PRESIDENTIAL ADDRESS: From the Great Society to Continuous Improvement Government: Shifting from "Does It Work?" to "What Would Make It Better?"

Douglas J. Besharov

Abstract

In the 1960s, various social programs were started (like Head Start) or dramatically expanded (like AFDC). Loosely, this period of expansion is called the Great Society. Too many Great Society social programs, unfortunately, have been disappointments—at least when compared to the high hopes of the '60s. Even if they "work," most of us wish that they worked much better. Some people take such statements to mean that the Great Society's social programs should be defunded. Most Great Society programs, however, are surely here to stay, for they serve important social functions. How many of us really think there could be an America without a social safety net? It is now time to do the difficult and unglamorous work of systematic program improvement. Instead of testing program efficacy over and over again, we should engage in ongoing and evidence-based efforts to increase program effectiveness (in both large and small ways). © 2009 by the Association for Public Policy Analysis and Management.

I am delighted to be here today and to have had the opportunity to serve APPAM, an organization I respect and care deeply about. Over the years, I've learned a lot from APPAM—from its meetings, from *JPAM*, and from participating in its governance. So, thank you all very much.

Earlier this week, Barack Hussein Obama[1] was elected president of the United States. But even those who did not vote for Barack Obama should wish him—and our nation—a successful presidency.

In important respects, my talk today is based on my own hopes for Obama's presidency. With his mandate—and the large Democratic majorities in Congress—President Obama should have the political running room to engage in a candid appraisal of current domestic programs and take steps to improve them—as he promised in the campaign. (You could liken it to Nixon going to China.) And, judging from his campaign team, he should have the talent pool of appointees needed to do so.

POLICY ANALYSIS AND MANAGEMENT INTERTWINED

I am, however, worried that an Obama administration may approach government improvement as solely a management issue. In the campaign, Obama pledged to create a White House "SWAT team" made up of government professionals to review programs for waste and inefficiency. After such reviews, he said:

[1]The name he will use when he is sworn in as president (Parsons, McCormick, & Nicholas, 2008).

Journal of Policy Analysis and Management, Vol. 28, No. 2, 199–220 (2009)
© 2009 by the Association for Public Policy Analysis and Management
Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20423

> We will fire government managers who aren't getting results, we will cut funding for programs that are wasting your money, and we will use technology and lessons from the private sector to improve efficiency across every level of government—because we cannot meet 21st century challenges with a 20th century bureaucracy.[2]

Most programs, however, cannot simply be managed to better performance. To be made more effective, they usually also need a change in their structure or orientation. Thus, sound policy analysis and program evaluation matter as much as proper management. The title of this talk refers to "continuous improvement government" not because I like buzzwords, but to emphasize that program improvement is a step-by-step process—one that combines policy analytic and management tools.[3]

Policy analysis and management are, of course, APPAM's two dimensions—connected in many ways and at many levels. In preparing for this talk, I reread prior presidential addresses and I noticed how most of them, in different ways, grappled with both of these elements of public policy. So, today, I will do the same, but with my focus being on the policy side of program improvement. And, in so doing, I will concentrate on the need to speed up the process of knowledge building, that is, the process of hypothesis identification and testing through evaluations and implementation studies.

HERE TO STAY

In a nutshell, here's my argument: In the 1960s, various social programs were started (like Head Start) or dramatically expanded (like AFDC). Loosely, we call this period of expansion the Great Society. Most of these programs sought to ameliorate pressing social needs and most are still with us today (although many have changed in important ways).

Too many Great Society social programs, unfortunately, have been disappointments—at least when compared to the high hopes of the sixties. Even if they "work" (leaving to another time what this means), most of us wish that they worked a lot better.

Here, I would make a broad and important distinction between (1) the greater relative success of income support programs, which usually achieve their goal of giving people money, although perhaps at the cost of at least some moral hazard; and (2) the relative lack of success of social service/social welfare programs that seek to increase human or social capital, through either more education or skills or positive changes in behavior.

Some people take such statements as meaning that the Great Society's social service/social programs should be defunded. Some programs, such as Model Cities, were terminated because they seemed so disastrous[4]—and perhaps more should have been. Most, however, are surely here to stay. How many of us really think there could be an America without a social safety net? Sure, some programs might be trimmed or modified, like AFDC/TANF, but cash welfare is still with us, as are food stamps, Medicaid, WIC, Head Start, child welfare services, job training, and so forth.

Even if these programs did not serve important social functions, which they do, it would be nearly impossible to dismantle them. The federal government's family literacy program, Even Start, provides adult literacy and early childhood education services to families with children under age 8 at a cost of about $11,000 per family per year. In the past 15 years, three evaluations, two of which were rigorous random assignment designs, have found that Even Start had no significant impact on children.[5]

[2] Obama, 2008.
[3] See, for example, Loeb & Plank, 2008.
[4] DeMuth, 1976.
[5] St. Pierre et al., 1995.

After the publication of Even Start's third negative evaluation in 2003, the Congress mustered the resolve to reduce funding from $279 million to $63 million in 2008 (in 2007 dollars).[6] Yet the Congress seems unwilling to terminate the program completely. No less a duo than Senators Olympia Snowe (R-ME) and Hillary Clinton (D-NY) complained in 2007 that the budget cuts have "resulted in dozens of local programs being shuttered leaving thousands of children and adults without the local support to become productive, literate members of our communities."[7]

The seeming immortality of government programs is, of course, not limited to social programs. Remember that 1950s' relic of the Cold War, the mohair subsidy, designed to ensure a source of wool for military uniforms decades after they were made of everything but wool. It was ended in 1995, only to be reinstated four years later through the good work of industry lobbyists.

So perhaps Model Cities will be back. Actually, that would not surprise me.

Some Great Society programs are not well designed for contemporary problems; after all, much has changed since the 1960s. In fact, many were probably not the right approach even back then. But, forty-plus years later, we should declare that most elements of the Great Society are as permanent as any other government programs, and that it is time for us to step up to the plate and do the unglamorous work of program improvement.

Using more formal evaluation terminology, I am suggesting that, instead of testing program efficacy over and over again, we engage in systematic and evidence-based efforts to increase program effectiveness (in both large and small ways).[8]

HEAD START

Head Start, considered by many the gem of the Great Society, is a prime example of the need to improve an ongoing program. Since that first summer of 1965, about 25 million children have passed through the program, at a total cost of about $145 billion, and yet we are still arguing about whether Head Start "works." I've contributed some to this argument, so I know it well.[9]

Of course, it matters how children are raised. Romulus and Remus were suckled by a wolf, and they founded a city that became a great empire. The rest of us, though, need much more care and nurturing to reach our full potential. Thus, the real policy question is not whether there was a proper random assignment of those 123 Perry Preschool children back in 1962, but, rather: *Is Head Start doing everything a program like it can do to compensate for family and community deficits?*

Tragically, repeated studies have shown that the *current* Head Start program—*not the idea behind the program*—fails to achieve the vitally important goals assigned to it. In other words, regardless of the efficacy of the idea, the program, as implemented under real-world conditions, does not seem effective.

Spurred by a 1997 U.S. General Accounting Office (now Government Accountability Office) report concluding that there was "insufficient" research to determine Head Start's impact,[10] in 1998, Congress required the U.S. Department of Health and Human Services to conduct the first rigorous national evaluation of Head Start. To its credit, the Clinton administration took this mandate seriously and initiated a 383-site randomized experiment involving about 4,600 children. (In fact, throughout his presidency, Bill Clinton and his appointees were strongly supportive of

[6] McCallion, 2006; U.S. Department of Health and Human Services, 2007; U.S. Department of Health and Human Services, 2008c.
[7] Snowe & Clinton, 2007.
[8] "Efficacy" refers to whether an intervention works in achieving a particular outcome or outcomes under ideal circumstances, whereas "effectiveness" looks at program effects under real-world circumstances. See Flay et al., 2005.
[9] Nathan, 2007.
[10] U.S. General Accounting Office, 1997.

efforts to improve Head Start, even to the point of defunding especially mismanaged local programs.)

