each participant in federally funded research. However, agency program evaluations are generally exempt from the rule (see 45 C.F.R. 46.101(b)(5)). The exempted types of research include research designed to study, evaluate, or otherwise examine: public benefit or service programs; procedures for obtaining benefits or services under those programs; possible changes in or alternatives to those programs or procedures; or possible changes in methods or levels of payment for benefits or services under those programs. See CRS Report RL32909, *Federal Protection for Human Research Subjects: An Analysis of the Common Rule and Its Interactions with FDA Regulations and the HIPAA Privacy Rule*, by Erin D. Williams.

[126] Ibid. The Common Rule generally requires oversight by an institutional body for federally funded research on human subjects. However, as noted above, agency program evaluations are exempt from the rule.

certain that each proposal includes adequate protections for participants' privacy and ensures that the plan to obtain participants' informed consent is sufficient, among other things. Though this type of review may be time consuming and add additional costs to a research project, it can prove beneficial not only by protecting the study's participants, but also by incidentally improving a study's design. Congress and agencies have also instituted other oversight mechanisms for program evaluations, including competitions for grants and peer review of grant applications.

Scrutinizing, or Prospectively Assessing, Studies' Internal and External Validity. Whether making program evaluation policy or scrutinizing studies, Congress might also focus on issues of study interpretation and implementation (e.g., deciding whether to direct or fund evaluations in light of potential contamination risks and projected external validity, or judging how much confidence to put in the internal and external validity of a study presented during the budget or reauthorization processes).

Contamination and Internal Validity. Actors in the policy process will not necessarily advertise any defects in the studies they present to influence Congress. With that in mind, when actors in the policy process present Members or committees of Congress with program evaluations intended to influence a policy decision, Congress might consider the evaluation's realized, and not merely theoretical, internal validity. In addition, when Congress sets program evaluation policy for a given policy area, it might be possible to prospectively consider the probability of a study's successful design, implementation, and corresponding internal validity. A major threat to the internal validity of RCTs and quasi-experiments has been called *contamination*.[127] It should be noted that other designs that attempt to estimate impacts are subject to additional threats, including selection bias, as noted previously.

To avoid contamination, well-designed and implemented RCTs ideally insulate the treatment and control groups from events that might affect one group during the study differently from the other group in a way that will affect outcomes. Doing so is intended to ensure the only systematic difference in the experience of the two groups is whether or not they received the intended treatment. RCTs also ideally ensure that the intended treatment was administered properly. Because social science research usually does not occur under tightly controlled laboratory conditions, however, it is often difficult to insulate a study from, or control for, unforeseen variables that might systematically affect the treatment and control groups differently.

---

[127] See Lawrence B. Mohr, *Impact Analysis for Program Evaluation*, pp. 80-84. Observers have identified many types of contamination and sometimes use different terms to describe them (e.g., spillovers, disruptions). Another major threat is attrition of subjects that differs between the treatment and control groups.

Three examples might help illustrate the threat of contamination.[128] First, if an experiment is not double-blinded, subjects in the treatment group might be aware of their inclusion in a special program. If this awareness results in psychological effects that are not considered part of the treatment and subjects behave differently, the study's results might be contaminated. Similarly, some subjects in a control group might learn they are not in the treatment group and either decide to avail themselves on their own initiative of alternative treatments that are not associated with the intervention, or resent or undermine the treatment being given to the other group. Second, if a program to curb crime in a city were evaluated with an RCT, certain districts might be the units of analysis. Perpetrators of crimes might move from experimental districts to control districts, contaminating findings for both groups. Third, with regard to treatment delivery, it might be difficult in some studies to ensure that the intended treatment is delivered properly for all subjects or all sites. If some subjects in the control group get the treatment, for example, or if the intended treatment is not delivered properly, inferences about the intended intervention's impact might be contaminated.

In light of these considerations, several questions might be of concern. When Congress is presented with program evaluation findings, for example, what confidence should Congress have that contamination did not degrade a study's internal validity? Does the study adequately address these risks? Also, when setting program evaluation policy, how much confidence should Congress have that studies in certain policy areas or contexts will be able to avoid contamination? What is the track record in a given area? What are the implications for how Congress and agencies should allocate scarce evaluation resources and structure an agency's portfolio of evaluations?

