Rigorous Program Evaluations on a Budget:

How Low-Cost Randomized Controlled Trials Are Possible in Many Areas of Social Policy

Coalition for Evidence-Based Policy
A NONPROFIT, NONPARTISAN ORGANIZATION

March 2012

This publication was produced by the nonprofit, nonpartisan Coalition for Evidence-Based Policy, with funding support from the MacArthur Foundation and the William T. Grant Foundation.

We welcome comments and suggestions on this document (jbaron@coalition4evidence.org).

Purpose and Background

Overview: The increasing ability of social policy researchers to conduct randomized controlled trials (RCTs) at low cost could revolutionize the field of performance-based government. RCTs are widely judged to be the most credible method of evaluating whether a social program is effective, overcoming the demonstrated inability of other, more common methods to produce definitive evidence. In recent years, researchers have shown it is often possible to conduct high-quality RCTs at low cost, addressing a key obstacle to their widespread use. Costs are reduced by measuring study outcomes with administrative data already collected for other purposes (e.g., student test scores, criminal arrests, health care expenditures). These developments make it possible now, more than ever before, for policy officials to use scientific evidence about "what works" to increase government effectiveness.

Purpose of this document: To illustrate the feasibility and value of low-cost RCTs for policy officials and researchers, by providing concrete examples from diverse program areas. This paper summarizes five well-conducted, low-cost RCTs, carried out in real-world community settings. Study costs range from $50,000 to $300,000, with random assignment itself comprising only a small portion of this cost (between $0 and $20,000). The studies all produced valid evidence that is of policy and practical importance.

Background on RCTs: When well-conducted, RCTs are regarded as the strongest method of evaluating program effectiveness, per evidence standards articulated by the National Academy of Sciences,[1] Congressional Budget Office,[2] Institute of Education Sciences,[3] U.S. Preventive Services Task Force,[4] Food and Drug Administration,[5] and other respected scientific bodies.

- ***Definition of RCT***: A study that randomly assigns individuals or other units (such as schools or counties) to one group that is eligible to participate in a program, or to a "control group" that is not. The unique advantage of this process is that, with a sufficiently large sample, it ensures the two groups are highly similar in both observed and unobserved characteristics (e.g., demographics, motivation). Thus, any difference in outcomes between the groups can confidently be attributed to the program and not to other factors.

- ***Other, more common methods, while providing valuable information, cannot by themselves produce definitive evidence***. Quasi-experiments, outcome-based performance metrics, and randomized studies with small samples or other limitations are valuable for identifying potentially effective programs that merit more rigorous testing, but alone cannot produce strong evidence. Careful investigations show that, too often, their findings are not confirmed in sizable, high-quality RCTs, because they are unable to control reliably for all factors other than program participation.[6] Thus, a recent National Academies report recommends that evidence of program effectiveness generally "cannot be considered definitive" without ultimate confirmation in well-conducted RCTs, "even if based on the next strongest designs."[1]

Background on cost: Often seen as expensive, RCTs can now be conducted at low cost in many cases.

- ***Costs are reduced by measuring outcomes using administrative data already collected for other purposes*.** Traditionally, RCTs' largest cost has been data collection: locating program and control group members at various times over the course of the study, and administering tests or interviews to obtain their outcome data. If, instead, key outcomes for the two groups can be measured with administrative data that public agencies already collect for other purposes (e.g., student test scores, criminal arrests, health care expenditures), the cost of data collection – and often the study itself – can be dramatically reduced. Thus, as the quality and breadth of administrative databases have increased over time, opportunities to do low-cost RCTs have greatly expanded.

- ***To do a low-cost RCT, certain conditions must exist***, such as (i) low-cost access to administrative data of reasonable quality to measure key outcomes; (ii) a sizable number of individuals (or other units) available to participate in the study without special recruitment efforts; and (iii) the approval of key policy or program officials to do a randomized study. When carried out, low-cost RCTs may not be able to examine all questions studied in a traditional RCT, such as the program's effect on outcomes other than those measurable with administrative data; the nature of any services received by the control group and how they differ from those delivered by the program; and possible reasons *why* a program has (or does not have) an effect. But, increasingly, they can be used to answer the fundamental policy question: does the program actually improve people's lives?