Confirming the findings of earlier, smaller evaluations, the Head Start Impact Study (released in June 2005) found that the current Head Start program has little meaningful impact on low-income children.[11] For 4-year-olds (half the program), statistically significant gains were detected in only 6 of 30 measures of social and cognitive development and family functioning. Results were somewhat better for 3-year-olds, with statistically significant differences on 14 of 30 measures; however, the measures that showed most improvement tended to be superficial. For both age groups, the actual gains were in limited and overlapping areas and disappointingly small, making them unlikely to lead to later increases in school achievement. For example, even after spending about six months in Head Start, 4-year-olds could identify only two more letters than those who were not in the program, and 3-year olds could identify one and one-half more letters. No gains were detected in much more important measures such as early math learning, oral comprehension (more indicative of later reading comprehension), motivation to learn, or social competencies, including the ability to interact with peers and teachers.

In the few domains where the Head Start Impact Study found an impact, the effect sizes tended to be between 0.1 SD and 0.4 SD; the effects reported were based on parental reports as opposed to objective tests of the children. Some argue that even such small effect sizes can make a difference. Perhaps, although it is difficult to see how they would result in *socially significant* gains as the children mature. (Equivalent effect size changes for IQ would be from 1.5 to 6 points.) Surely reasonable people on both sides of the political spectrum can agree that these outcomes are simply not good enough, and that disadvantaged children deserve much better.

Head Start advocates counter these field results by citing econometric studies (based on surveys conducted in the late 1960s, 1970s, and 1980s) conducted by economists Janet Currie, Duncan Thomas, and Eliana Garces[12] and Jens Ludwig and Douglas Miller.[13] Whatever the validity of these statistical studies, they attempt to estimate Head Start's impact before it was doubled in size and before it had gone through another almost two decades of poor management—enabled by the political cover provided by key members of Congress.

Furthermore, more recent econometric studies using the Early Childhood Longitudinal Study (ECLS-K) come to as bleak a conclusion as the Head Start Impact Study. Although their results are subject to possible selection bias problems, according to Katherine Magnuson, Christopher Ruhm, and Jane Waldfogel: "Children who attended prekindergarten or preschool have the highest test scores, followed by those exclusively in parental care or receiving other types of nonparental care (for example, babysitters); Head Start enrollees have the lowest scores in math and reading."[14]

FUNDING PRIORITIES

Many observers argue that the problem is insufficient funds—in Head Start and other social programs. Money is certainly an issue. Just a cursory look at Figure 1 tells the story: Spending on Social Security and medical entitlements (Medicare, Medicaid, and all other forms of medical aid) is way up, as is spending on cash programs (many of which are also entitlements). In contrast, spending on most service- or treatment-oriented programs has been comparatively flat for three decades.

[11] Puma et al., 2005.
[12] Currie & Thomas, 1995, 1996; Garces, Thomas, & Currie, 2002.
[13] Ludwig & Miller, 2007.
[14] Magnuson, Ruhm, & Waldfogel, 2004, p. 17. See also Rumberger & Tran, 2006; Loeb et al., 2005.

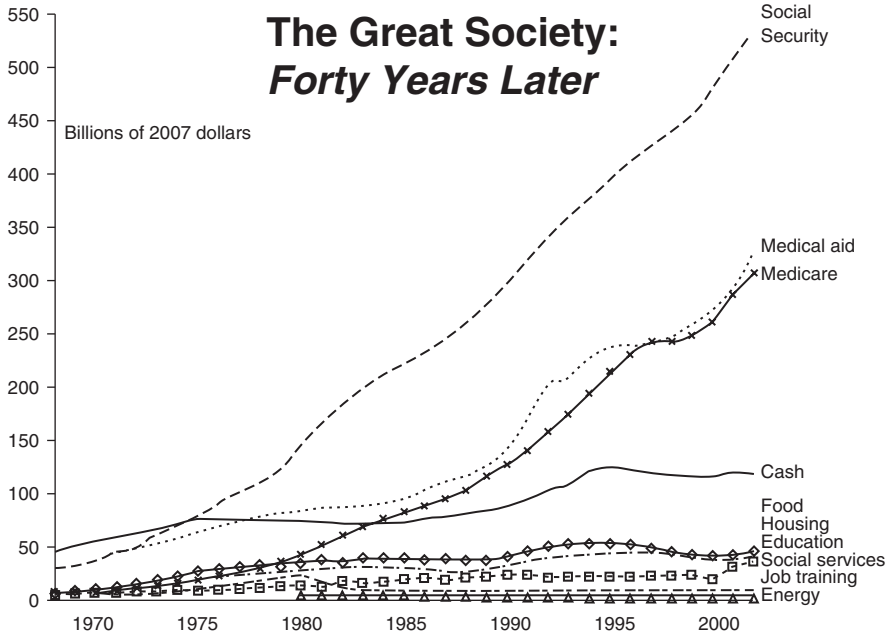

Figure 1. Total Spending for Income-Tested Benefits, Medicare, and Social Security, 1968–2002 (Millions of 2007 Dollars).

For years, the entitlement crisis (and it should be called that)—coupled with repeated tax cuts—has been eroding discretionary spending at all levels of government. Across a whole range of activities, the political pressure to feed Social Security, Medicare, and Medicaid (many would want to add defense spending to this list, even though it has been a declining percentage of GDP) has undercut government's ability to think, to plan, and to do. A government that underinvests in maintaining bridges and in the SEC's oversight of financial institutions is probably doing a pretty bad job of maintaining social welfare services for the disadvantaged, let alone improving and expanding them.

So more money would undoubtedly help, assuming it were spent wisely. A giant assumption, though. Head Start already costs about 50 percent more than high-quality, state-run pre-K programs—with much poorer results.[15]

KNOWLEDGE BUILDING IS TOO SLOW

There are many reasons for weak or ineffective programs, but let me focus on only one: the slow pace of knowledge accretion, needed for informed program development and improvement. Current R&D practices are too slow, too haphazard, and too often fail to factor in the dead ends that are inevitable in policy research and program development.

In many fields, of course, real advances are being made. Across the broad social welfare landscape, however, progress in social understanding and program development is excruciatingly slow. For the past year and a half, I have had the privilege of serving as the editor of JPAM's Policy Retrospectives section, and I have seen this

[15] Besharov, Myers, & Morrow, 2007.

up close. I cringe when responsible researchers say that it may take a generation to get things right. Most of us will be retired by then, and some of us will be dead.

We owe it to the recipients of these services and their families to do better, and do it faster.

One of the truly great accomplishments of modern social science has been the wide use of the randomized experiment to address important policy questions. For good reason, they are widely referred to as the "gold standard." In recent decades, such APPAM stalwarts as Abt Associates, MDRC, Mathematica Policy Research (MPR), and the Urban Institute—often funded by the U.S. Departments of Labor, Health and Human Services, and Housing and Urban Development, and more lately joined by the Department of Education—have shown time and again that rigorous social experimentation is both possible and fruitful.

There were many other players, of course. For example, the RAND Health Insurance Experiment (HIE) established that increased cost sharing by clients led to reduced medical care usage without any widespread effect on health status.[16] Although there remains some disagreement about the methodology and implications of the findings (but isn't there always?),[17] most of us are members of health insurance plans shaped by the RAND findings.

In program area after program area, well-planned and well-implemented randomized experiments, even if not definitive and sometimes controversial, have made major contributions to public policy. An abbreviated list would include evaluations of class size; early childhood education; food stamps cashouts; housing allowances or vouchers; Job Corps; teen mother programs; the 21st Century after-school program; vouchers for housing, job training, and K–12 education; welfare reform programs (such as welfare-to-work); and, of course, the Negative Income Tax.

These very real successes, however, should not blind us to the glacial slowness of current R&D practices. It took, for example, more than seven years (ten years if you include when the 30-month impacts were released) to complete Abt's very fine evaluation of the Job Training Partnership Act (JTPA). The JTPA study found modestly positive results for adult men and women,[18] but negative earnings effects for disadvantaged male youth and no earnings effects for disadvantaged female youth. These findings led Congress to cut JTPA's budget for youth programs by 80 percent. By the time the results were released, however, the JTPA's youth programs had been revamped, with, among other things, the creation of a separate youth program and targeted services to those with multiple employment barriers. But none of the changes were assessed by Abt before the youth program was all but eliminated.

