What Constitutes Strong Evidence of a Program's Effectiveness?

The Program Assessment Rating Tool (PART) was developed to assess the effectiveness of federal programs and help inform management actions, budget requests, and legislative proposals directed at achieving results. The PART examines various factors that contribute to the effectiveness of a program and requires that conclusions be explained and substantiated with evidence. The PART assesses if and how program evaluation is used to inform program planning and to corroborate program results.

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 randomized controlled trial (RCT) as an example of the best type of evaluation to demonstrate actual program impact. Yet, RCTs are not suitable for every program and generally can be employed only 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.

Few evaluation methods can be used to measure a program's effectiveness, where effectiveness is understood to mean the impact of the program. Some of the most commonly used methodologies used to demonstrate such impact fall into the categories of Experimental or Randomized Controlled Trials, Direct Controlled Trials, Quasi-experimental Studies, and Non-Experimental Studies (Direct or Indirect). The following is intended to help agencies choose the right methodology for evaluations of program impact. As such, the following points are covered:

(i) How is program evaluation addressed in the PART?

(ii) What are the most common ways to evaluate program performance?

(iii) What sorts of tests provide strong evidence of a program's effectiveness?

(iv) The Application of RCTs: where they are / are not possible.

I. How is program evaluation addressed in the PART?

The PART includes two questions specifically related to program evaluation. An evaluation may also provide evidence for many of the questions in section 3, which assesses program management.

Question 2.6 asks whether there "[a]re independent evaluations of sufficient scope and quality conducted on a regular basis or as needed to support program improvements and evaluate effectiveness and relevance to the problem, interest, or need." The purpose of question 2.6 is to ensure that the program (or agency) conducts non-biased evaluations on a regular or as-needed basis to fill gaps in performance information. These evaluations should be of sufficient scope and quality to improve planning with respect to the effectiveness of the program.

Question 4.5 asks if "independent evaluations of sufficient scope and quality indicate that the program is effective and achieving results." The purpose of question 4.5 is to determine whether the program is effective based on independent and comprehensive evaluations. This question may be particularly important for programs that have substantial difficulty formulating quantitative performance measures. This question suggests that the quality of program evaluations presented in question 2.6 be strongly considered in answering this question.

II. What are the most common ways to evaluate program performance?

The most significant aspect of program effectiveness is *impact*—the outcome of the program, which otherwise would not have occurred without the program intervention. Where it is feasible to measure the *impact* of the program, RCTs are generally the highest quality, unbiased evaluation to demonstrate the actual impact of the program. However, these studies are not suitable or feasible for every program, and a variety of evaluation methods may need to be considered because Federal programs vary so dramatically. Other types of evaluations may provide useful information about the impact of a program (but should be scrutinized given the increased possibility of an erroneous conclusion) or can help address *how* or *why* a program is effective (or ineffective) (i.e., meeting performance targets, achieving efficiency, fulfilling stated purpose). Some of primary evaluation methods are listed and described below.

- *Randomized Controlled Trials* – An RCT is a study that measures an intervention's effect by randomly assigning, for example, individuals (or other units, such as schools or police precincts) into an intervention group, which receives the intervention, and into a control group, which does not. At some point following the intervention, measurements are taken to establish the difference between the intervention group and the control group. Because the control group simulates what would have happened if there were no intervention, the difference in outcomes between the groups demonstrates the "outcome" or impact one would expect for the intervention more generally. There are many programs for which it would not be possible to conduct an RCT. To carry out an RCT, there must be a possibility of selecting randomized intervention and control groups—those who will receive a program intervention and those who will not (or will receive a different intervention). For practical, legal, and ethical reasons, this may not always be possible. (See examples in Section IV.D. of some types of programs for which RCTs may not be possible.)

- *Direct Controlled Trials* – A Direct Controlled Trial is a study where various factors that might influence test results are directly controllable to such a degree that potentially undesirable or external influences are eliminated as significant uncertainties in the outcome of the trial. Such trials are most often possible in technology or engineering programs. For example, in weapon system tests in the Department of Defense, a newly developed weapon will have a test plan that measures the performance of the new weapon under a hostile or adverse environment which simulates a battlefield situation. The performance of the weapon will be measured, analyzed using appropriate statistical and other analytic tools, and the results of that analysis will be compared to the pre-existing but demanding test performance thresholds. In such a case, this evaluation can provide the full measure of rigor needed for evaluation of the development program and for use in acquisition decisions. Another example of this type of evaluation may be a National Aeronautics and Space Administration program to develop a satellite. The test plan would employ appropriate measures and standards of performance so that the satellite subsystem or system could be tested in an appropriate and representative variety of environments and evaluated directly using proper analytical techniques to determine if the development effort has met its goals.

- *Quasi-Experimental* -- Like randomized controlled trials, these evaluations assess the differences that result from a Federally supported activity and the result that would have occurred without the intervention. For example, for a welfare program, the comparison may be between an intervention group that receives the benefits of a program and a comparison group that does not. However, the control activity (comparison group) is not randomly assigned. Instead, it is formed based on the judgment of the evaluator as to how to minimize any differences between the two groups, or it may be a pre-existing group. Quasi-experimental evaluations often are called

"comparison group studies." Under certain circumstances, well-matched comparison group studies can approach the rigor of randomized controlled trials and should be considered if random assignment is not feasible or appropriate. However, use of comparison group studies does increase the risk of misleading results because of the difficulty in eliminating bias in the selection of the control group. Awareness of this risk is crucial to the design of such evaluations. (Also see Section III.B.3.)

- *Non-Experimental Direct Analysis* -- These evaluations examine only the intervention subject (e.g., group)—the subject (group) receiving the program intervention (e.g., for groups, the intervention may be benefits); there is no comparison subject (group). A common example of this type of evaluation, the "pre-post study," examines only an intervention group (no separate comparison group is selected), with outcomes compared both before and after program benefits are received. "Longitudinal studies," which also examine changes over time and relate those changes back to the original condition of the intervention group, are another example.[1] Other examples of non-experimental tools and methods include correlation analyses, surveys, questionnaires, participant observation studies, implementation studies, peer reviews, and case studies. These evaluations 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 (Also see discussion at end of Section III.B.3.).

