Backgrounder
No. 2578
July 18, 2011
Published by The Heritage Foundation

Evaluating Federal Social Programs: Finding Out What Works and What Does Not

David B. Muhlhausen, Ph.D.

Abstract: *Federal social programs are rarely evaluated to determine whether they are actually accomplishing their intended purposes. As part of its obligation to spend taxpayers' dollars wisely, Congress should mandate that experimental evaluations of every federal social program be conducted. The evaluations should be large-scale, multisite studies to guard against mistakenly assuming that a program that works in one location or with one population will automatically work in other situations. Congress should place substantially less emphasis on funding evaluations based on less rigorous types of research designs, because their conclusions are much less reliable. Finally, Congress should exercise strict oversight to ensure that the evaluations are conducted and the results reported in a timely manner.*

The notion that public policy should be informed by social science has gained widespread acceptance. The evaluation of federal social programs, using scientific techniques, offers policymakers and the public ample opportunities to learn about the effectiveness of government programs. Despite the availability of evaluation methods, the effectiveness of federal social programs is often unknown in far too many cases. Many programs operate for decades without ever undergoing thorough scientific evaluations.

With the enormous federal debt increasingly shaping policy debates in Washington, D.C., Congress should subject all federal social programs to rigorous evaluations to determine what works and what does not work.

Talking Points

- The evaluation of federal social programs, using scientific techniques, offers ample opportunities for policymakers and the public to learn about the effectiveness of government programs.
- Despite the ready availability of evaluation methods, the effectiveness of federal social programs is often unknown in far too many cases. Many programs exist for decades without ever undergoing thorough scientific evaluations.
- Congress should mandate large-scale, multisite experimental evaluations of every federal social program that it funds.
- Experimental evaluations, which randomly assign individuals to the intervention and control groups, are the "gold standard" of evaluation designs.
- With the growing national debt increasingly shaping policy debates in Washington, D.C., Congress should subject all social programs to rigorous evaluations to determine what works and what does not work.
- Implementation of rigorous impact evaluation offers Congress excellent opportunities to exercise oversight of government programs.

This paper, in its entirety, can be found at:
http://report.heritage.org/bg2578

Produced by the Center for Data Analysis

Published by The Heritage Foundation
214 Massachusetts Avenue, NE
Washington, DC 20002–4999
(202) 546-4400 • heritage.org

The Heritage Foundation
LEADERSHIP FOR AMERICA

Evidence-Based Policy

The social sciences can make important contributions to policymaking. Perhaps the greatest contribution is the evidence-based policy movement that seeks to inform and influence policymakers through scientifically rigorous evaluations of the effectiveness of government programs.[1] The evidence-based policy movement, in other words, seeks to inform policymakers about what works and what does not work.

Scientifically rigorous impact evaluations are necessary to determine whether these programs actually produce their intended effects. Thus, the implementation of rigorous impact evaluation offers policymakers excellent opportunities to exercise oversight of government programs. Policymakers are shirking their responsibilities to taxpayers if they continue to fund social programs that are not known to work or that do not work at all. Obviously, there is little merit in continuing programs that fail to ameliorate their targeted social problems.

However, there is disagreement over what can be counted as evidence.[2] For example, should high-quality quasi-experiments be given the same level of scientific credibility as experimental evaluations? Despite such disagreements, this paper argues that experimental evaluations are the most credible and accurate method by which to assess effectiveness.

The Advantages of Experimental Evaluations

The impact of programs cannot be estimated with 100 percent certainty. All such impact evaluations face formidable control problems that make successful estimates difficult. As a general rule, the more rigorous the research methodology is, the more reliable the evaluation's findings are.

Determining the impact of social programs requires comparing the conditions of those who had received assistance with the conditions of an equivalent group that did not experience the intervention. However, evaluations differ by the quality of methodology used to separate the net impact of programs from other factors that may explain differences in outcomes between comparison and intervention groups.

Experimental evaluations are the "gold standard" of evaluation designs.

Broadly speaking, there are three types of research designs: experimental designs, quasi-experimental designs, and nonexperimental designs.[3] Experimental evaluations, often called randomized field or control trials, randomly assign individuals to the intervention and control groups.

Experimental evaluations are the "gold standard" of evaluation designs. Random assignment helps to ensure that the control group is equivalent to the intervention group in composition, predispositions, and experiences.[4] In other words, randomization eliminates any systematic association between intervention status and the observed and unobserved participant characteristics, thus largely eliminating the selection bias that potentially contaminates other evaluation designs.[5] Weaker evaluation designs are often plagued by unobserved

1. See Karen Bogenschneider and Thomas J. Corbett, *Evidence-Based Policymaking: Insights from Policy-Minded Researchers and Research-Minded Policymakers* (New York: Routledge, 2010).
2. Stewart I. Donaldson, "In Search of the Blueprint for an Evidence-Based Global Society," in Stewart I. Donaldson, Christina A. Christie, and Melvin M. Mark, eds., *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* (Thousand Oaks, Cal.: SAGE Publications, 2009), pp. 2–18.
3. William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin Company, 2002). A fourth research design is the natural experiment. Natural experiments use naturally occurring differences between intervention and comparison groups.
4. Peter H. Rossi, Mark W. Lipsey, and Howard E. Freeman, *Evaluation: A Systematic Approach*, 7th ed. (Thousand Oaks, Cal.: SAGE Publications, 2004).
5. Gary Burtless, "Randomized Field Trials for Policy Evaluations: Why Not in Education?" in Frederick Mosteller and Robert Boruch, eds., *Evidence Matters: Randomized Trials in Education Research* (Washington, D.C.: The Brookings Institution, 2002), pp. 179–197.

differences between the intervention and control groups, which makes drawing reliable causal conclusions impossible.