Ten years later, we are only now beginning an evaluation of its replacement, the Workforce Investment Act (WIA). Final results are not scheduled for release until 2015—six years from now. That's halfway through Barack Obama's second term, assuming that there is one. Can you see him waiting until then before deciding whether to put more money in the program or to radically restructure it?

The WIA evaluation is not an exceptional case:

- It has been eight years since the Head Start Impact Study was initiated,[19] but the 54-month results are not expected for another year.[20]

[16] Manning et al., 1987.
[17] See, for example, Nyman, 2007.
[18] Bloom et al., 1997, p. 560. Average earnings impacts per enrollee over the 30-month follow-up period were $1,837 for adult women, $1,599 for adult men (both statistically significant), but they were not statistically significant for female or male youth, with the exception of male youth arrestees, who experienced a statistically significant loss of $6,804, according to survey data on earnings.
[19] All the dates are based on when the contracts were awarded.
[20] U.S. Department of Health and Human Services, 2008d.

- It has been 14 years since the Moving to Opportunity study was initiated, and it will be another year before the final, ten-year follow-up results are available.[21]
- It has been ten years since the Employment Retention and Advancement evaluation was initiated to help welfare recipients and low-wage workers succeed in the labor market, and it will be another year before the final two-year impact results from all sites are available.[22]
- It has been six years since the Building Strong Families project assessing the effectiveness of healthy marriage education services for low-income unwed parents at or near the birth of their child was initiated, and it will be another year before interim findings are available, and three more years before the final three-year impact results are published.[23]
- It has been 15 years since the National Job Corps Study was initiated, and although four-year findings were available after seven years, the nine-year findings only recently became available this year.[24]

Many evaluations take a much shorter time, of course, but such long periods before results are available are all too common, and they have slowed program improvement to a slow crawl.

We urgently need to speed up and expand the processes of program design and testing, program implementation and evaluation, and, when necessary, program redesign—as the process necessarily repeats itself over time.

FIND MORE PROMISING IDEAS TO TEST

Putting aside questions of generalizability, some rigorous evaluations have shown that particular social programs seem to "work," at least modestly in the short term. These include supported work; welfare-to-work programs; training for adults; Job Corps for youth; and nurse home-visiting programs for pregnant, young women. Far too many evaluations, however, obey Pete Rossi's "Iron Law of Evaluation," namely, that: "The expected value of any net impact assessment of any large scale social program is zero."[25]

Why? As Pete would patiently explain, the experiment may have been underpowered, or poorly designed, or poorly implemented. But, as a college-aged Trotskyite turned social liberal (no Neocon, he), Pete hated to acknowledge a more fundamental problem: Sometimes the program being tested was simply a bad idea.

Too often, the political and administrative process leads to a research design that seems simply, well, wrongheaded. Consider, for example, the Comprehensive Child Development Program (CCDP), meant to test the effect of well-coordinated services and parental education on the growth and development of young children.[26]

Hopes were high for this $300 million program that served low-income, single mothers for as long as five years. It spent about $19,000 per family per year (that's on top of AFDC, food stamps, WIC, and other safety-net programs), and about $58,000 total per family. But a closer look at the project design suggests that it never had a good chance of succeeding:

[21] U.S. Department of Housing and Urban Development, 1996.
[22] U.S. Department of Health and Human Services, 2008b.
[23] U.S. Department of Health and Human Services, 2008a.
[24] Schochet, Burghardt, & McConnell, 2008.
[25] Rossi, 1987.
[26] St. Pierre et al., 1997.

- Program sites were all but prohibited from funding their own services—because this was a test of the impact of using free services from existing community sources (programs could create their own services "when necessary," but few did);
- Center-based child care was not authorized unless the mother was working—because the program sought to teach parents how to educate their children; and
- The sites were also prohibited from changing their approach even as their experience suggested that a mid-course correction was urgently needed—because there was to be no deviation from the planned intervention.[27]

I could go on.

When Abt announced that the program had had no positive impact on the young mothers or their children, many people concluded that these mothers were beyond the reach of our programs, rather than that the CCDP was a bum intervention.[28] (Of course, some advocates just blamed the evaluators, but that's nothing new, either.)[29]

Given how little was learned from this effort, it is difficult to see the $300 million spent on the CCDP as anything but a waste of money. I don't mean to single out the CCDP for special criticism. I could give many other examples of poorly conceived program designs. The CCDP just happens to be an experiment that I lived through—and have the scars to show for it.

Competing priorities and concerns frequently intrude on the process of program design. The original Nurse Home Visitation Program demonstrations used specially trained registered nurses who worked at a local hospital or Department of Health.[30] The principal investigator, David Olds, considered them a key element to the program's success. But, wanting to lower the cost of the intervention and to involve the community, many replications used paraprofessionals (often from the neighborhood) instead of nurses. Whether or not for that reason alone, they all failed—and the result was to discredit at least partially an otherwise credible program intervention.[31]

At the risk of further offending, let me also say that, as a field, we are not very good at coming up with good ideas to test. It's one thing to advocate for a new "program" to combat a particular serious social problem. It's quite another to specify what particular elements should be in the program. Truth be told, in many areas, we suffer a dearth of good ideas that can be appropriately tested. But, really, what are the new approaches to reducing teen pregnancy that should be tested? To job retention for welfare leavers? To helping troubled youth? To making schools "work"?

The chances are small that some government or foundation committee will come up with a program that will work much better than all the other programs that have gone before. We could decide, for example, that a major cause of teen pregnancy is unprotected sex, and we could then decide that there should be a program to encourage safe sex. But we would have great difficulty coming up with reasonable approaches to doing so—approaches that have not already been tried and been found wanting.

Do not misunderstand. There are many untested ideas deserving of serious examination. As Grover Whitehurst, former director of the Department of Education's Institute of Education Sciences, reminds us in his Rossi Award lecture: "The program that has a substantial positive impact may be rare but it exists. The probability of finding it will be remote unless we search widely, frequently, and intelligently."[32]

[27] St. Pierre et al., 1999.
[28] Samuelson, 1998.
[29] Gilliam et al., 2000.
[30] Child Trends, 2008.
[31] Sweet & Applebaum, 2004.
[32] Whitehurst, 2007.

INNOVATORS AND OUTLIERS

Recognizing this dearth of good, testable ideas, recent R&D planning efforts have actively sought to identify new ideas—usually by reaching out to a broad array of experts and frontline program operators, often using a snowball technique. This is a good, but insufficient, process. It depends on finding respondents who are able to identify promising new program ideas based on more than mere appearances and reputation. We need to expand and sharpen the process.

In most operating programs, and most demonstrations as well, there is a fair degree of variation from the mean. Yes, some Head Start centers are really good; so are some job training centers, and so forth. A number of researchers have therefore concluded that one of the best ways to find promising new program ideas is to look for "high performers"—or outliers—and then try to learn what they are doing that seems to work. Mark Kleiman of UCLA explains how the existence of such outliers can be exploited:

> The way we evaluate now, we measure the average performance of all the line operators, and take that average as a measure of how good the program is. But in a continuous-improvement world, we ought to be able to move average performance in the direction of the performance of the best operators (not all the way to, given regression toward the mean), either by figuring out what they're doing and teaching the others or by sheer filtering. So evaluators ought to be sensitive to the variance as well as the mean of the operator-level results.[33]

Call it the bottom-up generation of ideas: a process that identifies promising ideas that higher-level planners might never have imagined.[34]

That's essentially what happened in the welfare reform experiments of the 1980s and 1990s. The federal Work Incentive Program (WIN) authorized under OBRA 1981 gave states the freedom and the incentive to begin trying difficult approaches for moving recipients from welfare to work. Some were voluntary; some mandatory; and many were evaluated by MDRC. It soon became apparent that both approaches could either lower caseloads or increase earnings, at least modestly.[35] Subsequently, in California's state-funded Greater Avenues for Independence Program (GAIN) program, MDRC found that the "labor force attachment strategies" followed in Riverside County, California (mostly job search, some short-term training, and strict enforcement practices) stood out as outliers (in earnings increases and caseload declines) compared to other programs in the GAIN experiment.[36] Then, in a series of randomized experiments under the new JOBS Program, MDRC found that Riverside-like "labor force attachment strategies" outperformed human capital strategies.[37] Other random assignment studies confirmed this finding, as did simple pre/post analyses.[38]

Before this sequence of state-level experiments and evaluations, few experts proposed *mandatory* job search, work first, and other welfare-to-work practices coupled with strict enforcement.[39] Now, of course, such strategies characterize most welfare programs.