* * *

Criminal Justice Example: Hawaii's Opportunity Probation with Enforcement (HOPE) Program

I. Program description: HOPE is a supervision program for drug-involved probationers that provides swift and certain sanctions for a probation violation. The program requires probationers to appear before a judge, who issues a clear warning in open court that any probation violation – including a failed drug test or failure to show up for a probation appointment – will result in immediate jail time. The probationer is then frequently drug tested at random times during probation. If the probationer commits a probation violation, he or she is arrested and jailed briefly – usually for a few days – after which the probationer resumes participation in HOPE. Probationers can request a drug treatment referral at any time, and repeat violators are mandated to drug treatment.

II. Evaluation method: Randomized controlled trial (RCT) with follow-up one year after random assignment.[7] The study sample was 493 probationers in Honolulu identified as being at high risk of violating probation. The probationers were randomly assigned to either a group that participated in HOPE during their probation term or a control group that received usual probation – i.e., a monthly drug test and appointment with a probation officer, with no automatic sanctions for probation violations. The study meets widely-accepted criteria for a well-conducted RCT.[8]

III. Key findings on program impact: During the one-year follow-up period, HOPE group members (compared to control group members) –

- Were 55% less likely to be re-arrested (21% of the HOPE group were re-arrested during the one-year follow-up versus 47% of the control group);
- Were sentenced to 48% fewer days of incarceration (138 days vs. 267 days);
- Had a much lower rate of failed drug tests – HOPE group members failed an average of 13% of their drug tests, compared to 46% for members of the control group; and
- Were much less likely to have their probation revoked (7% vs. 15%).

All of these impacts were statistically significant at the 0.05 level.

IV. Cost of measuring program impact: About $150,000.[a] The low cost was achieved by measuring study outcomes using administrative data (e.g., arrest records) that the state already collects for other purposes, rather than engaging in costly new data collection (e.g., surveys).

The $150,000 cost included such items as (i) meeting with judges to gain their support for the study; (ii) hiring part-time employees through the state attorney general office to extract and prepare the administrative data for analysis; and (iii) analyzing and reporting study results. The study's lead author says that substantial work was needed to ensure the administrative data's accuracy and usability, and that if cleaner data had been available, the cost of measuring program impact on the above outcomes would have been about $100,000.

Random assignment contributed only a small portion of the study's overall cost ($10,000- $15,000). Most of the costs described above, such as data extraction/preparation, would have been incurred even if the study had used a non-randomized design. Gaining judges' agreement with random assignment did not require major effort, because a lead judge supported a randomized design and helped bring fellow judges on board. In addition, researchers were able to use administrative data to monitor whether judges were complying with probationers' randomized condition (e.g., providing warning hearings and frequent drug tests to the HOPE group but not the control group).

[a] This does not include the cost of program delivery, but rather is the added ("marginal") cost of the evaluation. The study also included a process evaluation, which.is not included in the above cost.

Child Welfare Example: Recovery Coaches for Substance-Abusing Parents

I. **Program description: This program provides case management services of a Recovery Coach to substance-abusing parents who have temporarily lost custody of their children to the state.** The Recovery Coach works with the parent, child welfare caseworker, and substance-abuse treatment agencies to (i) remove barriers to treatment, (ii) engage the parent in treatment, (iii) provide outreach to re-engage the parent if necessary, and (iv) provide ongoing support to the parent and family for the duration of the child welfare case.

II. **Evaluation method: A large randomized controlled trial (RCT) in Illinois, conducted over 2000-2009.**[9] 60 child welfare agencies, working with 2,763 eligible parents, were randomly assigned to (i) a group that provided Recovery Coach services, or (ii) a group that provided services as usual. The study The study meets widely-accepted criteria for a well-conducted RCT.[8]

III. **Key findings on program impact:** Effects of the program an average of three to five years after families entered the study (compared to the control group)[a] –

- 15% increase in the likelihood children returned home to live with their parent (31% of program group children returned home vs. 27% of control group children);
- 14% increase in the likelihood of children's foster care cases being closed within three years (50% vs. 44%); and
- 29% reduction in the likelihood of the mothers delivering a substance-exposed infant (15% vs. 21%).

Through higher reunification rates and quicker case resolution, the program was found to produce net savings to Illinois of $6.7 million – or $2,400 per parent – over the course of the study.

The study found no significant difference between the two groups in the percent of children experiencing a new child maltreatment allegation, suggesting the above effects were achieved without adversely affecting children's safety.

IV. **Cost of measuring program impact: About $100,000.**[b] The low cost was achieved by measuring study outcomes using administrative data (e.g., on foster care closure rates) that the state already collects for other purposes, rather than engaging in costly new data collection (e.g., surveys).