Randomized evaluations ensure that pre-program differences between the intervention and control groups do not confound or obscure the true impact of the programs being evaluated. Random assignment allows the evaluator to test for differences between the experimental and control groups that are due to the intervention and not to pre-intervention discrepancies between the groups. Because they draw members of the interaction and comparison groups from the same pool of eligible participants, these experimental evaluations are superior to other evaluations that use weaker designs.

Congress has seldom supported experimental evaluation of federally funded grant programs.

In addition, this design's methodology is easier to describe to policymakers and laymen than other evaluation methods that use sophisticated statistical modeling techniques,[6] which often have significant weaknesses in determining program impact. Further, the results of an evaluation using sophisticated statistical modeling techniques can "become entangled in a protracted and often inconclusive scientific debate about whether the findings of a particular study are statistically valid." Alternatively, the results of experimental evaluations are more straightforward and can be easily grasped: "Compared to the control group, the intervention group that participated in the program experienced a 10% increase in the outcome measure."[7]

In both quasi-experimental and nonexperimental designs, failure to remove the influence of differences that affect program outcomes leaves open the possibility that the underlying differences between the groups, not the program, caused the net impact. While quasi-experimental and nonexperimental designs use sophisticated techniques, experimental evaluations are still considered better at producing reliable estimates of program effects. Evidence in criminal justice policy indicates that quasi-experimental and nonexperimental evaluations are more likely to find favorable intervention effects and less likely to find harmful intervention effects.[8]

Given that experimental evaluations produce the most reliable results, Congress should promote the use of experimental evaluations to assess the effectiveness of federal programs. Congress has a responsibility to ensure that experimental evaluations are used to assess the impact of federal social programs. Quasi-experimental and nonexperimental designs, no matter how well designed, may be incapable of controlling for non-program factors that influence how participants respond to the intervention.

Given the importance of criminal justice policy, Professor David Weisburd of George Mason University argues that researchers have a moral imperative to conduct randomized experiments[9] because of their "obligation to provide valid answers to questions about the effectiveness of treatments, practices, and programs."[10] This moral imperative also applies to Congress, which spends hundreds of billions of dollars on social programs. Yet Congress has seldom supported experimental evaluation of federally funded grant programs.

6. *Ibid.*
7. *Ibid.*, p. 183.
8. After conducting a meta-analysis of 308 criminal justice program evaluations, Professor David Weisburd of George Mason University and his colleagues found that weaker evaluation designs are more likely to find favorable intervention effects and less likely to find harmful intervention effects. They caution that quasi-experimental and non-experimental designs, no matter how well designed, may be incapable of controlling for the unobserved factors that make individuals more likely to respond favorably to the intervention. See David Weisburd, Cynthia M. Lum, and Anthony Petrosino, "Does Research Design Affect Study Outcomes in Criminal Justice?" *Annals of the American Academy of Political and Social Sciences*, No. 578 (November 2001), pp. 50–70.
9. David Weisburd, "Ethical Practice and Evaluation of Interventions in Crime and Justice," *Evaluation Review*, Vol. 27, No. 23 (June 2003), pp. 336–354.
10. *Ibid.*, p. 350.

Major Experimental Evaluation of Federal Social Programs

Despite the trillions of dollars that Congress has spent on federal social programs, only a few programs have undergone large-scale experimental impact evaluations. These evaluations include:

- Negative Income Tax Experiments (1968–1978);
- National Health Insurance (1972–1982);
- Supported Work (1974–1980);
- MDRC Welfare to Work (1985–2001);
- National Job Training Partnership Act (1986–1993);
- Even Start (1991–1994);
- Upward Bound (1992–2004);
- Job Corps (1993–2003);
- Early Head Start (1996–present);
- Abstinence Education (1997–2007);
- Employment Retention and Advancement (2000–2007);
- Head Start (2002–2008); and
- Building Strong Families (2002–2011).

The welfare-to-work evaluations of job training and job search programs for Aid to Families with Dependent Children (AFDC) participants were highly influential in efforts to reform the nation's welfare system during congressional debates on the Family Support Act of 1998 and helped to pave the way for further reforms that occurred with the Personal Responsibility and Work Opportunity Reconciliation Act of 1996.[11]

Can Effective Programs Be Replicated?

Policymakers and advocates often assume that a social program that is effective in one setting will automatically produce the same results in other settings. Policymakers should not make this assumption.

Many advocates of social programs have adopted the language of the "evidence-based" policy movement. Under the evidence-based policy movement, programs found to be effective using rigorous scientific methods are deemed "effective" or "evidence-based" and held up as "model" programs.