FLEXIBILITY TO INNOVATE

Essential to the bottom-up generation of ideas is the ability of individual programs to innovate, or at least to do things a little differently from their counterparts.

[33] M.A.R. Kleiman, personal communication, September 28, 2008.
[34] Mead, 2005.
[35] Gueron & Pauly, 1991. See also Gueron, 1988.
[36] Riccio et al., 1994.
[37] Hamilton et al., 2001.
[38] Grogger, Karoly, & Klerman, 2002.
[39] See, for example, Mead, 1990; Bane, 1989.

The progressive leaning in welfare policy was the direct result of the programmatic flexibility available through the welfare reform waiver process[40]—combined with rigorous program evaluation.

Too many programs, however, are straightjacketed by rules and regulations—most having been put in place to control the program and some to protect the program from the feared predations of conservative administrations. After welfare reform proved how much could be learned by the waiver/evaluation process, waiver rules under the Food Stamp Program (now the Supplemental Nutrition Assistance Program) were broadened somewhat, but they are still too restrictive and make serious experimentation all but impossible.[41]

The original Egg McMuffin was a violation of McDonald's rigid rules about only serving lunches and dinners from preapproved menus and precise recipes. It was developed surreptitiously by one franchisee—who then tricked Ray Kroc, legendary president of McDonald's, into tasting it. What if Kroc had refused to taste it?[42]

Within reason (and with all appropriate safeguards), program flexibility should be encouraged. Program administrators need the freedom to change what they are doing in response to operational experiences and changing conditions—and policy planners need the tools to measure any resulting differences in performance or outcomes.

In the coming years, I hope that the Obama administration can revisit the issue of waivers, which proved such a powerful tool in welfare reform. Even for the Obama team, though, that will not be easy. Through much of the period since the Johnson presidency, attempts to gain the cooperation of programs with efforts to evaluate and "improve" them have been stymied by the unfriendly political atmosphere. Putting aside the danger that an outside evaluation might get it wrong (a real problem, we must acknowledge), program operators rightly feared that any negative findings could be used to reduce funding (as happened to the JTPA youth program after the evaluation). Hence, for too many years and for too many programs, there has been an entirely understandable defensive tendency to circle the wagons.

In 2003, for example, the Bush administration proposed an eight-state waiver experiment that would have explored different approaches to integrating Head Start into the wider world of child care by giving states control over Head Start funds. Even with the most stringent controls, the Republican Congress refused to approve even this limited experiment, because of vociferous opposition from the Head Start lobby and its allies. In one particularly colorful phrase, Congressman George Miller (D-CA) said that handing control of Head Start over to states was "like handing your children over to Michael Jackson."[43]

In effect, they wouldn't taste the Egg McMuffin.

Why was there so much opposition to an experiment in even a few states—unless the fear was that the experiment would be successful, and would demonstrate a better model for providing early education to disadvantaged children? Or that the Bush administration would have distorted or misused the results? Perhaps, just perhaps, the Head Start community and the Democratic Congress will trust a President Obama more than they have trusted President Bush. But we should remember how an earlier Democratic Congress ignored Jimmy Carter's proposal to include Head Start in the new Department of Education.

[40] See Rogers-Dillon, 2004; Harvey, Camasso, & Jagannathan, 2000; U.S. Department of Health and Human Services, 1997.

[41] For example, a state SNAP waiver project that might reduce benefits by more than 20 percent to more than 5 percent of the proposed project recipients may not include more than 15 percent of state SNAP recipients and may not last longer than five years. Food and Nutrition Act. 7 U.S.C. § 2026 (2008). See generally Besharov & Germanis, 1999.

[42] Kroc & Anderson, 1987.

[43] Miller, 2003.

OUTCOME-ORIENTED PERFORMANCE MANAGEMENT

To push this point a step further: The search for good ideas should be ongoing and systematic—wholesale rather than retail. In this regard, I have high hopes for the development of outcome-oriented performance management systems that are capable of monitoring not just program outputs (like clients trained) but also short-term outcomes or longer-term impacts (like clients employed after a certain period of time).[44] To the extent that such performance management systems could accurately identify outliers, they could become engines of "continuous program improvement."[45]

Up to now, the tendency has been to use performance management systems to identify the outliers on the left hand of the distribution—and then work to either improve or defund them. That kind of high-stakes management has not made these systems popular with most service providers. Who likes to be judged, especially if the yardstick seems unfair? No better and more visible example exists than the No Child Left Behind Act.

Performance management systems, however, can also be—and, increasingly, are—used to identify outliers on the right hand of the distribution. These outliers should then be studied to see what it is about them that seems to work better than average. Although accountability is always controversial, such an approach should be less threatening to program operators.

In the future, we can expect significant payoffs as analysts develop more sophisticated techniques to plumb evolving performance management systems, such as that of the Workforce Investment Act. The Department of Labor funded Carolyn Heinrich of the University of Wisconsin–Madison, Peter Mueser of the University of Missouri–Columbia, Ken Troske of the University of Kentucky, and colleagues to use propensity scoring to explore the four-year employment and earnings impacts of early cohorts of WIA participants. Their report should be available in 2009, as opposed to 2015 for the randomized experiment about WIA's effectiveness, and at much less cost, too.[46]

ATTRIBUTING CAUSATION

Identifying what seems to be a promising approach is only the first step. Appearances can be deceiving. Next come the challenging tasks of (1) determining whether outcomes really are superior, and then (2) attributing causality to any apparent programmatic differences.

Identifying such outliers and attributing their better outcomes to the program rather than some other factor is no easy task. Burt Barnow of Johns Hopkins University and Peter Schochet and John Burghardt of MPR both established the discordance between the rankings generated by existing performance management systems (JTPA and Job Corps, respectively), and the results of randomized experiments at the same sites.[47]

I raise this in a separate discussion for emphasis. It is simply too easy to let appearances and ideological preferences drive the process. Think of the controversies over "schools that work." People see what seems to be a successful school, but there is no way to tell how much (or even whether) the school is actually contributing to the success of students.

[44] Heinrich, 2007.
[45] Such performance management systems, by the way, need not be national. Statewide and even local systems have been successfully used to identify outliers.
[46] C. J. Heinrich, personal communication, November 20, 2008.
[47] Barnow, 2000; Schochet & Burghardt, 2008.

I expect that statistical techniques will *sometimes* establish that the program is providing a "value added" (to a sufficient level of approximation, at least),[48] but, more often than impatient advocates and policymakers would like, a definitive determination will require rigorous and, I am afraid, time-consuming experimentation—as promising ideas are actually put to the test.

The challenge involved in obtaining useful results is illustrated by a 2003 Abt study that sought to identify the instructional and classroom management practices associated with high-performing classrooms in high-poverty schools.[49] The idea was to compare them to lower-performing classrooms in all the schools, as a way to identify for value-added. The Abt researchers first identified "high-performance" schools, using the 1999 Longitudinal Evaluation of School Change and Performance (LESCP) survey with a sample of 60 schools with high levels of poverty. They identified a subset of 18 schools that, in the 1997–1999 period, were operating close to the national mean, "or at least at or above what would be expected given their poverty level" (based on school-level SAT-9 reading and math results for grades 3, 4, and 5).[50] Then, within each of these schools, teachers whose students performed above the national mean on the SAT-9 were selected as the "high performers." Data from the LESCP and site visits were then used to identify the instructional and classroom management practices in the high-performing classrooms compared to the other classrooms.