Generalizability (External Validity). When Congress is presented with a program evaluation to justify a policy position, the program that was evaluated presumably operated at a specific time and place, and under specific conditions. Without further analysis to gauge an evaluation's external validity, however, it will not always be clear whether the intervention itself or the study's findings can be generalized to other circumstances (e.g., future conditions, other subjects), as noted earlier in this report.

If generalizability were of concern, Congress might in the first place consider whether the intervention itself (i.e., as it was actually implemented) is replicable at a different time or place. If the intervention was highly customized to a particular time or locale, for example, an evaluation's findings might not be generalizable

---

[128] Some illustrations here are drawn from Ibid. Researchers can attempt to design studies ahead of time to avoid contamination. They can also make efforts to avoid and monitor contamination during a study's implementation. For an example regarding study design, in an RCT of high-school curricula, study researchers might randomize schools instead of individual students (i.e., schools are the randomized units of analysis). If one of the schools being studied in the intervention group happened to be in a neighborhood that underwent an isolated crime wave, the crime wave could have adversely influenced the findings for that group. However, randomization of schools would neutralize the threat of contamination, because one would expect schools in the control group to have a comparable probability of experiencing a crime wave.

elsewhere unless the intervention were fully replicated in all important respects.[129] Alternatively, if an actual intervention were not clearly documented (e.g., how it operated, whom it served), it might be unclear how to replicate the intervention. In that case, if the intervention were evaluated, it might not be reasonable to expect the evaluation's findings would be repeated elsewhere.[130]

With regard to generalizability of findings (as opposed to the intervention itself), if a single-site RCT finds that an intervention had an impact for a group of subjects in one instance, it might not necessarily follow that it will have a similar impact for other subjects, times, and circumstances. Congress might therefore look for multiple impact analysis studies on the subject. Large scale, multi-site RCTs are often considered more generalizable than single-site RCTs, because they estimate an intervention's impact in a potentially wider array of geographic locations and populations. However, multi-site RCTs are also typically more expensive and difficult to successfully implement compared to single-site RCTs. In addition, complementary studies are often considered necessary to make assessments of generalizability. Complementary studies might include observational or qualitative evaluations, which can potentially be used to better understand an intervention's mechanism of causation, potential unintended consequences, and conditionality (i.e., the conditions that are required for the intervention to work as intended).

If these considerations were a source of concern, Congress might scrutinize an evaluation's findings for external validity to other times, conditions, and subjects. Alternatively, if Congress is setting evaluation policy for an agency or program, or is directing that specific studies occur, Congress might provide direction or guidance regarding the evaluation methods that might be necessary for establishing a program's generalizability to other circumstances.

Issues When Directing or Scrutinizing Program Evaluations

Congress might also consider issues that apply to program evaluation generally. Because an RCT is one of many types of program evaluation, these issues might be important when Congress (1) considers making program evaluation policy for specific agencies or programs (e.g., what types of evaluations to direct or fund) or (2) scrutinizes individual evaluations, including RCTs, when making policy decisions and conducting oversight. For example, if Congress considers legislation that provides for program evaluation in one or more policy areas, to what extent should RCTs be the focus of these evaluation policies? To what extent should other

---

[129] However, full replication might or might not be appropriate at other times or places, if conditions in another environment dictated otherwise.

[130] If the intervention resulted in a favorable impact but was not clearly documented, it might not be clear under what conditions the program could be expected to result in similar findings, if Congress were considering expanding the program. Conversely, if the intervention did not result in a favorable impact and was not clearly documented, it might not be clear whether the program would do better if implemented under different conditions. For example, it might be possible that a favorable impact could be achieved if the intervention were modified or targeted at different subjects.

evaluation methods be the focus? Should multiple methods fit into a broader evaluation framework? The potential issues discussed below raise these and further questions.

What Types of Evaluations are Necessary? Given the nature of a policy area and the diverse needs of stakeholders — including agency program managers, the President, citizens, and notably Congress — what different types of evaluations should be directed, funded, and considered? This report began with a subsection titled "Key Questions about Government Programs and Policies," which outlined a number of questions that stakeholders often want to be informed about. In response, a wide array of program evaluation techniques have been developed.

