

Which Study Designs Can Produce Rigorous Evidence of Program Effectiveness? A Brief Overview

This paper is addressed to policymakers and researchers who are seeking to build scientifically-valid evidence about the effectiveness of a program, policy, or practice (“intervention”) – evidence that is strong enough to help inform policy decisions. The paper seeks to provide brief, concrete advice on the types of study designs that are capable of generating such evidence – preferably randomized controlled trials or, if not feasible, prospective, closely-matched comparison-group studies.* Specifically, this paper discusses:

- (i) Randomized controlled trials – what they are; why they are considered the gold standard for evaluating an intervention’s effect; areas where they are feasible and where they are not; and practical resources that can help you sponsor or conduct such studies; and
- (ii) Prospective, closely-matched comparison-group studies – what they are; when they should be considered; and how to design them so as to maximize the chances that they will produce valid results.

I. Wherever possible, we recommend that policymakers and researchers seeking to rigorously evaluate an intervention do so in a randomized controlled trial.

A. **Definition:** Randomized controlled trials measure an intervention’s effect by randomly assigning individuals (or groups of individuals) to an intervention group or a control group.

For example, suppose that a government agency wants to rigorously evaluate how effective a new program to facilitate the re-entry of prisoners into the community is in reducing recidivism. The agency might sponsor a randomized controlled trial which randomly assigns prisoners to either an intervention group, which participates in the new program, or to a control group, which receives the usual (pre-existing) services for prisoners. The study would then measure outcomes – such as re-arrest and re-incarceration rates – for both groups over a period of time. The difference in outcomes between the two groups would represent the effect of the new program compared to the services normally provided.

B. **Well-designed trials are recognized as the gold standard for evaluating an intervention’s effectiveness in many diverse fields, such as welfare and employment, medicine, psychology, and education.¹**

In medicine, for example, randomized controlled trials have provided the conclusive evidence of effectiveness for most of the major medical advances over the past 50 years, including: (i) vaccines for polio, measles, and hepatitis B; (ii) interventions for hypertension and high cholesterol, which in turn have helped bring about a decrease in coronary heart disease and stroke by more than 50% over the past half-century; and (iii) cancer treatments that have dramatically improved survival rates from leukemia, Hodgkin’s disease, breast cancer, and many other

* This is not meant to deny the value of less rigorous study designs, which may be very useful for other purposes – for example, in generating hypotheses about what works that merit confirmation in more rigorous studies. One research strategy, in fact, is to sponsor or carry out low-cost, less rigorous studies of a wide range of interventions, to identify areas where an additional research investment, using more rigorous methods, is warranted. This paper, however, limits its discussion to the more rigorous methods that are themselves capable of producing evidence that might help inform policy.

cancers. Notably, for over 40 years the Food and Drug Administration (FDA) has required that any new pharmaceutical drug – and, since 1976, medical device – be demonstrated effective in such trials before the FDA will approve it for marketing.

In U.S. welfare policy, randomized controlled trials over the past 25 years have built a valuable knowledge base on what works in moving people from welfare to work and in improving family well-being.² This knowledge was a key to the political consensus behind the 1988 welfare reform law and also helped shape the 1996 act, which led to dramatic changes in state and federal programs resulting in major reductions in welfare rolls and gains in employment among low-income Americans.³

Randomized controlled trials have been undertaken in many other, diverse policy areas, demonstrating:

- effective approaches to increasing tax compliance among delinquent taxpayers;
- effective policing, prosecution, and sentencing strategies;
- effective foreign aid programs to reduce poverty in developing countries;
- effective incentives for retirement savings by low and middle income taxpayers;
- effective approaches to improving racial tolerance among college students; and
- effective strategies to increase voter turnout,

to name a few illustrative examples (see the Appendix for additional discussion and references).

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

Specifically, the process of randomly assigning a sufficiently 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 group participates in the intervention, and the control group does not. Therefore, assuming the randomized controlled trial is properly carried out, the resulting difference in outcomes between the two groups can confidently be attributed to the intervention and not to other factors.

D. There is persuasive evidence that well-designed randomized controlled trials are superior to other study designs in measuring an intervention’s true effect.

There is also strong evidence that the most commonly-used nonrandomized methods – including “pre-post” studies and “comparison-group” studies without careful matching – often produce erroneous conclusions and can lead to practices that are ineffective or harmful.⁴ The following discussion elaborates.