However, many of the programs labeled as "evidence-based"—often by program advocates—have been evaluated in only a single setting, so the results cannot necessarily be generalized to other settings. In addition, these evidence-based programs have often been implemented by highly trained professionals operating under ideal conditions. These programs are carefully monitored to ensure that the participants receive the intended level of treatment. In the real world, program conditions are often much less than optimal.

The success of replicating evidence-based programs often depends on implementation fidelity—the degree to which programs follow the theory underpinning the program and how correctly the program components are put into practice. Incorrect implementation often accounts for the failures of previously successful or model programs when implemented in other jurisdictions.

Reconnecting Youth. A good example of a "successful" program that has not been found to be effective when replicated in the real world is Reconnecting Youth, a school-based substance abuse program. Reconnecting Youth was designated as a "model program" by the Substance Abuse and Mental Health Services Agency (SAMHSA)[12] and as a "research-based" program by the National Institute on Drug Abuse.[13] These classifications are important because schools receiving Safe and Drug-Free Schools and Communities grants under the No Child Left Behind Act of 2001 must select only drug

11. Ron Haskins, "Congress Writes a Law: Research and Welfare Reform," *Journal of Policy Analysis and Management*, Vol. 10, No. 4 (1991), pp. 616–632; David Greenberg, Donna Linksz, and Marvin Mandell, *Social Experimentation and Public Policymaking* (Washington, D.C.: Urban Institute Press, 2003); and Judith M. Gueron and Edward Pauly, *From Welfare to Work* (New York: Russell Sage Foundation, 1991).
12. Steven Schinke, Paul Brounstein, and Stephen E. Gardner, *Science-Based Prevention Programs and Principles, 2002*, U.S. Department of Health and Human Services, Center for Substance Abuse Prevention, Substance Abuse and Mental Health Services Administration, 2002, at *http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED474651* (June 29, 2011).

prevention programs that have been previously designated as effective.[14]

Denise Hallfors, a senior research scientist at the Pacific Institute for Research and Evaluation, and her colleagues evaluated the effectiveness of Reconnecting Youth in real-world conditions.[15] In a random experiment, 1,370 high-risk youths in nine high schools in two large urban school districts were assigned to intervention and control groups. Overall, Reconnecting Youth had no effect on academic performance, truancy, or substance abuse. However, Reconnecting Youth participants showed statistically significant decreases in conventional peer bonding and pro-social weekend activities (e.g., doing homework, club or church activities, and family activities) and a statistically significant increase in high-risk peer bonding.[16]

Hallfors and her colleagues concluded that "Reconnecting Youth failed to meet the requirement to do more good than harm."[17] Further, programs found to be effective in a single location "do not provide adequate evidence for widespread dissemination or designation as 'model' programs."[18]

Many of the programs labeled as "evidence-based"—often by program advocates—have been evaluated in only a single setting, so the results cannot necessarily be generalized to other settings.

MST. Multisystemic therapy (MST) is another program model that has been labeled effective. It has shown promise in reducing the delinquency of youth who display serious antisocial behavior. As a highly intensive and tailored counseling program aimed at individuals, not groups, MST recognizes that antisocial behavior is influenced by three areas where youth interact: family, school, and peer associations.[19] Highly trained MST counselors work with parents, usually in the home, to improve discipline, enhance family relationships, increase youth interactions with pro-social peers, and improve school performance.[20]

Several randomized experiments by its developers have linked MST to reductions in offending by participants.[21] However, there is some debate about

13. Elizabeth B. Robertson, Susan L. David, and Suman A. Rao, *Preventing Drug Use Among Children and Adolescents: A Research-Based Guide for Parents, Educators, and Community Leaders*, National Institutes of Health, National Institute on Drug Abuse, October 2003, at *http://drugabuse.gov/pdf/prevention/RedBook.pdf* (June 29, 2011).
14. Denise Hallfors, Hyunsan Cho, Victoria Sanchez, Sheren Khatapoush, Hyung Min Kim, and Daniel Bauer, "Efficacy vs. Effectiveness Trial Results of an Indicated 'Model' Substance Abuse Program: Implications for Public Health," *American Journal of Public Health*, Vol. 96, No. 12 (December 2006), pp. 2254–2259. See also 20 U.S. Code §§ 7112 and 7115.
15. Hallfors *et al.*, "Efficacy vs. Effectiveness Trial Results of an Indicated 'Model' Substance Abuse Program."
16. *Ibid.*, p. 2257.
17. *Ibid.*, p. 2258.
18. *Ibid.*
19. Scott W. Henggeler, Gary B. Melton, and Linda A. Smith, "Family Preservation Using Multisystemic Therapy: An Effective Alternative to Incarcerating Serious Juvenile Offenders," *Journal of Consulting and Clinical Psychology*, Vol. 60, No. 6 (December 1992), pp. 953–961.
20. Cynthia Cupit Swenson, Scott W. Henggeler, Ida Taylor, and Oliver W. Addison, *Multisystemic Therapy and Neighborhood Partnerships: Reducing Adolescent Violence and Substance Abuse* (New York: Guilford Press, 2005).
21. Charles M. Borduin, Scott W. Henggeler, David M. Blaske, and Risa J. Stein, "Multisystemic Treatment of Adolescent Sexual Offenders," *International Journal of Offender Therapy and Comparative Criminology*, Vol. 34, No. 2 (September 1990), pp. 105–113; Charles M. Borduin, Barton J. Mann, Lynn T. Cone, Scott W. Henggeler, Bethany R. Fucci, David M. Blaske, and Robert A. Williams, "Multisystemic Treatment of Serious Juvenile Offenders: Long-Term Prevention of Criminality and Violence," *Journal of Consulting and Clinical Psychology*, Vol. 63, No. 4 (August 1995), pp. 569–578; Scott W. Henggeler, W. Glenn Clingempeel, Michael J. Bronding, and Susan G. Pickrel, "Four-Year Follow-Up of Multisystemic Therapy with Substance-Abusing and Substance Dependent Juvenile Offenders," *Journal of the American Academy of Child and Adolescent Psychiatry*, Vol. 41, No. 7 (July 2002), pp. 868–874; and Scott W. Henggeler, Gary B. Melton, and Linda A. Smith, "Family Preservation Using Multisystemic Therapy: An Effective Alternative to Incarcerating Serious Juvenile Offenders," *Journal of Consulting and Clinical Psychology*, Vol. 60, No. 6 (December 1992), pp. 953–961.