- *Non-Experimental Indirect Analysis* – In some cases, such as with the results of basic research, the results may be so preliminary in the near-term or so predominantly long-term in nature that a review by a panel of independent experts may be the most appropriate form of assessment. The use of such surrogate analysis must be justified for a specific program based on the lack of viable alternative evaluations that would provide for more meaningful conclusions. Nevertheless, in some cases, such a review may be the best type of assessment available.

When it is not possible to use RCTs to evaluate program impact, agencies should consult with internal or external program evaluation experts, as appropriate, and OMB to identify other suitable evaluation methodologies to demonstrate a program's impact. Some sources of evaluation expertise may include the peer-reviewed literature for the relevant discipline, scientific organizations such as the National Academy of Sciences, think tanks, and research organizations. In addition, to assist in the decision of what type of evaluation will provide the most rigorous evidence appropriate and feasible, the PART guidance provides several links to references on program evaluation. For convenience, they are listed below. They are not intended to be exhaustive, but should be helpful when considering various evaluation methodologies:

- Program Evaluation Methods: Measurement and Attribution of Program Results; *Treasury Board of Canada, Secretariat*; 1998. *(a book available online)*
  http://www.tbs-sct.gc.ca/eval/pubs/meth/pem-mep_e.pdf

- Understanding Impact Evaluation; *The World Bank Group*. (a web site)
  http://www.worldbank.org/poverty/impact/index.htm

- "Program Evaluation: An Evaluation Culture and Collaborative Partnerships Build Agency Capacity;" GAO-03-454; *U.S. General Accounting Office*; May 2003.
  http://www.gao.gov/docdblite/summary.php?recflag=&accno=A06797&rptno=GAO-03-454

- "Performance Measurement and Evaluation: Definitions and Relationships;" GAO/GGD-98-26; *U.S. General Accounting Office*; April 1998.
  http://www.gao.gov/docdblite/summary.php?recflag=&accno=160204&rptno=GGD-98-26

- "Designing Evaluations;" GAO/PEMD-10.1.4; *U.S. General Accounting Office*; May 1991.
  http://161.203.16.4/t2pbat7/144040.pdf

- Randomized Controlled Trials: A User's Guide; Jadad, Alejandro A.; BMJ Books; 1998. *(a book available online)*
  http://www.bmjpg.com/rct/contents.html

- Experimental and Quasi Experimental Designs for Generalized Causal Inference; Cook T.D., Shadish, William, and Campbell, D.T.; Boston: Houghton Mifflin; 2001.

- Research Methods Knowledge Base; Trochim, William M.; Cornell University.
  (a web site)
  http://www.socialresearchmethods.net/kb/index.htm

- "Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide;" U.S. Department of Education; December 2003.
  http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

III. What sorts of tests provide strong evidence of a program's effectiveness?

One of the central challenges in developing strong evidence of a program's effectiveness is valid measurement of the difference between (i) the outcomes when the program is in place, and (ii) what the outcomes would have been in the absence of the program. In a welfare-to-work program, for example, this means measuring the difference between outcomes for program participants (e.g., employment, earnings, welfare dependency) and what their outcomes would have been in the absence of the program.

Because in some cases, particularly with social programs, one cannot directly measure what participants' outcomes would have been in the absence of the program intervention, many evaluation studies seek to *simulate* what would have happened in the absence of the intervention.

For some programs, like those that produce an effect on a physical system or develop a physical or intellectual asset, the effect may be directly producible and testable. On the other hand, some programs that cannot directly isolate the results of the intended action from other possible significant variables (e.g., in a social study) may create a control group by assembling a group of individuals that is as similar as possible to the group of participants, but does not receive the intervention. These studies then estimate the intervention's effect by comparing the outcomes for the two groups. The main studies of this type are RCTs and "comparison-group" studies. What follows is a brief summary of RCTs and some of the advantages over other study designs for evaluations in which the interaction of many intended and unintended effects may influence the outcome, including how likely the RCT is to produce valid estimates of a program intervention's true effect.

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.[2] Some of the advantages of RCTs are summarized as

follows. (Further information can be found in the references provided or at the Council on Excellence in Government website, http://excelgov.org/displayContent.asp?Keyword=prppcHomePage.)

A. The unique advantage of random assignment: It enables you to evaluate whether the intervention itself, as opposed to other factors, causes the observed outcomes.

Randomly assigning a large number of individuals into either an intervention group or a control group ensures, to a high degree of confidence, that there are no systematic differences between the groups in any characteristics (observed and unobserved) except one – namely, the intervention. Therefore, assuming the trial is properly carried out (as described in the Appendix) – the resulting difference in outcomes between the intervention and control groups can confidently be attributed to the intervention and not to other factors.

B. The RCT, when properly designed and implemented, is often better than other study designs in measuring an intervention's true effect.

1. Properly designed, RCTs are the only method that can eliminate the risk of bias, which can adversely affect the results of the evaluation.

Often, the problem with other study designs is that any knowledge of who may be selected for intervention and control groups may *bias*, or influence the resulting selection and introduce known or unknown differences between the groups being studied. When the comparability of the intervention and control groups cannot be assured, this can harm the validity of the results. What this means is that where there is bias, there is the potential for erroneous conclusions about the effectiveness of the intervention.

2. "Single group pre-post" study designs often produce erroneous results.

Definition: A "single group pre-post" study examines whether participants in an intervention improve or become worse off during the course of the intervention, and then attributes any such improvement or deterioration to the intervention.

The problem with this type of study is that, without reference to a randomly-assigned control group, it cannot answer whether the participants' improvement or deterioration would have occurred anyway, even without the intervention. This often leads to erroneous conclusions about the effectiveness of the intervention.