Certain evaluation types often address, or help address, multiple kinds of questions. Indeed, many evaluation types are considered complementary to each other. At the same time, the different types of evaluations bring their own sets of practical capabilities and limitations, and experts and practitioners sometimes disagree on the nature and importance of these capabilities and limitations. Because only finite resources are available for evaluation activities and staff, it can be challenging and controversial to determine the appropriate methods to be used to help answer certain stakeholder questions. Furthermore, when scrutinizing evaluations, it can be challenging to discern potential gaps in the perspectives provided by an actor in the policy process. For example, it can be challenging to discern clues that might suggest other complementary evaluation types are needed.

Given these considerations, many issues might be of congressional concern. For example, in Congress's view, how should the executive branch be pursuing evaluations under the PART initiative, which particularly highlighted RCTs? How should Congress oversee and respond to the Administration's efforts to achieve "budget and performance integration" through use of the PART? In the education policy arena, to what extent is ED appropriately implementing the program evaluation aspects of NCLB and ESRA? To what extent is the ED priority, which elevated RCTs and quasi-experiments above other evaluation types, consistent with congressional intent? How is ED implementing the priority? As Congress considers legislation in other policy areas, should Congress provide direction or guidance regarding program evaluation policy or methods? If so, what kinds of studies might agencies, Congress, and outside stakeholders need? When actors in the policy process present evaluations to influence Congress, are the actors presenting the full story, or are there gaps in the presentation? Are the evaluations they present capturing the key questions that need to be answered? Any of these multiple questions might be ripe for attention.

What Definitions and Assumptions are Being Used? As noted earlier, many actors in the policy-making process use program evaluations to help justify their policy recommendations and to attempt to persuade Congress to make decisions consistent with their policy objectives. Therefore, many observers have considered it important that policy makers, including Members and committees of Congress, be informed consumers of evaluation information. Unfortunately, however, the vocabulary of program evaluation can sometimes be confusing. For example, many observers and practitioners use the same terms, but with differing definitions for

those terms. When someone says a program is "effective," in what sense is the term being used? As noted previously, the term *effectiveness* might refer to (1) a program's overall merit or worth, (2) the extent to which a program is accomplishing its goals, (3) the program's impact on a particular outcome of interest, or (4) an ambiguous mix of all three prior definitions. These are very different concepts. In any of these senses, determining whether a program is "effective" can hinge upon an observer's views and assumptions about the program's mission, objectives, appropriate outcome(s) of interest, and progress. Thus, should these definitional aspects of program evaluations be of concern, Congress might scrutinize them closely.

Furthermore, assumptions that are implicit in an evaluation might go unstated. Some stakeholders or evaluators might implicitly argue that their preferred way of evaluating a program is "best" and that other methods are comparatively inappropriate or less appropriate. However, there is not always consensus among well-respected experts regarding when certain methods are best or most appropriate. Should these matters be of concern to Congress when they occur, Congress might investigate a number of questions. For example, if one method is claimed to be "best" compared to others, what are the stated and unstated reasons for that opinion? What would other stakeholders and evaluators say? To what extent, if any, might opinions of appropriateness be due to an underlying agenda (e.g., to support policy views) or self-interest (e.g., to get funding for a program or a type of evaluation)? These questions might also be applied to the cases discussed in this report. For example, in justifying the ED priority for RCTs, which claimed RCTs are "best" for determining "effectiveness," which definition for *effectiveness* is ED using? Is the ED definition consistent with how ED intends to use RCTs? Should ED employ additional or alternative types of evaluation for these intended uses of RCTs? In its strategic planning, budgeting, and operations, is ED using multiple definitions? What definition is being used for *overall effectiveness*, for purposes of the PART?

How Should Congress Use Evaluation Information When Considering and Making Policy? When Congress is presented with evaluation information in the policy process by various actors (e.g., lobbyists, experts, think tanks, academics, or agencies, among others), how should Congress use the information and findings? There is widespread consensus that program evaluations can help policy makers gain insights into policy problems and make better-informed decisions regarding ways to improve government performance, transparency, accountability, and efficiency. Nonetheless, the use of evaluation in the policy-making process can be controversial.