whether MST is truly effective and can be replicated successfully across the nation.

Professor Julia H. Littell of Bryn Mawr College and her colleagues have pointed out that some MST experimental evaluations have suffered from attrition in which subjects in the evaluation dropped out of treatment.[22] Evaluations, even random experiments, that exclude dropouts from outcome assessments may inadvertently engage in "creaming of the crop," in which those least likely to succeed drop out, leaving behind an intervention group composed of individuals most likely to succeed. This type of attrition breaks equivalence between the intervention and control groups, thus biasing the impact estimates.

Programs found to be effective in a single location "do not provide adequate evidence for widespread dissemination or designation as 'model' programs."

Further, the successful MST effects have yet to be replicated in other settings. An experimental evaluation of MST in Ontario, Canada, included intervention dropouts in its final outcome measures to avoid the problem of attrition. This evaluation, unbiased by attrition, found that MST failed to reduce delinquency.[23] In Norway, MST was found to be effective based on intermediate measures, but delinquency was not measured.[24]

After conducting a meta-analysis of MST, Littell and her colleagues concluded that "it is not clear whether MST has clinically significant advantages over other services."[25] While the debate over MST's effectiveness is not yet settled, evaluations suggest that MST has had little success when replicated in other settings.

These examples illustrate why programs should be evaluated in multiple settings before being labeled "evidence-based." Generalizing from a single evaluation conducted in one setting is at best premature.

What Congress Should Do

Congress can take several steps to ensure that federal social programs are properly assessed using experimental evaluations. The Appendix presents model legislative language that Congress could use to mandate experimental evaluation of the social programs that it authorizes and funds.

Step #1: When authorizing a new program or reauthorizing an existing program, Congress should specifically mandate experimental evaluation of the program.

Congressional mandates are necessary because federal agencies often resist performing experimental evaluations.

Local recipients of federal funding may also resist participating in experimental evaluations for a variety of reasons. They may not want to deny services to members of the control group or may not want the final results to reflect negatively on the program. For example, many jurisdictions receiving funding through the Job Training Partnership Act and Job Opportunities and Basic Skills programs refused to cooperate with large-scale experimental evaluations of these programs.[26]

Interest groups and some Members of Congress may also oppose experiments. For example, Upward Bound is a program intended to help disadvantaged

22. Julia H. Littell, Melanie Popa, and Burnee Forsythe, "Multisystemic Therapy for Social, Emotional, and Behavioral Problems in Youth Aged 10–17," *Campbell Systematic Reviews*, September 21, 2005.
23. Alan Leschied and Alison Cunningham, *Seeking Effective Interventions for Young Offenders: Interim Results of a Four-Year Randomized Study of Multisystemic Therapy in Ontario, Canada* (London, Ontario: Centre for Children and Families in the Justice System, 2002).
24. Terje Ogden and Colleen A. Halliday-Boykins, "Multisystemic Treatment of Antisocial Adolescents in Norway: Replication of Clinical Outcomes Outside of the US," *Journal of Child and Adolescent Mental Health*, Vol. 9, No. 2 (2004), pp. 77–83.
25. Littell *et al.*, "Multisystemic Therapy for Social, Emotional, and Behavioral Problems in Youth Aged 10–17."
26. Fred Doolittle and Linda Traeger, *Implementing the National JTPA Study* (New York: Manpower Demonstration Research Corporation, 1990), and Judith M. Gueron, "The Politics of Random Assignment: Implementing Studies and Affecting Policy," in Mosteller and Boruch, *Evidence Matters*, pp. 15–49.

high school students prepare for college. Many Upward Bound centers and the Council for Opportunity in Education (COE), which lobbies on behalf of Upward Bound centers, opposed Department of Education efforts under the George W. Bush Administration to conduct an experimental evaluation of Upward Bound.[27] A previous experimental evaluation found that Upward Bound had no impact on whether most participants attended college.[28]