Example. *If a Department of Health and Human Services' Comprehensive Child Development Program that assigned trained case workers to connect poor families with a variety of social services through periodic home visits was evaluated using a single group pre-post design, it would have found that the program was broadly effective in improving participant's lives. However, an RCT design that randomly assigned poor families to the program's services and to a control group that received no services found the program to be ineffective because participants fared no better in all major outcomes than the members of the control group.*

Specifically, the families in both the program and control groups showed significant improvement during the course of the program in the following outcomes: children's vocabulary and achievement scores, mothers' employment and income, families' reliance on welfare and food stamps, and percentage of mothers who were depressed. A single group pre-post study would have attributed the participants' improvement to the program, whereas in fact it was the result of other factors, as evidenced by the equal improvement for families in the control group.[3]

Examples such as the one above – in which the results of RCTs show that single group pre-post studies produce erroneous conclusions – are common; they can be found in almost any area where RCTs have been carried out.

It is important to note that while RCTs can be implemented so that the experimental and control groups are compared after the experimental group receives the intervention (post-only comparison), other designs also are possible as illustrated in the previous example. Often it is useful to measure characteristics of program participants prior to or immediately after random assignment to experimental and control conditions, and then assess impacts after the intervention. This kind of RCT pre-post design allows one to measure changes in both groups and to examine relative changes over time in each. Similarly, both groups can be assessed at multiple points in time to look at longer-term effects. Having baseline data also allows for assessment of potential differences between the groups before the intervention. This kind of analysis enables the study to verify that random assignment equalized the groups across some important variables. It also statistically controls, potentially, for differences that may appear between the groups.

3. The most common "comparison group" study designs also often lead to erroneous conclusions.

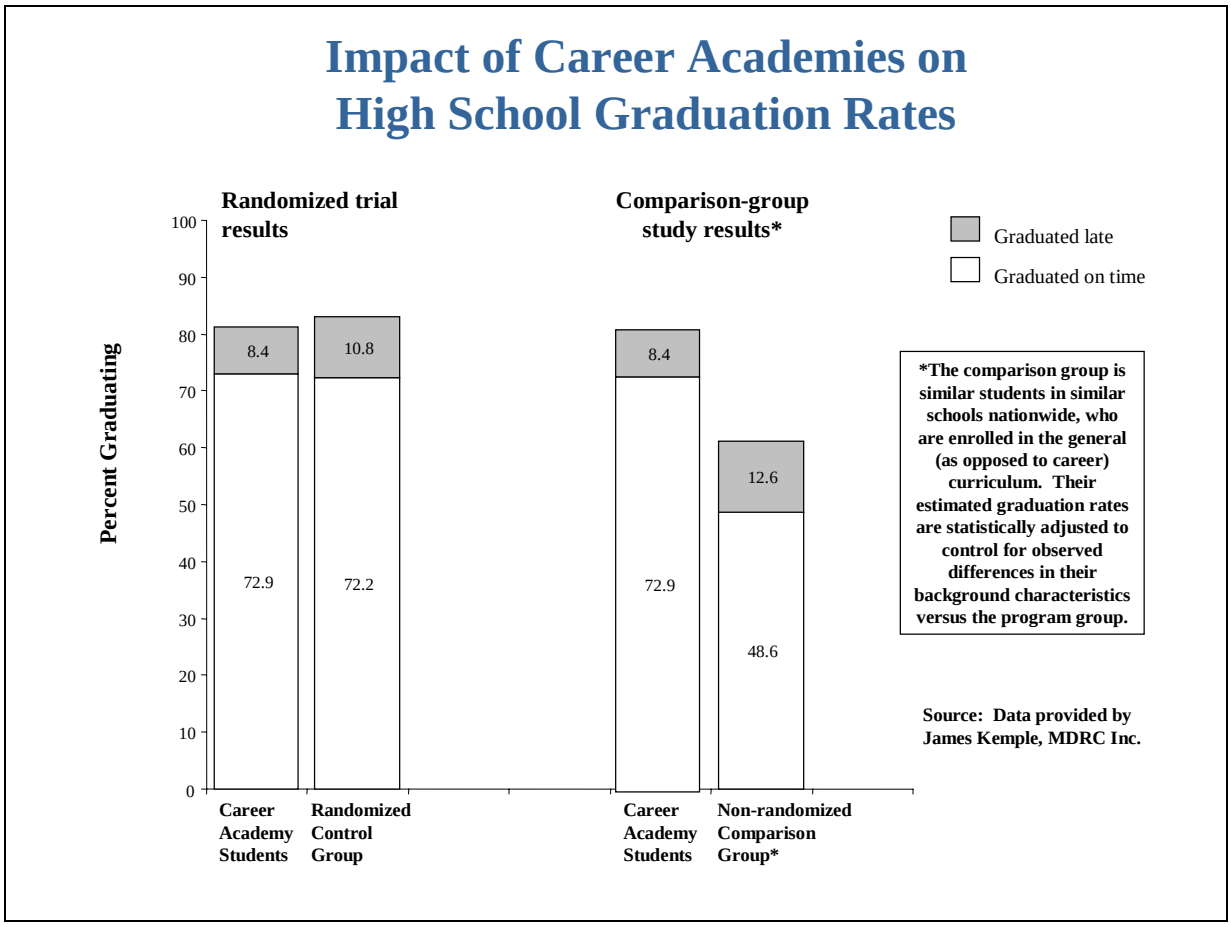
Definition: A "comparison group" study compares outcomes for intervention participants with outcomes for a comparison group chosen through methods other than randomization. For example, comparison-group studies often compare intervention participants with individuals having similar demographic characteristics (age, sex, race, socioeconomic status) who are selected from state or national survey data.

- **Investigations underscore the limitations of comparison-group studies.**

  In social policy, a number of "design replication" studies have been carried out to examine whether and under what circumstances comparison-group studies can replicate the results of RCTs.[4] These investigations first compare participants in a particular intervention with a control group, selected through randomization, in order to estimate the intervention's impact in an RCT. The investigations then compare the same intervention participants with a comparison group selected through methods other than randomization, in order to estimate the intervention's impact in a comparison-group design. Any systematic difference between the two estimates represents the inaccuracy produced by the comparison-group design.