The $100,000 cost included such items as (i) obtaining access to the state administrative data and preparing it for analysis; (ii) hiring a part-time person to monitor that parents stayed within their randomly assigned group; and (iii) analyzing and reporting study results. The costs of persuading program officials to agree to random assignment were small, because Illinois' Director of Child Welfare supported a randomized design based on a successful earlier RCT the state had conducted.

Random assignment contributed only a small portion of the study overall cost – about $5000. Most of the costs described above, such as gaining access to state data, would have been incurred even if the study had used a non-randomized design.

[a] Based on our review of the study reports, the first two impacts (on family reunification and foster care case closure) appear to be statistically significant at the 0.05 level in tests that account for the fact that groups, rather than individuals, were randomly assigned. The third impact (on substance-exposed births) may not be.

[b] This does not include the cost of program delivery, but rather is the added ("marginal") cost of the evaluation. The study cost also does not include a process analysis in which researchers interviewed child welfare caseworkers.

Example of a Community-Wide Intervention: The Triple P (Positive Parenting Program) System

I. Program description: A system of parenting interventions for families with children ages 0-8,[a] **which seeks to strengthen parenting skills and prevent child maltreatment.** System services include various combinations of parenting seminars, parent skills-training sessions, and individual consultations, with the type and amount of service (i.e., service "levels") tailored to the severity of each family's dysfunction and child behavioral problems. Sessions are delivered by a variety of trained service providers from different settings (e.g., healthcare, preschools, elementary schools, mental health, social services). The system also includes media strategies promoting positive parenting practices community-wide (e.g., news stories, radio announcements).

II. Evaluation method: Randomized controlled trial (RCT) evaluating county-wide implementation of Triple P in a sample of 18 rural and semi-urban South Carolina counties.[10] The trial randomly assigned these counties to (i) a group that implemented the Triple P System county-wide; or (ii) a control group that provided usual county services without implementation of Triple P. The study meets widely-accepted criteria for a well-conducted RCT.[8]

III. Key findings on program impact: Effects of Triple P two years after random assignment (compared to the control counties) –

- 25% reduction in the rate of substantiated child maltreatment (11.6 substantiated cases each year per 1,000 children in Triple P counties vs. 15.5 cases in control counties);
- 33% reduction in the rate of out-of-home placements – e.g., in foster homes (3.4 placements each year per 1,000 children vs. 5.1 placements); and
- 35% reduction in the rate of hospitalizations or emergency room visits for child maltreatment injuries (1.3 each year per 1,000 children age vs. 2.0).

The first two impacts were statistically significant at the 0.05 level, the third at the 0.10 level.

IV. Cost of measuring program impact: $225,000-$300,000.[b] The low cost was achieved by measuring study outcomes using administrative data (e.g., child maltreatment records) that the state already collects for other purposes, rather than engaging in costly new data collection (e.g., surveys).

The above cost included such items as (i) paying a programmer to extract the administrative data and prepare it for analysis; (ii) paying a statistician to conduct the analyses; and (iii) reporting study results.

Random assignment's contribution to the above cost was negligible. Because counties (rather than individuals) were randomly assigned, random assignment required neither the assent of policy officials nor monitoring to ensure its integrity. The researchers simply assigned the 18 counties in the sample randomly to the treatment versus control group, and began implementing Triple P in the treatment counties. Thus, the study costs described above would have been incurred even if the study had used a non-randomized design.

[a] Other versions of Triple P serve children up to age 12 and/or provide specific components ("levels") of Triple P, but not the full System.

[b] This does not include the cost of program delivery, but rather is the added ("marginal") cost of the evaluation. The study also included phone surveys as part of a process analysis, which are not included in the above cost.

K-12 Education Example: New York City Teacher Incentive Program

I. **Program description: This program provided low-performing schools that increased student achievement and other key outcomes with an annual bonus, to be distributed to teachers.** The program, funded at $75 million, was offered to a subset of low-performing public schools (elementary, middle, and high) in New York City, more than 80 percent of which participated for three years. Each year, approximately half to three-quarters of the schools were awarded the full bonus (approximately $3000 per teacher), and a small additional number received a partial bonus.

II. **Evaluation method: Large randomized controlled trial (RCT) with a sample of 396 of the city's lowest-performing schools, conducted 2008-2010.**[11] The trial randomly assigned these schools to (i) a group that was offered the incentive program, or (ii) a control group that was not. The study meets widely-accepted criteria for a well-conducted RCT.[8]

III. **Key findings on program impact: During the three years after random assignment, the program had no effect on student achievement, attendance, graduation rates, behavior, or GPA** (compared to control schools). Based in part on these results, the city ended the program, freeing up resources for other efforts to improve student outcomes.