For example, some advocates of *performance-based budgeting* and *evidence-based policy* have argued that a program's future funding or existence should be "based" on its "performance" or on "evidence" of its "effectiveness." The terms *performance-based budgeting* and *evidence-based policy*, however, do not have consensus definitions, because different actors typically have, among other things, different definitions for what constitutes "performance" and "evidence," conceptions of what it means to "base" decision making on performance or evidence, views about

whether a decision should be "based" on past performance or evidence, and views about what other factors should legitimately help drive decision making.[131]

Another fundamental question might be of concern to policy makers and stakeholders. What role should evaluation of *past* events play in forming *future* strategies and plans? A prominent argument in favor of using evaluations to shape future strategies is that past performance (by a person, program, agency, etc.) is usually the best predictor of future performance. However, a prominent counter-argument is that focusing primarily on the past can be compared to "driving a car using only the rearview mirror." Although evaluation of past performance is widely considered helpful for informing thinking and decisions, the process of strategic decision making has been found in the social science and management literatures to be legitimately driven by many more factors. These have included, among others, basic and applied research, forecasting and scenario planning, risk assessment, professional judgment from individual and group experience, theoretical extrapolation, intuition (especially when information is incomplete, consensus interpretations of information are lacking, the future is uncertain, or synthesis is necessary), and values. In addition, in spite of efforts to make decisions as rational as possible in the face of uncertainty and limited information, a wide body of social science has found that there are practical limits to rationality in decision making.[132] Should these considerations be of concern to Congress when considering policy questions or conducting oversight, several questions might be asked. For example, in an RCT context, when considering or scrutinizing RCT studies, how should these studies be used by Congress to inform thinking and decision making? How should they be used by agencies and OMB? Are agencies and OMB using them in appropriate ways? What other factors can or should be considered?

---

[131] Past efforts to use performance information, evidence, and analysis to drive policy decision making (e.g., planning-programming-budgeting systems in the 1960s and zero-based budgeting in the 1970s) were similarly subject to uncertainties relating to whether, how, and to what extent evidence or analysis can be used to drive or influence complex decisions. These decisions are usually made about priorities, policy changes, and resource allocation in the face of tradeoffs, uncertainty, and limited information — all of which are subject to diverse views and conflicting values. In response to these kinds of issues, some observers have proposed dropping "-based" from a term's name and replacing it with "-informed" (e.g., *performance-informed budgeting*). The term *performance-based budgeting* appears to have its post-World War II roots in the work of the First Hoover Commission. For discussion, see CRS Report RL32164, *Performance Management and Budgeting in the Federal Government: Brief History and Recent Developments*, by Virginia A. McMurtry. The term *evidence-based policy* appears to have its roots in *evidence-based practice* (a term typically concerning behavioral health disciplines such as psychiatry and social work) and ultimately *evidence-based medicine*. For discussion of the latter two terms, see Richard N. Rosenthal, "Overview of Evidence-Based Practice," in Albert R. Roberts and Kenneth R. Yeager, eds., *Evidence-Based Practice Manual: Research and Outcome Measures in Health and Human Services* (New York: Oxford University Press, 2004), pp. 20-29.

[132] This has been described as "bounded rationality." See Herbert A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, 3rd ed. (New York: Free Press, 1976), pp. 240-247.

How Much Confidence Should One Have in a Study in Order to Inform One's Thinking and Decisions? If Congress is presented in a decision-making situation with a single program evaluation, is the evaluation enough to have high confidence in the findings? Program evaluations can provide helpful insight into policy problems and the manner and extent to which federal policies address those problems. Unfortunately, however, one or more program evaluations do not always produce information that is comprehensive, accurate, credible, or unbiased.

For example, a program evaluation can be designed to answer a certain question, but the way someone frames the original evaluation question can influence how a program is ultimately portrayed. Most prominently, this can be the case when setting criteria for "success," such as a program's goals or the preferred outcomes of interest. However, actors in the policy process often have varying views on how to judge a program's or policy's success.[133] It is not always clear, therefore, that a study's research question will be viewed by most observers as covering what should have been covered to validly or comprehensively evaluate a program.

In addition, a program evaluation will not always necessarily be well designed or implemented. In such cases, a study might produce results that are flawed or inaccurate. Even in the best case — if an evaluation is appropriate for the research question being studied, is well designed and implemented, and there is widespread consensus on how to judge "success" — it is still possible that random chance or unforeseen events might result in an evaluation that produces information that is inaccurate or flawed. For example, it is possible that an evaluation might provide a "false positive" or "false negative" result (e.g., the study finds the program successful when actually it was not, or finds a program unsuccessful when it actually was successful). If this situation were the case, the study findings might not reveal it. The subject of data or information quality is also oftentimes a subject of concern when conducting or interpreting program evaluations.[134]

How might Congress cope with these possibilities? How confident must a Member or committee be in evaluation information, including from RCTs, in order

---

[133] For example, the Internal Revenue Service has multiple goals that are implicit in its mission statement, including service, enforcement, and fairness. Different stakeholders might disagree regarding which of these implicit goals is the most important. Furthermore, there might be implicit tradeoffs among the goals that prevent simultaneous maximization of all three. For more on IRS's mission and goals, see U.S. Department of the Treasury, Internal Revenue Service, *IRS Strategic Plan, 2005-2009* (Washington: 2004), available at [http://www.irs.gov/pub/irs-utl/strategic_plan_05-09.pdf].