  These investigations have shown that comparison-group studies in social policy (employment, training, welfare-to-work, education) often produce inaccurate estimates of an intervention's effects, because of unobservable differences between the intervention and comparison groups that differentially affect their outcomes. Even when statistical techniques have been used to adjust for observed differences between the two groups, problems have been found. In a sizeable number of cases, the inaccuracy produced by the comparison-group designs is large enough to result in erroneous overall conclusions about whether the intervention is effective, ineffective, or harmful.

Example. *Career Academies is an educational program that enrolls middle and high school student applicants in academic and technical courses in small learning communities with a career theme and partnership with local employers. Participants' high school graduation rates are one of the outcome measures of interest. A well-designed RCT of over 1,700 students that randomly assigned student applicants into an Academy or into a non-Academy control group that continued regular schooling found that the intervention did not result in increased graduation rates at the eight year follow-up. By contrast, if the evaluation had used a comparison group design comprised of like students from similar schools, the evaluation would have concluded erroneously that Career Academies increased the graduation rate by a large and statistically significant 33 percent. The following chart illustrates[5]:*



- **Examples from medicine also show the important limitations of comparison-group studies.**

Example: Hormone replacement therapy. *Over the past 30 years, more than two dozen comparison-group studies have found hormone replacement therapy for postmenopausal women to be effective in reducing the women's risk of coronary heart disease, typically by 35-50 percent. But when hormone therapy was recently evaluated in two large-scale RCTs – medicine's gold standard – it was actually found to do the opposite – namely, it increased the risk of heart disease, as well as stroke and breast cancer.*[6]

The field of medicine contains many other important examples of interventions in which the effect as measured in comparison-group studies has been subsequently contradicted by well-designed RCTs. If RCTs in these cases had never been carried out and the comparison-group results had been relied on instead, the result could have been needless death or serious illness for millions of people. This is why the Food and Drug Administration and National Institutes of Health generally use RCTs as the final arbiter of which medical interventions are effective and which are not.

As a final note, while RCTs are appropriate for addressing questions about causality (i.e., "what effect does an intervention have on outcomes?"), they are not the appropriate tool for answering *how* or *why* a program is effective, or addressing other research needs, such as obtaining descriptive data on:

- Trends (e.g., what is the prevalence of smoking among youth, and is it increasing or decreasing?); and

- Program implementation (e.g., how many people are receiving services under a particular federal program, what services are they receiving, and how satisfied are they with the services?).

Answering such questions, which is critical to making informed policy decisions, is best accomplished with other evaluation methods, such as surveys, questionnaires, implementation studies, etc.

IV. The Application of Randomized Controlled Trials: where they are / are not possible.

A. As a general guideline, RCTs can be carried out in a program where the following conditions apply:

1. Program participants and non-participants can be randomly assigned into two or more groups large enough to comprise a statistically-valid sample;

2. The groups each can be administered a distinct intervention (or non-intervention, which would be the control condition); and

3. For each of the groups, the program can measure the outcomes that the intervention(s) are designed to improve.

In cases where there is no suitable non-intervention group of subjects from which a control group can be selected, RCTs still may be used to test the effectiveness of different interventions, provided that each group is large enough to comprise a statistically-valid sample. In this case, one of the interventions will serve as the control group against which the other interventions will be compared. Still, it would not be possible to measure the net outcome associated with any of the interventions—only the incremental outcome associated with one intervention over another. Such evaluations still may provide useful information about program impacts and may be considered in situations where it is not possible to assign a "non-intervention" control group.

RCTs also may be possible in programs that have partial coverage (for example, not everyone who is eligible for a program is currently able to receive services due to limited program funding). In these circumstances, random assignment of eligible persons to the limited number of

available "slots" may be possible and can provide the opportunity for a rigorous evaluation of the program.

B. RCTs have been carried out in many diverse policy areas.

There is a precedent for carrying out RCTs in a variety of policy areas. As illustrative examples, RCTs have been used with:

- **Medical patients to measure the effectiveness of medical interventions.** *See for example: J.E. Manson et al., "Estrogen Plus Progestin and the Risk of Coronary Heart Disease," New England Journal of Medicine, August 7, 2003, vol. 349, no. 6, pp. 519-522, http://content.nejm.org/cgi/content/short/349/6/523. Also, International Position Paper on Women's Health and Menopause: A Comprehensive Approach, National Heart, Lung, and Blood Institute of the National Institutes of Health, and Giovanni Lorenzini Medical Science Foundation, NIH Publication No. 02-3284, July 2002, pp. 159-160, http://www.nhlbi.nih.gov/health/prof/heart/other/menopaus/menopaus.pdf.*

- **Mentally ill patients to evaluate the effectiveness of drug treatments and psychotherapies.** *See for example: Jeanne Miranda et al., "Treating Depression in Predominantly Low-Income Young Minority Women: A Randomized Controlled Trial," Journal of the American Medical Association, vol. 290, no. 1, July 2, 2003, pp. 57-65, http://jama.ama-assn.org/cgi/content/abstract/290/1/57.*

- **Substance abusers to evaluate the effectiveness of substance-abuse treatment programs.** *See for example: Karen L. Sees et al., "Methadone Maintenance vs. 180-Day Psychosocially Enriched Detoxification for Treatment of Opioid Dependence: A Randomized Controlled Trial," Journal of the American Medical Association, vol. 283, no. 10, March 8, 2000, pp. 1303-1310, http://jama.ama-assn.org/cgi/content/abstract/283/10/1303.*

- **Students to measure the effect of educational interventions.** *See for example: James Kemple and Kathleen Floyd, "Why Do Impact Evaluations? Notes from Career Academy Research and Practice," presentation at a conference of the Coalition for Evidence-Based Policy and the Council of Chief State School Officers, December 10, 2003, http://www.excelgov.org/usermedia/images/uploads/PDFs/MDRC-Conf-12-09-2003.ppt. James J. Kemple and Judith Scott-Clayton, "Career Academies – Impacts on Labor Market Outcomes and Educational Attainment," Manpower Demonstration Research Corporation, March 2004, http://www.mdrc.org/publications/366/overview.html.*