1. “Pre-post” study designs often produce erroneous results.

Definition: A 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 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. Such studies should therefore not be relied upon to inform

policy decisions (but may still be useful in hypothesis-generation, as described in the footnote on page 1).

Example. A pre-post study of Even Start – a federal program designed to improve the literacy of disadvantaged families – found that the children in the program made substantial improvements in school readiness during the course of the program (e.g., an increase in their national percentile ranking on the Picture Peabody Vocabulary Test from the 9th to the 19th percentile). However, a randomized controlled trial of Even Start carried out by the same researchers found that the children in the *control* group improved by approximately the same amount over the same time period. Thus, the program had no *net* impact on the children’s school readiness. If the researchers had only carried out the pre-post study, and not the randomized controlled trial, their results would have suggested erroneously that Even Start is highly effective in increasing school readiness.⁵

2. **“Comparison group” study designs (also known as “quasi-experimental” designs) without close matching also lead to erroneous conclusions in many cases.**

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.

In social policy (e.g., welfare and employment, education), 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 randomized controlled trials. These investigations have shown that the most commonly-used comparison-group study designs – which do not include very close matching of the intervention and comparison groups – often produce inaccurate estimates of an intervention’s effects, because of unobserved differences between the intervention and comparison groups that differentially affect their outcomes. This is true even when statistical techniques are used to adjust for observed differences between the two groups.⁶ Therefore, such studies – like pre-post studies – should not be relied upon to inform policy decisions (but may still be useful in hypothesis-generation, as described in the footnote on page 1).

The field of medicine also contains important evidence of the limitations of most comparison-group studies. The following is an illustrative example:

Example. 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 randomized controlled trials – 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.⁷

- 3. The evidence suggests that, among nonrandomized studies, closely-matched comparison-group studies can be a second-best alternative when a randomized controlled trial is not feasible.**

Specifically, the design replication studies noted above suggest that comparison-group studies in which the intervention and comparison groups are *very closely matched* in key characteristics, as described in the next section, can be a second-best alternative to a randomized controlled trial. Among comparison-group studies, these closely-matched studies are the most likely to generate valid conclusions about an intervention's effectiveness. However, their estimates of the magnitude of an intervention's effect are often inaccurate, and in some instances they may still produce erroneous overall conclusions about whether the intervention is effective, ineffective, or harmful.

- E. The following are practical, easy-to-use resources that can help you decide whether and how a randomized controlled trial might be carried out in your program area:**

- 1. The Appendix discusses areas where randomized controlled trials are possible and where they are not, and where they can be conducted at low cost.**
- 2. *Social Programs That Work* (www.evidencebasedprograms.org) contains concise summaries of well-designed trials with important policy implications.**
- 3. Concrete advice on how to gain the cooperation of program officials, staff, and participants in a trial, and address their practical/ethical concerns, is contained in:**

Larry L. Orr, *Social Experimentation: Evaluating Public Programs With Experimental Methods: Implementation and Data Collection*, prepared for the Assistant Secretary of HHS for Planning and Evaluation, <http://aspe.hhs.gov/search/hsp/qeval/part5.pdf>, October 1997, pp. 2-9; and

How to Conduct Rigorous Evaluations of Mathematics and Science Partnerships (MSP) Projects: A User-Friendly Guide for MSP Project Officials and Evaluators, prepared by the Coalition for Evidence-Based Policy for the U.S. Education Department's MSP program, through a subcontract with the National Opinion Research Center, <http://www.ed.gov/programs/mathsci/mspbrief2.doc>, August 2005, pp. 11-14.

Although these documents focus primarily on welfare, employment, and education policy, their content is largely applicable in other policy areas as well.

- 4. An easy-to-use checklist of key items to get right when conducting a randomized controlled trial is contained in:**

Key Items To Get Right in a Randomized Controlled Trial in Education, prepared by the Coalition for Evidence-Based Policy for the What Works Clearinghouse of the U.S. Education Department's Institute of Education Sciences, http://www.evidencebasedpolicy.org/guide_RCT.pdf, December 2005. Although this document focuses on the evaluation of educational interventions, its contents are applicable to other policy areas as well.