Research suggests that Upward Bound serves a population that, while viewed as disadvantaged, is already very likely to attend college.[29] However, participants who originally had no expectation of attending college were more likely to enroll in college.[30] In response to these findings, the Bush Administration wanted to focus Upward Bound on students with low academic expectations, where the program appeared to be effective, and conduct a new experimental evaluation of the revised program's effectiveness.[31]

However, many Upward Bound centers opposed the policy change and the additional evaluation using random assignment to assess the revised program's effectiveness. COE President Arnold L. Mitchem compared the use of random assignment, which would ultimately deny some eligible students access to Upward Bound, to the infamous Tuskegee syphilis experiments, in which medical treatment was withheld from black men so that government scientists could learn about the negative effects of the disease.[32] In the end, a rider barring the Department of Education from using funds to perform the proposed evaluation was attached to the fiscal year 2008 omnibus appropriations law.[33]

As previously mentioned, Congress has the moral imperative to ensure that it allocates taxpayer dollars effectively. Experimental evaluations are the only way to determine to a high degree of certainty the effectiveness of social programs. Congress should not cave in to interest groups that are opposed to rigorous evaluation of their programs. Congress should mandate that all recipients of federal funding, if selected for participation, must cooperate with evaluations in order to receive future funding.

Step #2: The experimental evaluations should be large-scale, multisite studies.

When Congress creates programs, especially state and local grant programs, the funded activities are implemented in multiple cities or towns. Federal grants are intended to be spread out across the nation. For this reason, Congress should require national, multisite experimental evaluations of these programs.

While individual programs funded by federal grants may undergo experimental evaluations, these small-scale, single-site evaluations do not inform policymakers of the general effectiveness of national programs. Small-scale evaluations only assess the impact on a small fraction of people served by federal social programs. The success of a single program that serves a particular jurisdiction or population does not necessarily mean that the same program will achieve similar success in other jurisdictions or among different populations. Thus, small-scale evaluations are poor substitutes for large-scale evaluations.

27. Kelly Field, "Senate Votes to Block Upward Bound Evaluation," *The Chronicle of Higher Education*, November 2, 2007.
28. David Myers, Robert Olsen, Neil Seftor, Julie Young, and Christina Tuttle, *The Impacts of Upward Bound: Results from the Third Follow-Up Data Collection*, Mathematica Policy Research, 2004, at *http://www.eric.ed.gov/PDFS/ED483155.pdf* (April 29, 2011).
29. Neil Seftor, Arif Mamun, and Allen Schirm, *The Impacts of Regular Upward Bound on Postsecondary Outcomes 7–9 Years After Scheduled High School Graduation: Final Report*, Mathematica Policy Research, January 2009, at *http://www.policyarchive.org/handle/10207/bitstreams/15740.pdf* (April 29, 2011).
30. Myers *et al.*, *The Impacts of Upward Bound*.
31. Field, "Senate Votes to Block Upward Bound Evaluation."
32. Kelly Field, "Education Department Agrees to End Controversial Upward Bound Study," *The Chronicle of Higher Education*, February 25, 2008.
33. *Ibid.* and Consolidated Appropriations Act, 2008, Public Law 110–161, § 519, December 26, 2007.

In addition, a multisite experimental evaluation that examines the performance of a particular program in numerous and diverse settings can potentially produce results that are more persuasive to policymakers than results from a single locality.[34]

The case of police departments performing mandatory arrests in domestic violence incidents is a poignant example of why caution should be exercised when generalizing findings from a single evaluation. During the 1980s, criminologists Lawrence W. Sherman and Richard A. Berk, currently professors at the University of Pennsylvania, analyzed the impact of mandatory arrests for domestic violence incidents on future domestic violence incidents in Minneapolis, Minnesota.[35] Compared to less severe police responses, the Minneapolis experiment found that mandatory arrests lead to significantly lower rates of domestic violence. Sherman and Berk urged caution, but police departments across the nation adopted the mandatory arrest policy based on the results of one evaluation conducted in one city.

However, what worked in Minneapolis did not always work in other locations. Experiments conducted by Sherman and others in Omaha, Nebraska; Milwaukee, Wisconsin; Charlotte, North Carolina; Colorado Springs, Colorado; and Dade County, Florida, found mixed results.[36] Experiments in Omaha, Milwaukee, and Charlotte found that mandatory arrests lead to long-term *increases* in domestic violence. Apparently, knowing that they would automatically be arrested prompted repeat offenders to become more abusive. In a subsequent analysis of the disparate findings, Sherman postulated that arrested individuals who lacked a stake in conformity within their communities were significantly more likely to engage in domestic violence after arrest, while married and employed arrested individuals were significantly less likely to commit further domestic violence infractions.

The Building Strong Families (BSF) demonstration project sponsored by the U.S. Department of Health and Human Services provides a more recent example. BSF provided counseling services to unmarried couples who were expecting or had recently had a baby in eight sites (Atlanta, Georgia; Baltimore, Maryland; Baton Rouge, Louisiana; Orange and Broward counties, Florida; Houston, Texas; Allen, Marion, and Lake counties, Indiana; Oklahoma City, Oklahoma; and San Angelo, Texas). The marriage program's intent was to steer low-income unmarried couples with or expecting a child toward marriage.