- **Schools to measure the effect of school-wide reform programs.** *See for example: Thomas D. Cook, H. David Hunt, and Robert F. Murphy, "Comer's School Development Program in Chicago: A Theory-Based Evaluation," American Educational Journal, vol. 36, no. 3, fall 1999, pp. 543-59, http://www.northwestern.edu/ipr/publications/comer.pdf.*

- **Young children from disadvantaged backgrounds to evaluate the effectiveness of child care and preschool interventions.** *See for example: Frances A. Campbell et al., "Early Childhood Education: Young Adult Outcomes From the Abecedarian Project," Applied Developmental Science, vol. 6, no. 1, 2002, pp. 42-57, http://www.leaonline.com/doi/abs/10.1207/S1532480XADS0601_05. Also, Lawrence J. Schweinhart, H.V. Barnes, and David P. Weikart, Significant Benefits: The High/Scope Perry Preschool Study Through Age 27 (High/Scope Press, 1993), http://www.highscope.org/Research/PerryProject/perryfact.htm.*

- **Adolescents to measure the effect of violence prevention and substance-abuse prevention programs.** *See for example: Gilbert J. Botvin et al., "Long-Term Follow-up Results of a Randomized Drug Abuse Prevention Trial in a White, Middle-class Population," Journal of the American Medical Association, vol. 273, no. 14, April 12, 1995, pp. 1106-1112, http://jama.ama-assn.org/cgi/content/abstract/273/14/1106.*

- **High-crime areas within a city in order the measure the effectiveness of policing strategies.** *See for example: Anthony A. Braga et al., "Problem-Oriented Policing in Violent Crime Places: A Randomized Controlled Experiment," Criminology, vol. 37, no. 3, August 1999, pp. 541-580, http://www.ncjrs.org/rr/vol1_1/37.html.*

- **Criminal defendants to evaluate the effectiveness of prosecution and sentencing strategies.** *See for example: Denise C. Gottfredson, Stacy S. Najaka, and Brook Kearley, "Effectiveness of Drug Treatment Courts: Evidence from a Randomized Trial," Criminology and Public Policy, vol. 2, no. 2, March 2003, pp. 171-196, http://www.criminologyandpublicpolicy.com/search/abstrGottfredson03.php.*

- **Prison inmates to evaluate the effectiveness of programs to facilitate their re-entry into society.** *See for example: Harry K. Wexler et al., "Three-Year Reincarceration Outcomes for Amity In-Prison Therapeutic Community and Aftercare in California," The Prison Journal, vol. 79, no. 3, September 1999, pp. 321-336, http://www.amityfoundation.com/lib/libarch/99wexler_3yroutcom.pdf.*

- **Low-income families to evaluate the effectiveness of income maintenance, poverty reduction, welfare-to work, job training, food and nutrition, and related programs.** *See for example: Gayle Hamilton et al., "National Evaluation of Welfare-to-Work Strategies: How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs," prepared by MDRC and Child Trends for the U.S. Department of Health and Human Services and U.S. Department of Education, November 2001, http://aspe.hhs.gov/hsp/NEWWS/5yr-11prog01/. Also, Lisa A. Gennetian, "The Long-Term Effects of the Minnesota Family Investment Program on Marriage and Divorce Among Two-Parent Families," prepared by MDRC for the U.S. Department of Health and Human Services, October 2003, http://www.mdrc.org/publications/357/full.pdf.*

- **Public housing residents to evaluate the effectiveness of housing voucher programs.** *See for example: Lawrence F. Katz, Jeffrey R. Kling, and Jeffrey B. Liebman, "Moving To Opportunity in Boston: Early Results of a Randomized Mobility Experiment," Quarterly Journal of Economics, May 2001, pp. 606-654, http://ideas.repec.org/a/tpr/qjecon/v116y2001i2p607-654.html. Also, Jens Ludwig, Greg J. Duncan, and Paul Hirschfield, "Urban Poverty and Juvenile Crime: Evidence From a Randomized Housing-Mobility Experiment," Quarterly Journal of Economics, May 2001, pp. 655-679, http://ideas.repec.org/a/tpr/qjecon/v116y2001i2p655-679.html.*

- **Voters to measure the effect of voter turnout strategies.** *See for example: Alan S. Gerber and Donald P. Green, "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment," American Political Science Review, vol. 94, no. 3, September 2000, pp. 653-663, http://www.yale.edu/isps/publications/GerberGreen.pdf.*

- **College students to measure the effectiveness of strategies to improve racial tolerance.** *See for example: Greg J. Duncan et al., "Empathy or Antipathy? The Consequences of*

Racially and Socially Diverse Peers on Attitudes and Behaviors," Joint Center for Poverty Research working paper, May 16, 2003, http://www.jcpr.org/wpfiles/Duncan_et_al_peer_paper.pdf.

- **Health insurance enrollees to evaluate the effect of various health insurance plans on health, customer satisfaction, and cost.** *See for example: Willard G. Manning et al., "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," The American Economic Review, vol. 77, no. 3, June 1987, pp. 251-277, http://ideas.repec.org/a/aea/aecrev/v77y1987i3p251-77.html.*

- **Medicare and Medicaid beneficiaries to evaluate the effectiveness of various approaches to health care delivery.** *See for example: Leslie Foster, et al., "Improving The Quality Of Medicaid Personal Assistance Through Consumer Direction," Health Affairs Web Exclusive, March 26, 2003, http://content.healthaffairs.org/cgi/reprint/hlthaff.w3.162v1.pdf. Donald L. Patrick et al., "Cost and Outcomes of Medicare Reimbursement for HMO Preventive Services," Health Care Financing Review, vol. 20, no. 4, Summer 1999, pp. 25-43, http://www.cms.hhs.gov/review/99Summer/99Summerpg25.pdf.*

- **Whole communities in developing countries to evaluate the effectiveness of family planning programs and poverty reduction programs.** *See for example: Emmanuel Skoufias and Bonnie McClafferty, "Is PROGRESA Working? Summary of the Results of an Evaluation by IFPRI," International Food Policy Research Institute, Food Consumption and Nutrition Division Discussion Paper No. 118, July 2001, http://www.ifpri.org/divs/fcnd/dp/papers/fcndp118.pdf.*

- **Children in developing countries to evaluate the effectiveness of nutrition, health, and education interventions.** *See for example: John Newman, Laura Rawlings, and Paul Gertler, "Using Randomized Control Designs in Evaluating Social Sector Programs in Developing Countries," The World Bank Research Observer, vol. 9, no. 2, July 1994, pp. 181-201, http://poverty.worldbank.org/library/view/5668/.*

- **Taxpayers to evaluate the effectiveness of various tax compliance strategies.** *See for example: Lawrence W. Sherman, Edward Poole, and Christopher S. Koper, Preliminary Report to the Pennsylvania Department of Revenue on the "Fair Share" Project, Jerry Lee Center of Criminology, Fels Institute of Government, University of Pennsylvania, 2004.*

There are many other policy areas where RCTs have not yet been carried out but for which they may be feasible.