- 5. The Evidence-Based Policy Help Desk (www.evidencebasedpolicy.org) contains all of the above resources, and other helpful tools, in a user-friendly website.**

II. We suggest you consider a matched comparison-group study only if you have exhausted all options for conducting a randomized controlled trial and conclude it is not feasible.

As discussed in the Appendix, in some cases conducting a randomized controlled trial may not be possible for practical, legal, or ethical reasons – for example, when the intervention being evaluated does not have the administrative or legal discretion to randomly assign applicants or participants to intervention versus control groups. In such cases, comparison-group studies in which the intervention and comparison groups are *very closely matched* in key characteristics can be a second-best alternative, based on the evidence discussed above.

What follows is a discussion of key principles we suggest researchers adhere to when conducting a matched comparison-group study, to maximize chances that the study will produce valid results. This discussion is limited to key principles, and does not try to address all contingencies that may affect the study's success.

A. The study's intervention and comparison group members should be very closely matched on characteristics that may predict their outcomes.

For example, in a study of a substance abuse prevention program the groups should be matched on characteristics that are known to predict future substance abuse (e.g., prior drug use, age, sex).

There is persuasive evidence that when a comparison-group study does not include close matching on such characteristics, the study is unlikely to generate accurate results even when statistical methods (such as regression adjustment) are used to correct for the differences between the two groups in estimating the intervention's effect.

What follows is more specific advice on how to implement the above matching principle:

1. The best predictors of study participants' outcomes, and therefore the most important to match on, are pre-intervention measures of these outcomes.

For example, in an evaluation of an intervention to prevent future criminal activity among offenders being released from prison, the offenders in the two groups should be closely matched on one or more measures of their pre-intervention criminal activity, such as number of arrests, convictions, and severity of offenses.

2. Of secondary importance, the two groups should be matched on demographic and other characteristics that are likely predictors of their outcomes (e.g. age, sex, ethnicity, poverty level, geographic location).

For example, a study of a crime prevention program for youth should include age and gender as matching criteria because studies show young men are more likely to commit crimes than women or older men.

Characteristics that predict who will *participate* in an intervention are often also good predictors of participants' outcomes, and therefore should be considered as matching criteria. For example, a study of a job training + child care program that tends to attract single mothers as opposed to married mothers should use marital status as a matching criterion.

B. The comparison group should not be comprised of individuals that had the option to participate in the intervention but declined.

This is because individuals that choose not to participate in an intervention may differ systematically from the individuals who do choose to participate in their level of motivation and other important characteristics. The difference in motivation (or other characteristics) may itself lead to different outcomes for the two groups, thereby causing the study to produce inaccurate estimates of the intervention's true effect.

Therefore, the comparison group should be comprised of individuals that did not have the option to participate in the intervention, rather than individuals who had the option but declined.

C. The study should choose the intervention and comparison groups “prospectively” – i.e., before the intervention is administered.

This is because if the intervention and comparison groups are chosen by the researcher *after* the intervention is administered (“retrospectively”), the researcher may consciously or unconsciously select the two groups so as to generate his or her desired results. Indeed, in many cases there may be dozens of possible intervention and comparison groups that the researcher can choose from, and if the researcher looks at enough of these, he or she will likely be able to find one such combination that will produce the desired result just by chance.

Prospective comparison-group studies are, like randomized controlled trials, much less susceptible to this problem. In the words of the director of drug evaluation for the Food and Drug Administration, “The great thing about a [randomized controlled trial or prospective comparison-group study] is that, within limits, you don't have to believe anybody or trust anybody. The planning for [the study] is prospective; they've written the protocol before they've done the study, and any deviation that you introduce later is completely visible.” By contrast, in a retrospective study, “you always wonder how many ways they cut the data. It's very hard to be reassured, because there are no rules for doing it.”⁸

D. The study should follow the same practices that a well-designed randomized controlled trial should follow in order to produce valid results (other than the actual random assignment).

That is, the study should use valid outcome measures, prevent “cross-overs” to or “contamination of” the comparison group, have low sample attrition, use an “intention-to-treat” analysis, and so on. As noted above, an easy-to-use checklist of these key items is contained in the following document: *Key Items To Get Right When Conducting a Randomized Controlled Trial in Education*, at http://www.evidencebasedpolicy.org/guide_RCT.pdf. For a matched comparison-group study (as opposed to a randomized controlled trial), we suggest you add items A, B, and C above to the checklist, and disregard the checklist's item about protecting the integrity of the random assignment process.