[134] Certain kinds of information disseminated by a federal agency, often including program evaluations, might be covered by statutory provisions informally called the "Information Quality Act" (Section 515 of the FY2001 Treasury and General Government Appropriations Act, 114 Stat. 2763A-153). Under OMB guidelines, agencies are required to issue their own guidelines to ensure and maximize the quality, objectivity, utility, and integrity of certain kinds of disseminated information. The information quality guidelines of the Department of Health and Human Services (HHS) cover program evaluations, among other things, within their scope (see [http://aspe.hhs.gov/infoquality/Guidelines/part1.shtml]). See also CRS Report RL32532, *The Information Quality Act: OMB's Guidance and Initial Implementation*, by Curtis W. Copeland and Michael Simpson.

to use the information to inform thinking and conclusions about a policy? In response, social science researchers have recommended that consumers of evaluation information be aware of the practical capabilities and limitations of various program evaluation methods and also scrutinize a study's claims of internal, external, and construct validity. They have also suggested looking for multiple studies and, if available, systematic reviews. Other observers have suggested using the resources of GAO or other congressional support agencies to help interpret or validate conclusions and scrutinizing these matters through hearings and oversight. Finally, Congress might consider whether federal agencies have sufficient capacity and independence to conduct, interpret, and objectively present program evaluations to Congress.

Do Agencies Have Capacity and Independence to Properly Conduct, Interpret, and Objectively Present Program Evaluations? At times, Congress and other actors have expressed concern over the capacity of agencies to adequately perform certain tasks, including management functions that range from procurement to financial management. One management function that has been frequently cited as a topic of concern is program evaluation. Over a long period of time, GAO has "found limited (and diminishing) resources spent on ... program evaluation" and "reason to be concerned about the capacity of federal agencies to produce evaluations of their programs' effectiveness."[135] Even with recent emphasis on program evaluation under GPRA and the Bush Administration's PART, it is unclear the extent to which agencies have capacity to properly conduct, interpret, or use program evaluations.[136] Many, if not all, of the issues discussed in this report could apply equally to organizations and decision makers within federal agencies.

Should program evaluation capacity in federal agencies be seen as a topic of concern, several questions might be considered by Congress. Given the complex issues and debates involved in the production, interpretation, and use of program evaluations — as well as complex debates about the appropriateness of different evaluation types in certain circumstances — do agencies have capacity to use evaluation information to soundly inform strategic and operational decisions? In addition, do agencies have the capacity to make objective, methodologically sound presentations and interpretations of evaluation information to Congress, including information from RCTs?

---

[135] U.S. General Accounting Office, *Program Evaluation: Agencies Challenged by New Demand for Information on Program Results*, GAO/GGD-98-53, Apr. 1998, p. 1; and *Performance Budgeting: Opportunities and Challenges*, GAO-02-1106T, Sept. 2002, p. 16. See also U.S. Government Accountability Office, *Program Evaluation: OMB's PART Reviews Increased Agencies' Attention to Improving Evidence of Program Results*, GAO-06-67, pp. 15-16, 28.

[136] Evaluations of federal programs are also conducted by outside contractors. Nevertheless, because program evaluations are undertaken in these cases under the oversight of agencies and presumably interpreted by agency leaders and analytical staffs (i.e., because agencies are consumers of evaluation information), the agencies might need analytical competency in several types of program evaluation.

Furthermore, do agency program evaluation offices and personnel have the necessary independence from politics (e.g., partisan or institutional) and self-interest to, without undue hindrance, raise potentially uncomfortable issues and surface objective, valid, and reliable findings for consideration by policy makers, including Congress?[137] Should they have this kind of independence?