C. The costs of conducting RCTs are not always prohibitive.

RCTs can cost anywhere from $50,000 to $50 million. At one end, large, multi-site RCTs – which have the potential, by themselves, to yield strong evidence of an intervention's effectiveness – may typically cost in the range of $10 to $50 million. Endnote 7 shows the cost of large RCTs that have been carried out in K-12 education in recent years.[7]

Importantly, however, small-scale RCTs – which may contribute to strong evidence of a program's impact – can sometimes cost far less than large-scale ones. In addition, the cost of an RCT (small or large) can sometimes be reduced dramatically by measuring outcomes with data that is *already being collected* for other purposes. This reduction in cost can be dramatic because the primary expense in most RCTs is the cost of collecting outcome data.

Example. *Recently, an RCT was carried out to evaluate Fast Forward, a computerized reading intervention, in the Hartford, Connecticut school district. The trial randomly assigned approximately 500 students to an intervention or control group, and measured many (but not all) educational outcomes with achievement tests that the schools were already administering for other purposes. The trial found that the intervention was not effective in improving reading skills. The cost of the trial was approximately $300,000 - $350,000.*[8]

Example. *The Pennsylvania state government recently commissioned a large RCT to evaluate the effectiveness of various approaches to improving tax compliance by businesses that were late in paying their sales taxes. The trial randomly assigned 7000 such businesses to receive one of seven letters, ranging from threatening to pleading, and made use of outcome data that the state already collected for other purposes – namely, whether the businesses paid their taxes. The trial found that a letter containing a short (1/3 page) statement that tax is due and that the business is liable produced significantly more tax revenue than the state's existing letter (full-page, detailed letter with boxes that the businesses check to indicate why they have not paid the tax). The trial results indicated that the state's use of the short letter for all late-paying businesses could generate $6 million annually in increased revenue. The cost of the trial was $102,000.* [9]

The above also are examples in which RCTs produced valuable results in a very short time frame – within a year or two. Some trials take much longer to produce results. But even trials of interventions that are designed to have a long-term effect (e.g., early childhood programs) often begin producing valuable information on short-term outcomes (e.g., language skills) within 2-3 years. Sometimes, but not always, these short-term outcomes are a harbinger of the longer-term outcomes (e.g., high school graduation rate, employment, welfare dependency) that are of the greatest policy significance.

Even in cases in which the costs of conducting RCTs appear quite significant, it is important to recognize that other evaluations that also attempt to measure impact also may have similar costs. For example, well-designed comparison group studies have data requirements that are quite similar to that of RCTs and do not necessarily offer cost savings.

D. There are many programs for which it is not possible or practical to carry out RCTs.

As discussed earlier, there are many programs for which it is not possible to conduct an RCT. For an agency or program to conduct an RCT, there must be a possibility of selecting randomized intervention and control groups. The agency or program must have sufficient discretion in the administration of the program to permit random assignment of groups who will receive a program intervention and those who will not (or will receive a different intervention). For practical, legal, and ethical reasons, this may not always be possible. Where a program is broadly providing a public good, for example, such as clean air, homeland security, and basic research benefits—no one can be excluded from the intervention. Other examples:

- One cannot carry out an RCT to evaluate whether reducing carbon emissions will prevent global warming, because there is only one planet earth. (However, it may be possible to randomize industrial sites in order to evaluate the effectiveness and cost of various methods of reducing carbon emissions.)

- One cannot carry out an RCT to evaluate the effectiveness of manned space flight, because we can only afford to carry out one such program.

- One cannot carry out an RCT to evaluate military assistance to NATO countries, because of the political impossibility of randomizing countries as well as the lack of sufficient numbers of countries to form valid statistical groupings.

- One cannot choose a random sample of military operations in which to use particular operational strategies, because once a particular operation is approved, any tool or strategy that might help under changing conditions must be available for use.

- One cannot carry out an RCT to evaluate the effectiveness of Federal disaster assistance because of the legal and/or ethical problems associated with denying benefits to some victims or providing different types of benefits to different groups of victims suffering from the same kind of disaster.

- One cannot carry out an RCT to evaluate the effectiveness of a health, safety, or financial regulation program because of the legal and/or ethical problems associated with denying protection to people covered by the law. (However, it may be possible to use RCTs to test new approaches to improving health, safety, and financial outcomes, the results of which can inform future regulatory action.)

In cases where it is not possible to use an RCT to evaluate the effectiveness of a program intervention, other approaches may be needed to evaluate: *What difference does the program make?* To approach an assessment of impact, the analysis must make every effort to compare the effect of the program with a baseline of what would have occurred in the absence of the program—an extremely difficult test. Finally, if it is not possible to evaluate the impact of a program, other evaluation approaches may shed light on *how* or *why* a program is effective (or ineffective), or may provide other information that is needed for the management of the program.

Appendix – The Quality and Quantity of Randomized Controlled Trials Needed To Establish Strong Evidence of an Intervention's Effectiveness

This appendix sets out key principles for evaluating the quality and quantity of randomized controlled trials (RCTs) needed to establish evidence of a program's impact.

Well-designed and implemented evaluations.

RCTs generally provide "strong" evidence of a program intervention's effectiveness. However, in order to do so, they also must be well-designed and implemented in order to constitute strong evidence. What follows are some key elements of a well-designed and implemented trial:

1. **The study should clearly describe the program intervention**, including: (i) who administered it, who received it, and what it cost; (ii) how it differed from what the control group received; and (iii) the logic of how it is supposed to affect outcomes.