The eight-site demonstration project is undergoing an experimental evaluation by Mathematica Policy Research, a leading research firm that specializes in conducting impact evaluations of government programs. More than 5,000 couples were randomly assigned to a relationship counseling group or a control group that could not participate in the program.

In 2010, Mathematica released the initial findings from a 15-month follow-up study.[37] Overall, the authors found that "BSF did not make couples more likely to stay together or get married. In

34. Erica B. Baum, "When the Witch Doctors Agree: The Family Support Act and Social Science Research," *Journal of Policy Analysis and Management*, Vol. 10, No. 4 (Autumn 1991), pp. 603–615, and Gueron, "The Politics of Random Assignment," pp. 15-49.
35. Lawrence W. Sherman and Richard A. Berk, "The Specific Deterrent Effects of Arrest for Domestic Assault," *American Sociological Review*, Vol. 49, No. 2 (April 1984), pp. 261–272.
36. Lawrence W. Sherman, *Domestic Violence: Experiments and Dilemmas* (New York: Free Press, 1992); Lawrence W. Sherman, Douglas A. Smith, Janell D. Schmidt, and Dennis P. Rogan, "Crime, Punishment, and Stake in Conformity: Legal and Informal Control of Domestic Violence," *American Sociological Review*, Vol. 57 (October 1992), pp. 680–690; Lawrence W. Sherman, Janell D. Schmidt, Dennis P. Rogan, Douglas A. Smith, Patrick R. Gartin, Ellen G. Cohn, Dean J. Collins, and Anthony R. Bacih, "The Variable Effects of Arrest on Criminal Careers: The Milwaukee Domestic Violence Experiment," *The Journal of Criminal Law & Criminology*, Vol. 83, No. 1 (1992), pp. 137–169.
37. Robert G. Wood, Sheena McConnell, Quinn Moore, Andrew Clarkwest, and JoAnn Hsueh, *Strengthening Unmarried Parents' Relationships: The Early Impacts of Building Strong Families*, Mathematica Policy Research, May 2010, at *http://www.mathematica-mpr.com/publications/pdfs/family_support/BSF_impact_finalrpt.pdf* (March 14, 2011). A long-term follow-up study will be conducted when the couples' children reach the age of three.

addition, it did not improve couples' relationship quality."[38] For example, 17 percent of all couples participating in the program eventually married, while 18 percent of the couples excluded from the program were married 15 months after random assignment—a statistically indistinguishable difference of 1 percentage point.[39]

While the evaluation of the eight demonstration sites found federally funded marriage promotion programs to be ineffective overall, the results from Baltimore and Oklahoma City were contradictory. (The results from the other six sites were largely consistent with the overall finding that BSF had no effect on improving the relationships of participating couples.) In Baltimore, compared to couples in the control group, unmarried couples participating in the program were less likely to be still romantically involved.[40] In addition, couples in the program reported less support and affection in their relationships, and fathers were less likely to provide financial support for their children and less likely to engage in cognitive and social play with their children.[41]

In Oklahoma City, the opposite occurred. While unmarried couples in the program were no more likely to marry than were the control group couples, Oklahoma participants reported improvements in relationship happiness, support and affection, use of constructive conflict behaviors, and avoidance of destructive conflict behaviors.[42] Additionally, fathers participating in the program were more likely to provide financial support for their children than were their counterparts in the control group.[43]

If Baltimore were the only site evaluated, then the results would indicate that federally sponsored marriage counseling for unmarried couples with children has harmful effects. The somewhat positive Oklahoma City results would have led to the opposite conclusion.

Contradictory results from evaluations of similar social programs implemented in different settings are a product not only of implementation fidelity, but also of the enormous complexity of the social context in which these programs are implemented. Jim Manzi, a senior fellow at the Manhattan Institute, uses the conflicting results of experimental evaluations to explain the influence of "causal density" on the social sciences.[44] Causal density, a term coined by Manzi, is "the number and complexity of potential causes of the outcomes of interest."[45] Manzi postulates that as causal density rises, social scientists will find greater difficulty in identifying all of the factors that cause the outcome of interest.

Just as with the contradictory effects of mandatory arrest policies by location, the confounding influence of causal density may have contributed to the conflicting BSF findings in Baltimore and Oklahoma City. For this reason, experimental evaluations of federal social programs should be conducted in multiple sites.

Step #3: Congress should specify the types of outcome measures to be used to assess effectiveness.

A federal program that is intended to ameliorate a particular social problem should be assessed on its impact on that particular social problem. For example, when assessing the impact of prisoner reentry programs, the most important outcome measure is recidivism. Some have questioned the emphasis on recidivism as a measure of effectiveness compared to other measures that assess adjustment or reintegration of former prisoners into society,[46] but while intermediate measures, such as finding employ-

38. *Ibid.*, p. xii.
39. *Ibid.*, p. 12.
40. *Ibid.*, p. 16, Table 7.
41. *Ibid.*, p. 16, Table 7, and p. 22, Table 11.
42. *Ibid.*, p. 16, Table 7.
43. *Ibid.*, p. 22, Table 11.
44. Jim Manzi, "What Social Science Does—and Doesn't—Know," *City Journal*, Vol. 20, No. 3 (Summer 2010), pp. 14–23, at *http://www.city-journal.org/2010/20_3_social-science.html* (March 14, 2011).
45. *Ibid.*

ment and housing, are important, these outcomes are not the ultimate goal of reentry programs. If former prisoners continue to commit crimes after going through reentry programs, then any intermediate outcomes are irrelevant to judging whether the programs are effective.