Finally, and more broadly, Congress might consider issues of evaluation capacity that go beyond federal programs. In the past, Congress has established agencies that focus on evaluation issues in entire policy areas. For example, in 1989, Congress established a new agency within the Department of Health and Human Services to serve as a focal point in health care research. Congress reauthorized the agency, now called the Agency for Healthcare Research and Quality (AHRQ), in 1999.[138] Rather than focus only on evaluating federal programs, AHRQ's statutory mission is to conduct and support research in all aspects of health care, synthesis and dissemination of available scientific evidence for use by multiple stakeholders, and initiatives to advance health care quality.[139] As noted previously in this report, Congress also established IES within ED in 2002. IES has a multi-part mission "to provide national leadership in expanding fundamental knowledge and understanding in education from early childhood though post-secondary study" for many stakeholders, providing them with reliable information about "the condition and progress of education in the United States...," "educational practices that support learning and improve academic achievement and access to educational opportunities for all students," and "the effectiveness of Federal and other education programs" (116 Stat. 1944).

---

[137] According to OMB guidance, the PART requires evaluations to be "independent" (conducted by "non-biased parties with no conflict of interest"). OMB interpreted this guidance to not allow program evaluations to be conducted by programs themselves for purposes of the PART. Instead, OMB allows contracted-out evaluations by third parties to count for the PART and said that evaluations by inspectors general and agency program evaluation offices "might" also be considered independent. See U.S. Office of Management and Budget, *Guidance for Completing the Program Assessment Rating Tool*, Mar. 2005, pp. 4-5, available at OMB's website ([http://www.whitehouse.gov/omb/part/index.html]) under the link "Instructions for Completing the 2005 PART," at [http://www.whitehouse.gov/omb/part/fy2005/2005_guidance.pdf]. For discussion of disagreements between agencies and OMB about PART independence requirements, see U.S. Government Accountability Office, *Program Evaluation: OMB's PART Reviews Increased Agencies' Attention to Improving Evidence of Program Results*, pp. 25-26. It should be noted that OMB and the PART itself would not necessarily be considered independent under these criteria, because, among other things, the Administration and OMB sometimes established the goals and performance measures by which programs would be evaluated under the PART.

[138] See 42 U.S.C. § 299(b), as amended by P.L. 106-129, The Healthcare Research and Quality Act of 1999 (113 Stat. 1653).

[139] AHRQ's website says the agency's main functions are to "sponsor and conduct research that provides evidence-based information on health care outcomes; quality; and cost, use, and access" to help "health care decisionmakers — patients and clinicians, health system leaders, purchasers, and policymakers — make more informed decisions and improve the quality of health care services." For more about the agency's mission, customers, and goals, see [http://www.ahrq.gov/about/profile.htm].

Although there are differences in the missions of these agencies, one aspect that arguably makes them similar is the scope of their research and evaluation work. Specifically, their focus goes beyond federal programs to instead encompass research and evaluations throughout an entire policy area such as "education" or "health care," whether interventions are delivered by the federal government or another entity. Should Congress view program evaluation capacity as an issue for an entire policy area, Congress could move to consider whether establishment of a policy research and evaluation agency might be warranted.

Appendix A: Glossary of Selected Terms and Concepts

The Vocabulary of Program Evaluation

Unfortunately for consumers of evaluation information, the vocabulary of program evaluation can sometimes be technical and difficult. The field is multi-disciplinary and some concepts are complex. Sometimes it is not always clear in what sense a term is being used and whether the term is being used appropriately. Technical experts and actors in the policy process sometimes use the same terms for different concepts or use different terms for the same concept. Nonetheless, in program evaluations, understanding the meanings of and distinctions between key terms can make a significant difference in how to interpret study findings and limitations and in how to scrutinize evaluations to see if they are being represented objectively and forthrightly.

This appendix draws on the report to briefly define and, if necessary, explain several recurring terms and to briefly identify several definitions for the same term, as appropriate. However, the definitions provided below are illustrative only and do not necessarily indicate what an evaluation's author or what an actor in the policy process intends to communicate. More definitive assessments typically must be made on a case-by-case basis. The footnotes in this report provide written resources that can help with understanding evaluations and the terms they employ, and CRS analysts can provide additional assistance or refer readers to other resources. In each entry, a term that is included elsewhere in the glossary is written in *italics* the first time it is used.