2. **The study should use placebo controls if participants' beliefs that they are receiving a program intervention may plausibly affect their outcomes.** Placebo controls would be appropriate, for example, in a study of a drug treatment for a mental illness, where participants' beliefs that they are receiving treatment may plausibly alleviate their symptoms.

3. **The random assignment process should include safeguards to ensure it is not compromised.** For example, individuals (or other subjects) randomly assigned to the control group should not receive the program intervention, nor have an opportunity to cross over to the intervention group. Also, individuals unhappy with their prospective assignment to either the intervention or control group should not have an opportunity to delay their entry into the study until another opportunity arises for assignment to their preferred group.

4. **The study should provide data showing that, prior to the program intervention, the intervention and control groups do not differ systematically in their measured characteristics** (allowing that, by chance, there may be a few minor differences). The program intervention and control groups studied must be comparable (identical in key, measurable characteristics) before the program intervention and evaluation actually begin in order for the RCT to identify differences in outcomes that are the result of the intervention and not a difference between the two groups.

5. **The study should use outcome measures that are valid – – i.e., that accurately measure the true outcomes that the program intervention is designed to affect.** For example:

   - If the study uses tests to measure outcomes (e.g., tests of academic achievement or psychological well-being), the tests used should be ones whose ability to accurately measure true outcomes (e.g., true academic skills or psychological status) is well-established.

   - Wherever possible, a study should measure the final outcomes that the intervention is designed to affect (e.g., for an applied research program, whether the program yields commercially-successful technological innovations, rather than whether it merely funds research of high technical quality).

- If outcomes are measured through interviews or observation, the interviewers/observers preferably should be "blinded" – i.e., kept unaware of who is in the intervention and control groups.

- When study participants are asked to self-report outcomes, their reports should be corroborated, if possible, by independent and/or objective measures (e.g., for a crime prevention program, official arrest data).

6. **The study should have low overall attrition of study participants, and no differential attrition between the intervention and control groups.** As a general guideline, the study should obtain outcome data for at least 80 percent of the individuals (or other subjects) originally randomized. (Studies that choose to follow only a representative subsample of the randomized individuals should obtain outcome data for at least 80 percent of the subsample.) Furthermore, the attrition rate should be approximately the same for the intervention and the control groups. In addition, the study should test for potential bias due to attrition, especially if there is differential attrition or response rates over time fall below 80 percent.

7. **The study should use an intention-to-treat approach.** That is, it should collect outcome data for all individuals (or other subjects) who were randomly assigned, even those who do not participate in or complete the program intervention. The study also should use the outcome data for all randomly-assigned individuals in estimating the intervention's effect.

8. **The study should preferably obtain data on long-term outcomes of the program intervention** to enable policymakers to judge whether the intervention's effects were sustained over time.

9. **The study should conduct power analyses as part of its evaluation design to ensure that the sample sizes are adequate to be able to detect as statistically significant the magnitude of the effect(s) that the program is likely to have (based on previous studies or comparable investigations in the research literature).** If the study claims that the intervention improves one or more outcomes, it should report (i) the size of the effect, and (ii) tests showing that the effect is statistically significant.

10. **The study should report the intervention's effects on all the outcomes that the study measured, not just those for which there is a positive effect.** This is because if a study measures a large number of outcomes, the study may find, by chance alone, positive and statistically-significant effects on one or a few of those outcomes. Thus, the study should report – at least in summary form – the intervention's effects on all measured outcomes so that readers can judge whether the positive effects are the exception or the pattern.

Quantity of evidence needed.

1. **In order for RCTs to generate strong evidence of program impact, they generally require:**

   (i) that the program intervention be demonstrated effective through well-designed RCTs, in more than one site of implementation, and

   (ii) that these sites be typical settings where policymakers would seek to implement the program if it is found effective (e.g., community drug abuse clinics, public schools, job training program sites). Typical settings would not include, for example, specialized sites that researchers set up and administer at a university for purposes of the study.

Such a demonstration of effectiveness may require more than one RCT of the intervention, or one large trial with more than one implementation site.

2. **Different types of programs require different considerations in evaluating impact. For example, for:**

   - **Demonstration programs or programs that operate in a limited number of areas** - RCTs may demonstrate what can be achieved or whether an approach is promising enough to pursue further.

   - **Established national programs** - evaluations should ensure that sites and program participants are selected using probability methods to enable results to be statistically representative of population sites administering the programs and program participants. Only this kind of design can provide an accurate measurement of the actual effect of the program as it is currently operating.

3. **Main reasons why a demonstration of effectiveness in more than one site is needed:**

   - **A single finding of effectiveness can sometimes occur by chance alone.** For example, even if all program interventions tested in RCTs were ineffective, we would expect 1 in 20 of those trials to "demonstrate" effectiveness by chance alone at conventional levels of statistical significance. Two RCTs (or two sites of one large RCT) reduces the likelihood of such a false positive result to 1 in 400.

   - **The results of a trial in any one site may be dependent on site-specific factors and thus may not be generalizable to other sites.** It is possible, for instance, that an intervention may be highly effective in a site with an unusually talented individual managing the details of implementation, but would not be effective in another site with other individuals managing the detailed implementation.

4. **Pharmaceutical medicine provides an important precedent for the concept that strong evidence requires a showing of effectiveness in more than one instance.** Specifically, the Food and Drug Administration (FDA) usually requires that a new pharmaceutical drug or medical device be shown effective in more than one RCT before the FDA will grant it a license to be marketed. The FDA's reasons for this policy are similar to those discussed above.[10]

Notes

[1] Some evaluation literature refer to "pre-post" and "longitudinal" studies as forms of quasi-experimental evaluation, because a reflexive comparison is made between the group receiving the program intervention and a *control group* composed of the same group before the intervention. Other sources do not consider such studies to be quasi-experimental.