Unfortunately, as the Abt researchers acknowledge, there is no way to know whether their methodology adequately dealt with the other factors that could affect student performance, such as demographic and personal characteristics of the children and families, or earlier experiences in other classrooms or settings. (For example, classrooms were dropped when teachers reported that their students were "exceptional.") Moreover, the measure of high performance was too narrow, because it would not identify those teachers/classrooms that achieved substantial gains, but nevertheless remained below national norms.

LEARNING FROM FAILURE

That first Egg McMuffin that Kroc tasted was the culmination of almost one year's worth of recipe testing—which brings me back to the process of testing program or service ideas. The tendency to test only one program idea at a time (as happened in the JTPA and CCDP evaluations) elongates the learning process from years to decades—as we test and fail, and test and fail again, and again (hopefully getting a little closer to success each time).

R&D strategies should be planned with failure in mind, and they should be structured so lessons can be learned from those failures. Can you imagine telling a foundation official or political appointee that you really don't think this idea has more than a 50/50 chance of working? (The true figure, of course, is much lower.)

One important part of learning from failure is getting up close and identifying exactly what the program did (and did not do). Norton Grubb of the University of California–Berkeley describes just how close the evaluators must come to the program and how much they can learn by doing so:

> The existing evidence suggests two other conclusions. One is that the conventional evaluations of job training in the United States—based on random assignment methods, with outcomes compared for experimental and control groups—don't provide enough information. They say almost nothing about why such programs do or don't work, and

[48] For example, the Wisconsin Center for Education Research, a part of the University of Wisconsin–Madison, "develops, applies, and disseminates value-added and longitudinal research methods for evaluating the performance and effectiveness of teachers, schools, and educational programs and policies" (Wisconsin Center for Education Research, 2008; but see Amrein-Beardsley, 2008).
[49] Millsap et al., 2003.
[50] Millsap et al., 2003.

> therefore provide almost no guidance for administrators and program designers, or for policy-makers wanting to reform such programs. Only when there is supplementary information—for example, the interviews and information about depressive conditions available for New Chance, or the observational information available for CET—is there any possibility for learning why a program works or fails.
>
> Finally, these programs often look worse the closer one gets to them—for example, as one observes what goes on in classrooms, where it's often clear that the content is simplistic and the teaching quite mediocre. But without such understanding, it's impossible to know how to identify the reasons for failure, and therefore to recommend the appropriate changes.[51]

Planned variation designs (described below) are actually another way to learn from failure if they can parse out the separate impacts of particular program elements. Compare this to designs, such as the JTPA evaluation, where, as Orr and his colleagues explain:

> The study was designed to evaluate only the services normally provided by JTPA, not alternatives to JTPA. This means that the study is primarily *diagnostic*, not *prescriptive*. That is, although we can identify those parts of the program that have positive impacts and those that do not, we cannot say what alternative services would have worked better."[52]

TESTING MULTIPLE IDEAS

We should also do more testing of multiple ideas at once, perhaps through planned variation experiments. They test more than one idea at a time—thereby increasing the likelihood of finding positive impacts in a shorter period of time.

Under current practices, too much rides on individual, often long-running demonstration programs. Here's a particularly colorful description of the need to use the basketball rather than football approach to product development. It is a lesson for the development of social policy as well as American manufacturing.

> The key to making any manufactured product profitable these days is lowering the transactional costs of designing it. Look at Sony or Samsung or Apple or Honda. What these companies (really, groups of companies) have cultivated is the capacity to experiment. Product teams within each group design prototypes that will appeal to their niche customers. Company leaders then dump the likely losers, batting for singles and the odd home run. (Apple, remember, had no idea that the iPod would be a grand slam.) The point is, you don't want that much riding on each try. You want (if you'll pardon more sports metaphors) to transform your design-to-manufacturing paradigm from football to basketball—that is, to set yourself up to rush the basket many times per game, not painfully drive your way toward the end zone just a few times.[53]

One way to "rush the basket many times per game" is to test multiple ideas at the same time—and against each other as well as the status quo. Consider the Department of Education's 2004/2005 random assignment experiment to determine the effectiveness of 16 (yes, 16) educational software products. In this congressionally mandated study, a group of experts selected the 16 software products on prior evidence of effectiveness.[54] Then MPR and SRI tested the products in 33 school districts and 132 schools, with 439 teachers participating in the study. Within each school, teachers were randomly assigned to a treatment group, which used the study product, or to a control group, where teachers were to teach reading or math the way they otherwise would have. After one year, there were no statistically significant differences in test scores between classrooms using the selected reading and mathematics

[51] Grubb, 1999, p. 369.
[52] Orr et al., 1996, p. 214.
[53] Avishai, 2008, p. B03.
[54] Dynarski et al., 2007.

software products and those in the control group. These are unhappy results, but imagine where we would be if only one or a few products had been tested.

In essence, the welfare experiments accelerated learning because they tested two or more competing approaches against each other.[55] Although they did not reveal much about effective job training programs, they were able to compare the impact on one program component: "work first" (and other "labor force attachment strategies") with more traditional job training approaches (the "human capital development model"). The result, for better or worse, was the end of welfare as we knew it. Of course, other factors led to the passage first of JOBS and then TANF, but the impact of this research on the policy process has been well documented.[56]

A more formal planned variation experiment, that is, one that implements and compares two or more promising variations of a particular program,[57] can often offer a revealing look inside the proverbial programmatic black box. For example, under a Department of Labor contract, MPR conducted a random assignment study to assess the relative impact of three types of voucher programs for job training: (1) the Structured Customer Choice Model, where counselors played the major role in selecting training programs and the effective amount of the voucher for participants based on their views of program effectiveness; (2) the Guided Customer Choice Model, where participants had some counseling but made their own decisions about training choices, subject to a fixed voucher amount; and (3) the Maximum Customer Choice Model, where participants did not have to participate in any counseling and made their own decisions about training, subject to a fixed voucher amount.[58]

A total of 7,922 individuals at eight sites were randomly assigned between January 2002 and March 2004 to one of these three groups. There was no control group, so the evaluation could not assess the overall impact of any of these approaches. But what it did find was that, 15 months after random assignment, the employment, earnings, and other outcomes were essentially the same for all three groups, which suggests that the advice of counselors does not help make the WIA experience more valuable. My conclusion? Either we don't need the counselors, or that we need better counselors.

Note that these findings (which, by the way, are only the 15-month impacts) appeared seven long years after the contract was awarded.

Not all programs or questions lend themselves to planned variation, and the practical and cost requirements of mounting a successful planned variation have, no doubt, discouraged more attempts. But I hope we see more planned variations within existing programs, as it becomes progressively more difficult to isolate a meaningful zero-services control group. (In 2003, for example, about 25 percent of the control group in the Head Start Impact Study was in some other form of center-based care.)[59] In many cases, the current program can be the counterfactual. That presents another benefit, according to Whitehurst:

> Evaluations can be influential, even of widely deployed social programs, if the evaluations are designed not to disconfirm program effectiveness but to improve it. Thus the control group isn't the absence of the program but the current version of the program, while the intervention group is a planned variation that is hypothesized to improve outcomes. The threat of such an evaluation to advocates of the program is low because the results can't be used as an argument to shut down or reduce funding for the program. In short, focus on program improvement.[60]

[55] The utility of some welfare experiments, however, was compromised because two or more inconsistent program changes were tested at the same time on the same group.
[56] Baum, 1991; Haskins, 1991.
[57] See generally Rivlin & Timpane, 1975.
[58] McConnell et al., 2006.
[59] U.S. Department of Health and Human Services, 2005.
[60] Whitehurst, 2007.

DIVERSITY OF METHODS

I remain convinced that, all things being equal, randomized experiments are the most reliable way to attribute causation to particular programs or policies. But although randomized experiments may be the "gold standard" of program evaluation, diversification can be crucial for long-term returns (as the 2008 stock market declines reminded us all too well).