IV. **Cost of measuring program impact: About $50,000.**[a] The low cost was achieved by measuring study outcomes using administrative data (e.g., state test scores) that the school district already collects for other purposes, rather than engaging in costly new data collection (e.g., administering achievement tests as part of the study).

The $50,000 cost included such items as (i) extracting the administrative data and preparing it for analysis; (ii) analyzing and reporting study results; and (iii) paying overhead costs of the study author's research institution.

Random assignment's contribution to the above cost was negligible. This is because (i) district officials recognized that they did not have sufficient funds to provide the incentive program to all low-performing schools in the city, and therefore were amenable to allocating program participation by lottery (i.e., random assignment); and (ii) the study author personally conducted the lottery at the start of the study, eliminating the need for ongoing supervision of random assignment to ensure its integrity. Thus, the study costs described above, such as data extraction/preparation, would have been incurred even if the study had used a non-randomized design.

[a] This does not include the cost of program delivery – e.g., the bonus payments – but rather is the added ("marginal") cost of the evaluation. The study also included phone surveys as part of a process analysis, which are not included in the above cost.

Criminal Justice Example: Philadelphia Low-Intensity Community Supervision Experiment

I. **Program description: Low-Intensity Community Supervision for probationers or parolees at low risk of committing a serious crime (compared to the usual, more intensive/costly supervision).** This is a program in Philadelphia for criminal offenders on probation or parole identified as being at low risk of committing a serious crime, based on prior criminal behavior and other factors. These offenders receive low-intensity supervision by a probation officer (averaging 2.4 visits with the officer per year), instead of the usual higher-intensity supervision (averaging 4.5 visits per year). The purpose is to reduce the cost of supervision to Philadelphia County. Based on the encouraging study findings below, the county has adopted this approach for all low-risk offenders.

II. **Evaluation method: Randomized controlled trial (RCT) with follow-up one year after random assignment.**[12] This study randomly assigned 1,559 offenders on parole (from a short sentence in county jail) or probation in 2007-2008 to either (i) a treatment group that received low-intensity supervision or (ii) a control group that received standard supervision. The sample members were identified via a computer algorithm as being roughly in the lower half of active cases in likelihood of committing a serious crime. The study meets widely-accepted criteria for a well-conducted RCT.[8]

III. **Key findings on program impact:** During the one-year follow-up period, there were no statistically-significant differences between the treatment and control groups in –

- prevalence or frequency of new criminal charges for any offense, including violent offenses; or
- likelihood of incarceration in county jail, or days incarcerated.

There was also no pattern of *non-significant* differences between the two groups in these outcomes. These findings – i.e., no increase in crime – suggest the program is a viable way to reduce costs in the criminal justice system.

IV. **Cost of measuring program impact: Less than $100,000.**[a] The low cost was achieved by measuring study outcomes using administrative data (e.g., arrest records) that the county already collects for other purposes, rather than engaging in costly new data collection (e.g., surveys).

The above study cost included such items as (i) extracting the administrative data and preparing it for analysis; (ii) having a research assistant in the probation department monitor random assignment to ensure it was carried out faithfully; and (iii) analyzing and reporting study results. County officials had actually suggested an RCT, so no effort was needed to gain their support for the study design. (The officials felt that an RCT could generate evidence as to the program's safety that would be convincing in the public arena, and help defuse potential controversy over this type of reform.)

Random assignment contributed only a small portion of the study's overall cost (approximately $15,000-$20,000). Most of the costs described above, such as data extraction/preparation, would have been incurred even if the study had used a non-randomized design.

[a] This does not include the costs of developing and implementing the program – e.g. the algorithm for identifying low-risk offenders – but rather is the added ("marginal") cost of the evaluation.

References

[1] National Research Council and Institute of Medicine, *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities*, Mary Ellen O'Connell, Thomas Boat, and Kenneth E. Warner, Editors (Washington DC: National Academies Press, 2009), recommendation 12-4, p. 371, linked here.

[2] *CBO's Use of Evidence in Analysis of Budget and Economic Policies*, Jeffrey R. Kling, Associate Director for Economic Analysis, November 3, 2011, page 31, linked here.

[3] U.S. Department of Education's Institute of Education Sciences, *What Works Clearinghouse - Procedures and Standards Handbook*, Version 2.1, September 2011, pp. 11-12, linked here. 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, linked here.