[2] See, for example, the Food and Drug Administration's standard for assessing the effectiveness of pharmaceutical drugs and medical devices, at 21 C.F.R. §314.126. See also, "The Urgent Need to Improve Health Care Quality," Consensus statement of the Institute of Medicine National Roundtable on Health Care Quality, *Journal of the American Medical Association*, vol. 280, no. 11, September 16, 1998, p. 1003; and Gary Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, vol. 9, no. 2, spring 1995, pp. 63-84.

[3] Robert G. St. Pierre and Jean I. Layzer, "Using Home Visits for Multiple Purposes: The Comprehensive Child Development Program," *The Future of Children*, vol. 9, no. 1, spring/summer 1999, p. 134.

[4] 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, at http://www.mdrc.org/ResearchMethodologyPprs.htm. James J. Heckman et al., "Characterizing Selection Bias Using Experimental Data," *Econometrica*, vol. 66, no. 5, September 1998, pp. 1017-1098. Daniel Friedlander and Philip K. Robins, "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *American Economic Review*, vol. 85, no. 4, September 1995, pp. 923-937. Thomas Fraker and Rebecca Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources*, vol. 22, no. 2, spring 1987, pp. 194-227. Robert J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review*, vol. 176, no. 4, September 1986, pp. 604-620. Roberto Agodini and Mark Dynarski, "Are Experiments the Only Option? A Look at Dropout Prevention Programs," Mathematica Policy Research, Inc., August 2001, at http://www.mathematica-mpr.com/PDFs/redirect.asp?strSite=experonly.pdf. Elizabeth Ty Wilde and Rob Hollister, "How Close Is Close Enough? Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes," Institute for Research on Poverty Discussion paper, no. 1242-02, 2002, at http://www.ssc.wisc.edu/irp/.

This literature is systematically reviewed in Steve Glazerman, Dan M. Levy, and David Myers, "Nonexperimental Replications of Social Experiments: A Systematic Review," Mathematica Policy Research discussion paper, no. 8813-300, September 2002. The portion of this review addressing labor market interventions is published in "Nonexperimental versus Experimental Estimates of Earnings Impact," *The American Annals of Political and Social Science*, vol. 589, September 2003.

[5] James Kemple and Kathleen Floyd, "Why Do Impact Evaluations? Notes from Career Academy Research and Practice," presentation at a conference of the Coalition for Evidence-Based Policy and the Council of Chief State School Officers, December 10, 2003, http://www.excelgov.org/usermedia/images/uploads/PDFs/MDRC-Conf-12-09-2003.ppt. James J. Kemple and Judith Scott-Clayton, "Career Academies – Impacts on Labor Market Outcomes and Educational Attainment, " Manpower Demonstration Research Corporation, March 2004, http://www.mdrc.org/publications/366/overview.html. Although the study found that Career Academies had no effect on high school graduation rates, it did find program participants had significantly higher earnings than members of the control group over the eight-year follow-up period.

[6] J.E. Manson et al., "Estrogen Plus Progestin and the Risk of Coronary Heart Disease," *New England Journal of Medicine*, August 7, 2003, vol. 349, no. 6, pp. 519-522. *International Position Paper on Women's Health and Menopause: A Comprehensive Approach*, National Heart, Lung, and Blood Institute of the National Institutes of Health, and Giovanni Lorenzini Medical Science Foundation, NIH Publication No. 02-3284, July 2002, pp. 159-160. Stephen MacMahon and Rory Collins, "Reliable Assessment of the Effects of Treatment on Mortality and Major Morbidity, II: Observational Studies," *The Lancet*, vol. 357, February 10, 2001, p. 458. Sylvia Wassertheil-Smoller

et al., "Effect of Estrogen Plus Progestin on Stroke in Postmenopausal Women – The Women's Health Initiative: A Randomized Controlled Trial, *Journal of the American Medical Association*, May 28, 2003, vol. 289, no. 20, pp. 2673-2684. Recent reviews of this study have raised some questions about its conclusions because of the age range of the study group. See for example: Naftolin at al, The Women's Health Initiative could not have detected cardioprotective effects of starting hormone therapy during the menopausal transition, Fertility and Sterility, Vol.81, No.6, June 2004.

[7] The following are illustrative examples of the cost of large randomized controlled trials in K-12 education:

- In the 1990s, the U.S. Education Department funded randomized controlled trials to evaluate the Department's School Dropout Demonstration Assistance program and its Upward Bound program. The Dropout Demonstration study involved randomized controlled trials in each of 16 sites, and cost $7.3 million over 1991-1995. The Upward Bound study involved randomized controlled trials in each of 67 sites, and cost over $5.4 million during 1992-1996.

- Tennessee's Student-Teacher Achievement Ratio (STAR) Project was a large-scale randomized controlled trial, funded by the state of Tennessee, that examined the effect of reducing class size in early elementary school on student achievement. The trial, involving 79 schools and over 11,000 students, cost $12 million over 1985-1989.

- Success For All, a comprehensive school reform program, is currently being evaluated in a randomized controlled trial funded by the U.S. Education Department. The trial, involving 60 schools over a five-year period, is expected to cost over $6 million. The Department estimates that a larger initiative to evaluate 5 to 10 school reform models in similar trials would cost $42 million over a six-year period.

Source: Coalition for Evidence-Based Policy, *Bringing Evidence-Driven Progress To Education: A Recommended Strategy for the U.S. Department of Education*, http://www.excelgov.org/usermedia/images/uploads/PDFs/coalitionFinRpt.pdf, November 2002, p. 16.

[8] Cecilia Elena Rouse and Alan B. Krueger, "Putting Computerized Reading Instruction to the Test: A Randomized Evaluation of a 'Scientifically-based' Reading Program," *Economics of Education Review* (forthcoming), http://www.ers.princeton.edu/workingpapers/5ers.pdf. The estimated cost of the trial is based on correspondence with the authors.

[9] Lawrence W. Sherman, Edward Poole, and Christopher S. Koper, *Preliminary Report to the Pennsylvania Department of Revenue on the "Fair Share" Project,* Jerry Lee Center of Criminology, Fels Institute of Government, University of Pennsylvania, 2004.

[10] *Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products*, Food and Drug Administration, May 1998, pp. 2-5