After years of extolling the virtues of randomized experiments over other forms of evaluation, our field should be more explicit about their frequently serious limitations, including limited generalizability, inability to capture community effects and saturation effects, contamination and substitution, randomization bias, only testing the intent to treat, high cost, and the often long period between initiation and findings.[61]

Hence, before closing, I would like to recognize the growing contribution of nonexperimental methods in the policy process. Sometimes even a simple before-and-after comparison works just fine to attribute causation. In discussing alternatives to random assignment, Thomas Cook and Peter Steiner suggest the following thought experiment: "Imagine Oprah Winfrey discussed your book on her show and you had extensive monthly data on prior and subsequent sales. A large spike would occur immediately after the interview, and critics probably could not develop plausible alternatives to explain such a large occurrence at this precise time point."[62]

Using an interrupted times series approach, the U.S. General Accounting Office concluded that limits placed on earnings disregards in the Aid to Families with Dependent Children (AFDC)[63] and other rule changes reduced the average AFDC-Basic monthly caseload by 493,000 families and lowered average monthly expenditures by $93 million.[64] Other analyses came to similar conclusions about OBRA's impact.[65]

Recent years have seen an explosion of more sophisticated nonexperimental work, much of it as helpful as the average randomized experiment (and sometimes more so)—especially since it usually is much more timely. Let me mention a few, although with a warning that some were more successful than others:

- Regression discontinuity to determine the impact of the Early Reading First program.[66]
- Propensity scoring to determine the impact of grade retention in kinde garten.[67]
- Difference-in-differences to estimate the impact of school district accountability policies on standardized state test scores in Chicago.[68]
- Fixed effects to estimate the relative impacts of welfare reform and the economy on the reduction of welfare caseloads in the 1990s.[69]
- Instrumental variables to estimate the impact of compulsory schooling laws on dropout rates,[70] and, as mentioned above,
- Interrupted times series to estimate the impact of the unemployed parent program and more generous earnings disregard on welfare caseloads and expenditures.[71]

[61] See generally Nathan, 2008.
[62] Cook & Steiner, 2009, 165–166.
[63] The changes were contained in the Omnibus Budget Reconciliation Act of 1981 (OBRA).
[64] U.S. General Accounting Office, 1984.
[65] Moffitt, 1984; Research Triangle Institute, 1983; Institute for Research on Poverty, 1985.
[66] Jackson et al., 2007.
[67] Hong & Raudenbush, 2005.
[68] Jacob, 2005.
[69] Council of Economic Advisors, 1997.
[70] Angrist & Krueger, 1991.
[71] U.S. General Accounting Office, 1984; Riggin & Ward-Zukerman, 1995.

In the right hands, nonexperimental approaches can hasten and enrich the development of knowledge. I apologize for the qualifier, but nonexperimental evaluations can raise special worries. Replication is usually difficult, and the analysis is often less than transparent. Put bluntly, the low barriers to entry (all one really needs is a laptop, a data set, and a rudimentary knowledge of statistical analysis) invite mischief: An advocate group can issue a report (or at least a press release) and be in the newspapers and on the Internet without the traditional protections of peer review.

As time goes on, I hope that we will develop more tools for assessing the trade-off between the slower and more precise results of randomized experiments and the quicker but perhaps less precise results of nonexperimental methods.[72] In the end, though, an analytic approach that combines methods, as appropriate, will probably be the most powerful tool for testing programmatic ideas and attributing causation, as Judy Gueron, formerly president of MDRC, describes in her Rossi Award lecture:

> I want to share three final thoughts. The first is that, while I obviously believe in the value of social experiments, I also am convinced that they are far from the only source of insight and evidence. If we are to advance policy and practice, we will need to draw on the insights of practitioners, managers (the M in APPAM), and those researchers who use different methods to diagnose the problems and to understand why people behave as they do and how social programs work in practice.
>
> Experiments are not an alternative to these sources of insight and innovation, but a way to confirm or disprove the expert judgment they suggest. In my experience, the strongest programs and evidence emerge when these different types of experts work together, using the strengths of their different fields.[73]

NO STIRRING CALL TO ACTION

I wish I could close with the usual call for action by President Obama and the Congress. But major change will be difficult to achieve. The problems that I have described are deep seated, and many are understandable accommodations to competing social and political priorities. The president, nevertheless, has one tool at his disposal that might improve federal R&D practices: OMB's Program Assessment Rating Tool (PART).

Continuous improvement government requires the development of learning organizations, in which managers are rewarded for figuring out what isn't working and for trying something else. Such a process must be driven by a force outside individual programs. That was some of the reasoning behind the Bush administration's creation of PART, which had the announced purpose of holding agencies "accountable for accomplishing results."[74]

Under PART, OMB evaluates the performance of particular federal programs based on governmentwide criteria. OMB's ratings—"Effective," "Moderately Effective," "Adequate," "Ineffective," and "Results Not Demonstrated"[75]—are then supposed to inform funding and management decisions.[76]

After a rocky start, and for all its flaws, PART seems to accomplish at least some of its goals.[77] It has provided the beginnings of a framework for encouraging rigorous R&D efforts—governmentwide. According to many observers (who are often critics of the specifics of PART), it provides an important incentive for agencies to evaluate their key programs (and provides guideposts for doing so). Then-Comptroller General David Walker testified before Congress that "PART helped to create or strengthen an evaluation culture within agencies by providing external motivation

[72] See, for example, Cook, Shadish, & Wong, 2008; Heckman & Hotz, 1989.
[73] Gueron, 2008.
[74] U.S. Office of Management and Budget, 2003b, p. 47.
[75] U.S. Office of Management and Budget, 2008.
[76] U.S. Office of Management and Budget, 2003a, p. 4.
[77] See, for example, Bavier, 2006; Nathan, 2005; Walker, 2007; Posner, 2007.

for program review and focused attention on performance measurement and its importance in daily program management."[78]

The Obama campaign issued a statement sharply critical of PART, complaining that its rating process is not transparent and is ideologically driven, and that the OMB examiners do not have the sufficient knowledge and understanding to fairly assess the programs. The statement promised that Obama would "fundamentally reconfigure PART."[79]

I hope that the Obama administration will recognize PART's usefulness as a tool to help strengthen federal R&D practices, and that it will build on its strengths while correcting its most serious weaknesses.

In closing, I console myself—and, I hope, readers—with three upbeat thoughts. First, although slow, progress is being made in building knowledge about numerous domestic programs, as I have tried to highlight in this talk. Second, we in APPAM are major players in this evolving and expanding mosaic of program evaluation. Last, we have learned enough since Rossi first propounded his Iron Law of Evaluation to offer a friendly update.

As I suggested earlier, Pete was always ambivalent about his Iron Law. He was human, so he enjoyed the attention it (and he) received, and he considered the objective, scientific method that it reflected to be one of the highest forms of social science. Nevertheless, as an unrepentant and proud liberal, Pete was deeply saddened by the disappointing results of so many evaluations, and the fact that evaluation had become a profoundly conservative force.[80]

So, with what I hope is due modesty, but also based on my years of friendship and discussions with that great bear of a man, I would like to close by taking the liberty of offering a friendly amendment to Pete's Iron Law of Evaluation:

> The expected value of any net impact assessment of any large scale social program is zero . . .
>
> . . . unless it systematically assesses the impact of services provided by innovators and outliers in comparison to those of other providers.

I think he would have approved.

Thank you.

DOUGLAS J. BESHAROV is a professor at the University of Maryland School of Public Policy and the Joseph J. and Violet Jacobs Scholar at the American Enterprise Institute for Public Policy Research. Douglas M. Call assisted Douglas Besharov in the preparation of this address.

REFERENCES

Aaron, H. J. (1978). Politics and the professors: The Great Society in perspective. Brookings Studies in Social Economics. Washington, DC: The Brookings Institution.

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System [electronic version]. Educational Researcher, 37, 65–75.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? [electronic version]. Quarterly Journal of Economics, 106, 979–1014.

Avishai, B. (2008, November 23). Why Detroit can't keep up [electronic version]. Washington Post, B03.

Bane, M. J. (1989). Welfare reform and mandatory versus voluntary work: Policy issue or management problem [electronic version]. Journal of Policy Analysis and Management, 8, 285–289.

[78] Walker, 2007, p. 5.
[79] Obama for America, 2008, [6].
[80] See generally Aaron, 1978.

Barnow, B. S. (2000). Exploring the relationship between performance management and program impact: A case study of the Job Training Partnership Act [electronic version]. Journal of Policy Analysis and Management, 19, 118–141.