[4] U.S. Preventive Services Task Force, "Current Methods of the U.S. Preventive Services Task Force: A Review of the Process," *American Journal of Preventive Medicine*, vol. 20, no. 3 (supplement), April 2001, pp. 21-35.

[5] The Food and Drug Administration's standard for assessing the effectiveness of pharmaceutical drugs and medical devices, at 21 C.F.R. §314.126, linked here.

[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.

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.

Thomas D. Cook, William R. Shadish, and Vivian C. Wong, "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons," *Journal of Policy Analysis and Management*, vol. 27, no. 4, 2008, pp. 724-50.

John P. A. Ioannidis, "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association*, vol. 294, no. 2, July 13, 2005, pp. 218-228.

Mohammad I. Zia, Lillian L. Siu, Greg R. Pond, and Eric X. Chen, "Comparison of Outcomes of Phase II Studies and Subsequent Randomized Control Studies Using Identical Chemotherapeutic Regimens," *Journal of Clinical Oncology*, vol. 23, no. 28, October 1, 2005, pp. 6982-6991.

John K. Chan et. al., "Analysis of Phase II Studies on Targeted Agents and Subsequent Phase III Trials: What Are the Predictors for Success," *Journal of Clinical Oncology*, vol. 26, no. 9, March 20, 2008.

[7] Angela Hawken and Mark Kleiman, "Managing Drug Involved Probationers with Swift and Certain Sanctions: Evaluating Hawaii's HOPE." Dec 2009. *Submitted to the National Institute of Justice*.

[8] The study meets widely-accepted criteria for a well-conducted RCT (summarized in the Coalition's RCT Checklist), such as: (i) the program and control groups were highly similar in their pre-program characteristics; (ii) the study had no sample attrition (because the use of administrative made it possible to measure outcomes for all members of the program and control groups); (iii) the study appropriately measured outcomes for all individuals assigned to the program group, regardless of whether or how long they participated in the program (i.e., the study used an "intention-to-treat" analysis); (iv) study outcomes were assessed with valid measures; (v) where appropriate, the study's statistical analysis accounted for the fact that groups, rather than individuals, were randomly assigned; and (vi) the study evaluated the program as delivered in a typical community setting, thus providing evidence of its effectiveness under real-world implementation conditions.

[9] *Profiles of the Title IV-E Child Welfare Waiver Demonstration Projects*, James Bell Associates: Arlington, Virginia, June 2010, pp. 52-56. Mark F. Testa, Joseph P. Ryan, Pedro M. Hernandez, and Hui Huan, *Illinois AODA IV-E Waiver Demonstration Interim Evaluation Report*, University of Illinois at Urban-Champaign, School of Social Work, Children and Family Research Center, September 2009. Joseph P. Ryan, Jeanne Marsh, Mark Testa, and Richard Louderman, "Integrating Substance Abuse Treatment and Child Welfare Services: Findings from the Illinois AODA Waiver Demonstration," *Social Work Research*., vol. 30, no. 2, 2006, pp. 95-107. Joseph P. Ryan,

Sam Choi, Jun Sung Hong, Pedro Hernandez, and Christopher R. Larrison, "Recovery Coaches and Substance Exposed Births: An Experiment in Child Welfare," *Child Abuse & Neglect*, vol. 32, 2008, pp. 1072-1079.

[10] Ronald J. Prinz, Matthew R. Sanders, Cheri J. Shapiro, Daniel J. Whitaker, and John R. Lutzker. "Population-Based Prevention of Child Maltreatment: The U.S. Triple P System Population Trial," *Prevention Science*, 2009, vol. 10, no. 1, pp. 1-12. Michael E. Foster, Ronald J. Prinz, Matthew R. Sanders, and Cheri J. Shapiro, "The Costs of Public Health Infrastructure for Delivering Parenting and Family Support, " *Children and Youth Services Review*, 2008, vol. 30, pp. 493-501. Clarifying correspondence with Ronald Prinz (August 2009)

[11] Roland G. Fryer, "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," NBER Working Paper No. 16850, March 2011, linked here.

[12] Geoffrey C. Barnes, Lindsay Ahlman, Charlotte Gill, Lawrence W. Sherman, Ellen Kurtz, and Robert Malvestuto. "Low-Intensity Community Supervision for Low-Risk Offenders: a Randomized, Controlled Trial," *Journal of Experimental Criminology*, vol. 6, no. 2, pp. 159-189.