E. Endnote 9 contains an illustrative example of a prospective, closely-matched comparison-group study that was well-designed and implemented.⁹

This is a study of a preschool intervention for disadvantaged children, designed to improve educational outcomes. The study follows the principles outlined in this paper, with one exception discussed in the endnote.

Notes

¹ See, for example, Office of Management and Budget, *What Constitutes Strong Evidence of Program Effectiveness*, http://www.whitehouse.gov/omb/part/2004_program_eval.pdf, 2004; U.S. Department of Education, “Scientifically-Based Evaluation Methods: Notice of Final Priority,” *Federal Register*, vol. 70, no. 15, January 25, 2005, pp. 3586-3589; the Food and Drug Administration’s standard for assessing the effectiveness of pharmaceutical drugs and medical devices, at 21 C.F.R. §314.12; Institute of Medicine (IOM), *Knowing What Works in Health Care: A Roadmap for the Nation*, The National Academies Press, 2008; “Criteria for Evaluating Treatment Guidelines,” *American Psychologist*, vol. 57, no. 12, December 2002, pp. 1052-1059; and *Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination*, Society for Prevention Research, April 12, 2004, at <http://www.preventionresearch.org/sofetext.php>.

² See, for example, Manpower Demonstration Research Corporation, *National Evaluation of Welfare-to-Work Strategies: How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs* (U.S. Department of Health and Human Services and U.S. Department of Education, November 2001).

³ See, for example, statement of Ron Haskins, who in 1996 was the staff director of the House Ways and Means Subcommittee with jurisdiction over the welfare reform bill, in *Rigorous Evidence: The Key to Progress in Education? Lessons from Medicine, Welfare and Other Fields*, Coalition for Evidence-Based Policy, Council for Excellence in Government, November 18, 2002, pp. 67-69.

⁴ For a useful summary of the scientific evidence on which studies designs are most likely to produce valid estimates of an intervention’s effect, see Office of Management and Budget, *What Constitutes Strong Evidence of Program Effectiveness*, http://www.whitehouse.gov/omb/part/2004_program_eval.pdf, 2004, pp. 4-8.

⁵ Robert G. St. Pierre et. al., “Improving Family Literacy: Findings From the National Even Start Evaluation,” Abt Associates, September 1996.

⁶ Howard S. Bloom, Charles Michalopoulos, and Carolyn J. Hill, “Using Experiments to Assess Nonexperimental Comparison-Groups Methods for Measuring Program Effects,” in *Learning More From Social Experiments: Evolving Analytic Approaches*, Russell Sage Foundation, 2005, pp. 173-235. 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,” *Review of Economics and Statistics*, vol. 86, no. 1, 2004, pp. 180-194. 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, pp. 63-93.

⁷ 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.

⁸ Robert J. Temple, Director of the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration, quoted in Gary Taubes and Charles C. Mann, “Epidemiology Faces Its Limits,” *Science*, vol. 269, issue 5221, July 14, 1995, p. 169.

⁹ Liza M. Conyers, Arthur J. Reynolds, and Suh-Ruu Ou, “The Effect of Early Childhood Intervention and Subsequent Special Education Services: Findings from the Chicago Child-Parent Centers,” *Educational Evaluation and Policy*

Analysis, vol. 25, no. 1, Spring 2003, pp. 75-95. This study follows the principles outlined in this paper, with one exception: while the intervention and comparison groups were well-matched in many important predictors of educational outcomes (e.g., poverty status, single-parent household), they were not matched on pre-program measures of educational achievement. This may be because the sample members were 3 or 4 years old when they entered the study, and measuring the achievement of children that age is not straightforward. Preferably, however, the study would have made sure that the two groups were equivalent in early language skills, or similar measures.

Appendix

The Application of Randomized Controlled Trials: Where They Are Possible, and Where They are Not

The following is an excerpt from Office of Management and Budget, *What Constitutes Strong Evidence of Program Effectiveness*, http://www.whitehouse.gov/omb/part/2004_program_eval.pdf, 2004.

* * *

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/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.mdr.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.

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.