Baum, E. B. (1991). When the witch doctors agree: The Family Support Act and social science research [electronic version]. Journal of Policy Analysis and Management, 10, 603–615.

Bavier, R. (2006). Workshop on assessing impacts of the Food Stamp Program. Paper presented at the U.S. Department of Agriculture, Washington, DC.

Besharov, D. J., & Germanis, P. (1999). Making food stamps part of welfare reform [electronic version]. Policy and Practice, 57, 6–11.

Besharov, D. J., Myers, J. A., & Morrow, J. S. (2007). Costs per child for early childhood education and care: Comparing Head Start, CCDF child care, and prekindergarten/preschool programs (2003/2004) [electronic version]. College Park, MD: Welfare Reform Academy. Retrieved December 11, 2008, from http://www.welfareacademy.org/pubs/childcare_edu/costperchild.pdf.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., et al. (1997). The benefits and costs of JTPA Title II-A Programs: Key findings from the National Job Training Partnership Act Study [electronic version]. Journal of Human Resources, 32, 549–576.

Card, D., & Krueger, A. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania [electronic version]. American Economic Review, 84, 772–793.

Child Trends. (2008). Guide to effective programs for children and youth: Nurse-family partnership. Retrieved December 17, 2008, from http://www.childtrends.org/lifecourse/programs/NurseHomeVisitingProgram.htm.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons [electronic version]. Journal of Policy Analysis and Management, 27, 724–750.

Cook, T. D., & Steiner, P. M. (2009). Some empirically valid alternatives to random assignment [electronic version]. Journal of Policy Analysis and Management, 28, 165–166.

Council of Economic Advisors. (1997). Technical report: Explaining the decline in welfare receipt, 1993–1996 [electronic version]. Washington, DC: Executive Office of the President. Retrieved December 15, 2008, from http://clinton4.nara.gov/WH/EOP/CEA/Welfare/Technical_Report.html.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? [electronic version]. American Economic Review, 85, 341–364.

Currie, J., & Thomas, D. (1996). Does Head Start help Hispanic children? [electronic version]. National Bureau of Economic Research Series. Cambridge, MA: National Bureau of Economic Research. Retrieved July 17, 2008, from http://www.nber.org/papers/w5805.pdf.

DeMuth, C. (1976). Deregulating the cities [electronic version]. The Public Interest, 44, 115–128.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., et al. (2007). Effectiveness of reading and mathematics software products: Findings from the first student cohort [electronic version]. Washington, DC: U.S. Department of Education. Retrieved December 15, 2008, from http://ies.ed.gov/ncee/pdf/20074006.pdf.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination [electronic version]. Prevention Science, 6, 151–175.

Food and Nutrition Act. (2008). 7 U.S.C. § 2026.

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start [electronic version]. American Economic Review, 92, 999–1012.

Gilliam, W. S., Ripple, C. H., Zigler, E. F., & Leiter, V. (2000). Evaluating child and family demonstration initiatives: Lessons from the Comprehensive Child Development Program [electronic version]. Early Childhood Research Quarterly, 15, 41–59.

Gilmour, J. B., & Lewis, D. E. (2006). Does performance budgeting work? An examination of the Office of Management and Budget's PART scores [electronic version]. Public Administration Review, 66, 742–752.

Grogger, J., Karoly, L. A., & Klerman, J. A. (2002). Consequences of welfare reform: A research synthesis [electronic version]. Santa Monica, CA: RAND Corporation. Retrieved December 15, 2008, from http://www.rand.org/pubs/drafts/DRU2676/.

Grubb, W. N. (1999). Lessons from education and training for youth: Five precepts [electronic version]. In Preparing youth for the 21st century: The transition from education to the labour market: Proceedings of the Washington D.C. conference, February 23–24, 1999 (pp. 363–383). Paris: Organisation for Economic Co-operation and Development.

Gueron, J. M. (1988). State welfare employment initiatives: Lessons from the 1980s [electronic version]. Focus, 11, 17–24.

Gueron, J. M. (2008). Acceptance remarks: 2008 Rossi Award. Retrieved December 12, 2008, from http://www.welfareacademy.org/rossi/2008_gueron_speech.shtml.

Gueron, J. M., & Pauly, E. (1991). From welfare to work. New York: Russell Sage Foundation.

Hamilton, G., Freedman, S., Gennetian, L., Michalopoulos, C., Walter, J., Adams-Ciardullo, D., et al. (2001). National evaluation of welfare-to-work strategies: How effective are different welfare-to-work approaches? Five-year adult and child impacts for eleven programs [electronic version]. New York: MDRC. Retrieved December 15, 2008, from http://aspe.hhs.gov/hsp/NEWWS/5yr-11prog01/index.htm.

Harvey, C., Camasso, M. J., & Jagannathan, R. (2000). Evaluating welfare reform waivers under Section 1115 [electronic version]. Journal of Economic Perspectives, 14, 165–188.

Haskins, R. (1991). Congress writes a law: Research and welfare reform [electronic version]. Journal of Policy Analysis and Management, 10, 616–632.

Heckman, J. J., & Hotz, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. Journal of the American Statistical Association, 84, 862–877.

Heinrich, C. J. (2007). Evidence-based policy and performance management: Challenges and prospects in two parallel movements. American Review of Public Administration, 37, 255–277.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics [electronic version]. Educational Evaluation and Policy Analysis, 27, 205–244.

Institute for Research on Poverty. (1985). Measuring the effects of the Reagan welfare changes on the work effort and well-being of single parents [electronic version]. Focus, 8, 1–8.

Jackson, R., McCoy, A., Pistorino, C., Wilkinson, A., Burghardt, J., Clark, M., et al. (2007). National evaluation of Early Reading First: Final report [electronic version]. Washington, DC: U.S. Department of Education. Retrieved December 15, 2008, from http://ies.ed.gov/ncee/pdf/20074007.pdf.

Jacob, B. (2005). Accountability, incentives and behavior: Evidence from school reform in Chicago [electronic version]. Journal of Public Economics, 89, 761–796.

Kroc, R., & Anderson, R. (1987). Grinding it out: The making of McDonald's. New York: St. Martin's Press.

Loeb, S., & Plank, D. N. (2008). Learning what works: Continuous improvement in California's education system (PACE Policy Brief 08-4) [electronic version]. Berkeley, CA: Policy Analysis for California Education. Retrieved December 17, 2008, from http://pace.berkeley.edu/reports/PB.08-4.pdf.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. (2005). How much is too much? The influence of preschool centers on children's social and cognitive development [electronic version]. National Bureau of Economic Research Working Paper. Cambridge, MA: National Bureau of Economic Research. Retrieved November 3, 2008, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=875688#.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design [electronic version]. Quarterly Journal of Economics, 122, 159–208.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2004). Does prekindergarten improve school preparation and performance [electronic version]. National Bureau of Economic Research Working Paper. Cambridge, MA: National Bureau of Economic Research. Retrieved December 11, 2008, from http://www.nber.org/papers/w10452.pdf?new_window=1.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E., Leibowitz, A., & Marquis, M. S. (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment [electronic version]. American Economic Review, 77, 251–277.

McCallion, G. (2006). Even Start: Funding controversy [electronic version].Washington, DC: Congressional Research Service. Retrieved January 15, 2008, from http://www.ed.psu.edu/goodlinginstitute/pdf/CRS_even_start_funding_controversy.pdf.

McConnell, S., Stuart, E., Fortson, K., Decker, P., Perez-Johnson, I., Harris, B., et al. (2006). Managing customers' training choices: Findings from the Individual Training Account Experiment [electronic version]. Washington DC: Mathematica Policy Research, Inc. Retrieved December 15, 2008, from http://www.mathematica-mpr.com/publications/PDFs/managecust.pdf.

Mead, L. M. (1990). Should workfare be mandatory? What the research says [electronic version]. Journal of Policy Analysis and Management, 9, 400–404.

Mead, L. M. (2005). Policy research: The field dimension [electronic version]. Policy Studies Journal, 33, 535–557.