Selected Terms and Concepts

Construct validity. In practice, there are several definitions of this term: (1) in measuring outcomes, the extent to which a study actually evaluates what it is being represented as evaluating (e.g., does the study's *outcome of interest* actually measure "student achievement"?); and (2) in relation to a *program*, the extent to which the actual program reflects one's ideas and theories of how the program is supposed to operate, and the causal mechanism through which it is supposed to achieve outcomes.

Contamination. In an *RCT*, something aside from the intended treatment that might affect the *treatment group* or *control group* differently from the other group in a way that will affect observed outcomes. For example, *RCTs* should ideally insulate the treatment and control groups from contaminating events in order to ensure that the only difference in the experience of the two groups is whether or not they received the intended treatment. In addition, RCTs should ideally ensure that the treatment was administered properly; otherwise, the treatment might be considered contaminated.

Control group. In an *RCT*, a group of subjects chosen by *random assignment* that is comparable to the *treatment group* but that does not experience the *program* being studied.

Effect. Depending on usage, something that inevitably follows an antecedent (as a cause or agent); for example, the act of dropping a pen — a cause — is closely followed by a noise — an effect — when the pen strikes the floor. Also used sometimes as a synonym for *impact*.

Effective (effectiveness). A term with multiple possible definitions. In practice, it is used as a synonym for *impact*, *merit* and *worth*, or accomplishment of specific, intended goals.

Evaluation. An applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs, value, *merit*, *worth*, significance, or quality of a program, product, person, policy, proposal, or plan.

External validity. The extent to which an intervention being studied can be (1) applied to and replicated in other settings, times, or groups of subjects and (2) expected to deliver a similar *impact* on an *outcome of interest*. The terms generalizability, replicability, and repeatability are sometimes used as synonyms for external validity. Some usage of this term refers only to one of the two aspects noted here.

Government Performance and Results Act (GPRA) of 1993. A federal law that requires most executive branch agencies to develop five-year strategic plans, annual performance plans (including goals and performance indicators, among other things), and annual program performance reports. In GPRA's legislative history, it was contemplated that not all forms of *program evaluation* and measurement would necessarily be quantifiable because of the large diversity of federal government activities.

Impact. An estimated measurement of how a *program* intervention affected the *outcome of interest* for a large group of subjects, on average, compared to what would have happened without the intervention. For example, if the unemployment rate in a geographic area would have been 6% without an intervention, but was estimated to be 5% because of the intervention, the impact would be a 1% reduction in the unemployment rate (i.e., 6% minus 5% equals an impact of 1%), or, alternatively, a 16.7% reduction in the unemployment rate, if one characterizes the impact as a proportion of the prior unemployment rate. Depending on the chosen outcome of interest, the average impact across all subjects usually reflects the weighted average of the subjects who experienced favorable impacts, subjects who did not experience a change, and others who experienced unfavorable impacts. Some theorists and practitioners use the term *effect* as a synonym for impact.

Internal validity. In an *RCT*, the confidence with which one can state that the *impact* found or inferred by a study was caused by the intervention being studied.

Merit. The overall intrinsic value of a *program* to individuals. This term is usually paired with the term *worth*.

Meta-analysis. A type of *systematic review* that uses statistical methods to derive quantitative results from the analysis of multiple sources of quantitative evidence.

Observational design. A term that has been used in different ways, but that often refers to empirical and *qualitative* evaluations of many types that are intended to help explain cause-and-effect relationships but do not attempt to approximate an *RCT*.

Outcome of interest. Something, oftentimes a public policy goal, that one or more stakeholders care about (e.g., unemployment rate, which many actors might like to be lower). There can be many potential outcomes of interest related to a *program*. Actors in the policy process will not necessarily agree which outcomes are important. Outcomes of government programs need not always be quantitative (e.g., sending humans safely to the moon and back to earth).

Performance measurement (performance measure). A term that can mean many things but is usually considered to be different from *program evaluation*. Typically, the term refers to ongoing and periodic monitoring and reporting of *program* operations or accomplishments (e.g., progress toward quantitative goals) and sometimes also statistical information related to, but not necessarily influenceable by, a program. Occasional synonyms for "measure" are indicator, metric, and target. Sometimes the word "performance" is dropped, especially when stakeholders believe the measure does not necessarily indicate whether the program itself caused changes in favorable or unfavorable directions.

Program. A government policy, activity, project, initiative, law, tax provision, function, or set thereof, that someone might wish to evaluate. In *program evaluation*, synonyms for program include treatment and intervention.