Impact evaluations that rely solely on intermediate outcomes tell little about the effectiveness of federal social programs in ameliorating the targeted social problems. While federal social programs should be assessed on intermediate outcomes, these measures should never substitute for primary outcomes.

Step #4: Congress should institute procedures that encourage government agencies to carry out congressionally mandated evaluations, despite any entrenched biases against experimental evaluations.

Simply mandating an experimental evaluation does not necessarily guarantee that the evaluation will actually be made. The Department of Labor, for example, has a poor track record in implementing and disseminating experimental evaluations mandated by Congress.

The Workforce Investment Act (WIA) of 1998 mandated a large-scale, multisite evaluation of the Department of Labor's job-training programs and required the department to report the results by September 2005. Despite this mandate and deadline, the Department of Labor procrastinated.[47] In November 2007, nine years after the passage of the Workforce Investment Act and more than two years after the deadline, the department finally submitted a request for proposals for the evaluation.[48] The contract for the experimental evaluation was awarded in June 2008, almost four years after the deadline.[49] According to the U.S. Government Accountability Office, the evaluation will not be completed until June 2015—nearly 10 years after its original due date and 17 years after Congress mandated the evaluation.[50]

Congress needs to take steps to ensure that evaluations are completed in a timely manner. One recommended method is to require department heads, such as the Attorney General or Secretary of Labor, to submit annual progress reports, with the first report to be submitted no later than one year after Congress mandates the evaluation. The progress reports would go to the appropriations and oversight committees of both chambers of Congress. For example, the Department of Labor would be required to submit the report to the Senate and House Committees on Appropriations; the Senate Committee on Health, Education, Labor and Pensions; and the House Committee on Education and the Workforce. Thirty days after the report is submitted to Congress, it should be posted on the department's Web site.

Step #5: Congress should require that congressionally mandated evaluations be submitted to the relevant congressional committees in a timely manner after completion.

46. Christy A. Visher and Jeremy Travis, "Transitions from Prison to Community: Understanding Individual Pathways," *Annual Review of Sociology*, Vol. 29 (2003), pp. 89–113.
47. David B. Muhlhausen and Paul Kersey, "In the Dark on Job Training: Federal Job-Training Programs Have a Record of Failure," Heritage Foundation *Backgrounder* No. 1774, July 6, 2004, at *http://www.heritage.org/Research/Reports/2004/07/In-the-Dark-on-Job-Training-Federal-Job-Training-Programs-Have-a-Record-of-Failure*.
48. U.S. Department of Labor, "Requests for Proposals (RFP) 2007," at *http://www.doleta.gov/grants/rfp07.cfm* (July 18, 2010), and U.S. Government Accountability Office, *Employment and Training Administration: More Actions Needed to Improve Transparency and Accountability of Its Research Programs*, GAO–11–285, March 2011, at *http://www.gao.gov/new.items/d11285.pdf* (April 21, 2011).
49. U.S. Government Accountability Office, *Employment and Training Administration: More Actions Needed*.
50. U.S. Government Accountability Office, "Workforce Investment Act: Labor Has Made Progress in Addressing Areas of Concern, but More Focus Needed on Understanding What Works and What Doesn't," Statement of George A. Scott, Director, Education, Workforce, and Income Security, before the Subcommittee on Higher Education, Lifelong Learning, and Competitiveness, Committee on Education and Labor, U.S. House of Representatives, GAO–09–396T, February 26, 2009, at *http://www.gao.gov/new.items/d09396t.pdf* (July 18, 2010), and *Employment and Training Administration: More Actions Needed*.

Thirty days after any evaluation is submitted to Congress, the evaluation should be made available on the Web site of the federal government agency responsible for the evaluation. Requiring that Congress and the public be informed of evaluation results is important because government agencies are quick to release positive results but sometimes reluctant to release negative results.

For example, a cost-benefit analysis of the Job Corps that found that the program costs outweighed the benefits was finalized in 2003,[51] but the Department of Labor withheld it from the public until 2006.[52] The Government Accountability Office has criticized the Department of Labor for its history of delaying the release of its research findings.[53] Similarly, the Department of Health and Human Services has noticeably delayed the release of an evaluation of Head Start that reported underwhelming results.[54] Congress needs to be vigilant in ensuring that evaluation results are disseminated promptly.

Conclusion

With the federal debt reaching staggering heights, Congress needs to ensure that it is spending taxpayer dollars wisely. Multisite experimental evaluations are the best method for assessing the effectiveness of federal social programs. Yet to date, this method has been used on only a handful of federal social programs. Congress needs to reverse this trend.

—David B. Muhlhausen, Ph.D., is Research Fellow in Empirical Policy Analysis in the Center for Data Analysis at The Heritage Foundation.