Program evaluation. Under the *Government Performance and Results Act (GPRA) of 1993*, "an assessment, through objective measurement and systematic analysis, of the manner and extent to which Federal programs achieve intended objectives." Program evaluation has been seen as (1) informing conclusions at particular points in time and also (2) a cumulative process over time of forming conclusions, as more *evaluation* information is collected and interpreted. Practitioners and theorists categorize different types of program evaluation in several ways. The varying types are sometimes referred to generically as designs or methods. Typical synonyms for this term include evaluation and study.

Qualitative evaluation. A wide variety of *evaluation* types that judge the *effectiveness* of a *program* (e.g., whether it accomplishes its goals) by, among other things, conducting open-ended interviews, directly observing program implementation and outcomes, reviewing documents, and constructing case studies.

Quasi-experimental design. A type of *evaluation* that attempts to estimate a treatment's *impact* on an *outcome of interest* for a group of subjects but, in contrast with *RCTs*, does not have *random assignment* to *treatment* and *control groups*. Some quasi-experimental designs are controlled studies (i.e, with a control group and

at least one treatment group), but others lack a control group. Some quasi-experiments do not measure the outcome of interest before the treatment takes place. Some observers and practitioners consider quasi-experiments to be a form of *observational design*, but others put them in their own category.

Random assignment. The process of assigning subjects into a *control group* and one or more *treatment groups* by random chance.

Random selection. The process of drawing a sample by random chance from a larger population (e.g., to undertake a survey that is intended to be representative of a broader population). This term is different from, and sometimes confused with, *random assignment*.

Randomized controlled trial (RCT). In its basic form, an *evaluation* design that uses *random assignment* to assign some subjects to a *treatment group* and also to a *control group*. The treatment group participates in the *program* being evaluated and the control group does not. After the treatment group experiences the intervention, an RCT attempts to compare what happens to the two groups, as measured by the resulting difference between the two groups on the *outcome of interest*, in order to estimate the program's *impact*. The terms randomized field trial (RFT), random assignment design, experimental design, random experiment, and social experiment are sometimes used as synonyms for RCT, and vice versa. Use of the word "field" in this context is often intended to imply that an *evaluation* is being conducted in a more naturalistic setting instead of a laboratory or other artificial environment. Double-blind studies are those in which neither the subjects nor the researchers know which group gets the treatment. Single-blind studies are those in which the subjects do not know they are getting the treatment being investigated.

Statistical significance. In the context of an *RCT*, a finding of statistical significance is typically interpreted as a level of confidence (usually expressed as a probability, e.g., 95%, which is also referred to as "significance at the .05 level") that an estimated *impact* is not merely the result of random variation. Assuming the RCT suffered from no defects, this finding would indicate that at least some of the measured impact may with substantial confidence (e.g., 95% confidence) be attributed to the treatment as a cause. Stated another way, significance at the .05 level indicates that there is a 1 in 20 chance that the observed difference could have occurred by chance, if the program actually had no impact. However, simply because an estimated impact is found to be statistically significant does not necessarily mean the impact is large or important.

Systematic review. A form of structured literature review that addresses a question that is formulated to be answered by analysis of evidence, and involves objective means of searching the literature, applying predetermined inclusion and exclusion criteria to this literature, critically appraising the relevant literature, and extraction and synthesis of data from the evidence base to formulate findings. Although systematic reviews typically focus much attention on concerns about *internal validity* of various studies, judgments about *external validity*, or generalizability of findings, are often left to readers to assess, based on their implicit or explicit decision how applicable the systematic review's evidence is to their

particular circumstances. In *program evaluation*, systematic reviews have been performed under various names (e.g., evaluation synthesis, integrative review, research synthesis), in different ways, and usually in decentralized fashion. Some systematic reviews focus on *RCTs* (and might include *quasi-experiments*), and others include disparate types of studies.

Treatment group. In an *RCT*, the group of subjects chosen by *random assignment* that experiences or participates in a *program*; also sometimes called an experimental or intervention group.

Unit of analysis. In an *RCT*, the subjects of the study who are *randomly assigned* to one or more *treatment groups* and also a *control group*. Subjects are typically individual persons but sometimes might be things or organizations like schools, hospitals, or police stations.

Validity. See entries for *internal validity*, *external validity*, and *construct validity*.

Worth. The overall extrinsic value of a *program* to society. This term is usually paired with the term *merit*.