51. Peter Z. Schochet, Sheena McConnell, and John Burghardt, *National Job Corps Study: Findings Using Administrative Earnings Records Data: Final Report* (Princeton, N.J.: Mathematica Policy Research, October 2003).
52. David B. Muhlhausen, "Job Corps: A Consistent Record of Failure," Heritage Foundation *WebMemo* No. 1374, February 28, 2007, at *http://www.heritage.org/Research/Reports/2007/02/Job-Corps-A-Consistent-Record-of-Failure*.
53. U.S. Government Accountability Office, *Employment and Training Administration: More Actions Needed*.
54. Jennifer Marshall, David B. Muhlhausen, Russ Whitehurst, Nicholas Zill, and Debra Viadero, "Is Head Start Helping Children Succeed and Does Anyone Care?" video feed, The Heritage Foundation, March 22, 2010, at *http://www.heritage.org/Events/2010/03/Head-Start* (July 19, 2010).

APPENDIX
MODEL LEGISLATION FOR MULTISITE EXPERIMENTAL EVALUATIONS

SEC. <Insert number>. EVALUATIONS.

(a) PROGRAMS AND ACTIVITIES CARRIED OUT UNDER THIS TITLE.—For the purpose of improving the management and effectiveness of programs and activities carried out under this title, the Secretary shall provide for the continuing impact evaluation of the programs and activities, including those programs and activities carried out under section <Insert number>. Such impact evaluations shall address—

(1) Outcomes measures of the effectiveness of such programs and activities in relation to their cost, including the extent to which the programs and activities—

(A) Improve the <Insert outcome measures> of participants in comparison to comparably situated individuals who did not participate in such programs and activities;

(B) Increase the <Insert outcome measures> over the level that would have existed in the absence of such programs and activities; and

(C) Increase the <Insert outcome measures> of participants in comparison to comparably situated individuals who did not participate in such programs and activities;

(2) The effectiveness of the performance measures relating to such programs and activities;

(3) The effectiveness of the structure and mechanisms for delivery of services through such programs and activities;

(4) The impact of such programs and activities on the community and participants involved;

(5) The impact of such programs and activities on related programs and activities;

(6) The extent to which such programs and activities meet the needs of various demographic groups; and

(7) Such other factors as may be appropriate.

(b) OTHER PROGRAMS AND ACTIVITIES.—The Secretary may conduct impact evaluations of other federally funded programs related to <Insert policy area (e.g., employment, early childhood education)> and activities under other provisions of law.

(c) TECHNIQUES.—Impact evaluations conducted under this section shall use appropriate methodology and research designs, including the use of intervention and control groups chosen by scientific random assignment methodologies. For each impact evaluation, the Secretary shall fulfill all the notification and reporting requirements under subsections (d), (e), and (f). The Secretary shall conduct as least 1 multisite control group evaluation under this section by the end of fiscal year <Insert year>.

(d) NOTIFICATION OF IMPACT EVALUATION PROGRESS.—

(1) REPORTS TO CONGRESS.—Not later than 1 year after the date of the enactment of the <Insert name of Act>, and annually thereafter, the Secretary shall transmit to the <Insert two or more House committees> of the House of Representatives and the <Insert two or more Senate committees> of the Senate a report on the progress the Secretary is making in evaluating the programs and activities carried out under this section.

(2) AVAILABILITY TO GENERAL PUBLIC.—Not later than 1 year after the date of the enactment of the <Insert name of Act>, and annually thereafter not later than 30 days after the transmission of an annual report under paragraph (1), the Secretary shall make available the reports to the general public on the Internet website of the Department of <Insert name>.

(e) REPORTS.—The entity carrying out an impact evaluation described in subsection (a) or (b) shall prepare and submit to the Secretary a draft report and a final report containing the results of the evaluation.

(f) REPORTS TO CONGRESS.—Not later than 30 days after the completion of such a report described in subsection (e), the Secretary shall transmit the draft report to the <Insert House committees from subsection (d)> of the House of Representatives and the <Insert House committees from subsection (d)> of the Senate. Not later than 30 days after the completion of such a final report, the Secretary shall transmit the final report to such committees of the Congress. All reports must be made available to the general public on the Department's internet web site within 30 days of being transmitted to such committees of Congress.

(g) DEFINITIONS.—In this section:

(1) IMPACT EVALUATION—The term "impact evaluation" means an evaluative study that evaluates, in accordance with subsection (a), the outcomes of programs and activities carried out under this title, including the impact on social conditions such programs and activities are intended to improve.

(2) SCIENTIFIC RANDOM ASSIGNMENT METHODOLOGIES—The term "scientific random assignment methodologies" means research designs conducted in program settings in which intervention and control groups are—

(A) formed by random assignment; and

(B) compared on the basis of outcome measures for the purpose of determining the impact of programs and activities carried out under this title.

(3) CONTROL GROUP.—The term "control group" means a group of individuals—

(A) who did not participate in the programs and activities carried out under this title; and

(B) whose outcome measures are compared to the outcome measures of individuals in an intervention group.

(4) INTERVENTION GROUP—The term "intervention group" means a group of individuals—

(A) who participated in the programs and activities carried out under this title; and

(B) whose outcome measures are compared to the outcome measures of individuals in a control group.