Miller, G. (2003). Panel One: Overview of administration plan and reaction from Capitol Hill. Retrieved December 17, 2008, from http://www.brookings.edu/comm/events/20030507wrb.pdf.

Millsap, M. A., Brigham, N., Chase, A., & Layzer, C. (2003). Gaining ground: Instructional and management practices in high-performing, high-poverty elementary classrooms final report. Washington, DC: U.S. Department of Education.

Moffitt, R. (1984). Assessing the effects of the 1981 federal AFDC legislation on the work effort of women heading households: A framework for analysis and evidence to date. IRP Discussion Paper no. 742-A-84. Madison, WI: Institute for Research on Poverty.

Nathan, R. P. (2005). Presidential address: "Complexifying" performance oversight in America's governments [electronic version]. Journal of Policy Analysis and Management, 24, 207–215.

Nathan, R. P. (Ed.). (2007). Point/counterpoint: How should we read the evidence about Head Start? Three views [electronic version]. Journal of Policy Analysis and Management, 26, 673–689.

Nathan, R. P. (Ed.). (2008). Point/counterpoint: The role of random assignment in social policy research [electronic version]. Journal of Policy Analysis and Management, 27, 401–415.

Nyman, J. A. (2007). American health policy: Cracks in the foundation [electronic version]. Journal of Health Politics, Policy and Law, 32, 759–783.

Obama, B. (2008, September 22). The change we need in Washington [electronic version]. Retrieved December 17, 2008, from http://www.realclearpolitics.com/articles/2008/09/the_change_we_need_in_washingt.html.

Obama for America. (2008). Stop wasteful spending and curb influence of special interests so government can tackle our great challenges. Retrieved December 17, 2008, from http://www.governmentexecutive.com/pdfs/092208ts1.pdf.

Orr, L. L., Bloom, H. S., Bell, S. H., Doolittle, F., Lin, W., & Cave, G. (1996). Does training for the disadvantaged work? Evidence from the national JTPA study. Washington, DC: The Urban Institute Press.

Parsons, C., McCormick, J., & Nicholas, P. (2008, December 9). Barack Obama plans to reach out to Muslim world [electronic version]. Chicago Tribune. Retrieved December 17, 2008, from http://www.chicagotribune.com/news/politics/obama/chi-barack-obama-muslim-1210,0,5694976.story.

Posner, P. (2007). Performance budgeting: Preparing for a post-PART world. Testimony presented at Committee on the Budget, House of Representatives, Washington, DC.

Puma, M., Bell, S., Cook, R., Heid, C., & Lopez, M. (2005). Head Start impact study: First year findings [electronic version]. Washington, DC: U.S. Department of Health and Human

Services. Retrieved December 11, 2008, from http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/first_yr_finds.pdf.

Research Triangle Institute. (1983). Final report: Evaluation of the 1981 AFDC Amendments. Research Triangle Park, NC: RTI.

Riccio, J., Friedlander, D., Freedman, S., Farrell, M. E., Fellerath, V., Fox, S., et al. (1994). GAIN: Benefits, costs, and three-year impacts of a welfare-to-work program [electronic version]. New York: MDRC. Retrieved December 15, 2008, from http://www.mdrc.org/publications/175/full.pdf.

Riggin, L. J., & Ward-Zukerman, B. (1995). Effects on the AFDC-Basic caseload of providing welfare to two-parent families [electronic version]. Social Science Journal, 32, 265–278.

Rivlin, A. M., & Timpane, P. M. (Eds.). (1975). Planned variation in education: Should we give up or try harder? Washington, DC: The Brookings Institution.

Rogers-Dillon, R. (2004). The welfare experiments: Politics and policy evaluation. Palo Alto, CA: Stanford University Press.

Rossi, P. H. (1987). The iron law of evaluation and other metallic rules. In J. H. Miller & M. Lewis (Eds.), Research in social problems and public policy (pp. 3–20). Greenwich, CT: JAI Press.

Rumberger, R. W., & Tran, L. (2006). Preschool participation and the cognitive and social development of language minority students [electronic version]. CSE Technical Report. Los Angeles, CA: Center for the Study of Evaluation. Retrieved October 31, 2008, from http://lmri.ucsb.edu/publications/06_rumberger-tran.pdf.

Samuelson, R. J. (1998, February 23). Investing in our children: Sorry, government programs can't undo most of the ill effects of family breakdown [electronic version]. Newsweek, 131, 45.

Schochet, P. Z., & Burghardt, J. A. (2008). Do Job Corps performance measures track program impacts [electronic version]. Journal of Policy Analysis and Management, 27, 556–576.

Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does Job Corps work? Impact findings from the National Job Corps Study [electronic version]. American Economic Review, 98, 1864–1886.

Shoop, T. (2008, September 22). Obama pledges to fire managers, cut redundant programs [electronic version]. Government Executive. Retrieved December 17, 2008, from http://www.govexec.com/story_page.cfm?articleid=41022.

Snowe, O., & Clinton, H. R. (2007). Letter to U.S. Senate Appropriations Committee [electronic version]. Retrieved January 15, 2008, from http://www.evenstart.org/pdfs/Clinton-Snowe_Even_Start_letter_-_FINAL_w_sigs_-_05-03-07.pdf.

St. Pierre, R. G., Layzer, J. I., Goodson, B. D., & Bernstein, L. S. (1997). National impact evaluation of the Comprehensive Child Development Program: Final report. Cambridge, MA: Abt Associates Inc.

St. Pierre, R. G., Layzer, J. I., Goodson, B. D., & Bernstein, L. S. (1999). The effectiveness of comprehensive, case management interventions: Evidence from the national evaluation of the Comprehensive Child Development Program [electronic version]. American Journal of Evaluation, 20, 15–34.

St. Pierre, R., Ricciuti, A., Tao, F., Creps, C., Swartz, J., Lee, W., et al. (2003). Third national Even Start evaluation: Program impacts and implications for improvement. Cambridge, MA: Abt Associates Inc.

St. Pierre, R., Swartz, J., Gamse, B., Murray, S., Deck, D., & Nickel, P. (1995). National evaluation of the Even Start Family Literacy Program: Final report. Washington, DC: U.S. Department of Education.

Sweet, M. A., & Applebaum, M. I. (2004). Is home visiting an effective strategy? A meta-analytic review of home visiting programs for families with young children [electronic version]. Child Development, 75, 1435–1456.

U.S. Department of Health and Human Services. (1997). Setting the baseline: A report on state welfare waivers [electronic version]. Washington, DC: U.S. Department of Health and Human Services. Retrieved December 15, 2008, from http://aspe.hhs.gov/hsp/isp/waiver2/title.htm.

U.S. Department of Health and Human Services. (2005). Head Start Impact Study: First year findings. Washington, DC: U.S. Department of Health and Human Services.

U.S. Department of Health and Human Services. (2007). Fiscal year 2007 federal child care and related appropriations [electronic version]. Washington, DC: U.S. Department of Health and Human Services. Retrieved January 16, 2008, from http://www.acf.hhs.gov/programs/ccb/ccdf/appro07.pdf.

U.S. Department of Health and Human Services. (2008a). Building Strong Families (BSF): Overview. Retrieved December 17, 2008, from http://www.acf.hhs.gov/programs/opre/strengthen/build_fam/build_fam_overview.html.

U.S. Department of Health and Human Services. (2008b). Employment Retention and Advancement Project (ERA): Overview. Retrieved December 17, 2008, from http://www.acf.hhs.gov/programs/opre/welfare_employ/employ_retention/employ_reten_overview.html.

U.S. Department of Health and Human Services. (2008c). Fiscal year 2008 federal child care and related appropriations [electronic version]. Washington, DC: U.S. Department of Health and Human Services. Retrieved December 11, 2008, from http://www.acf.hhs.gov/programs/ccb/ccdf/appro08.pdf.

U.S. Department of Health and Human Services. (2008d). Head Start impact study and follow-up: Overview. Retrieved December 17, 2008, from http://www.acf.hhs.gov/programs/opre/hs/impact_study/imptstudy_overview.html#overview.

U.S. Department of Housing and Urban Development. (1996). Expanding housing choices for HUD-assisted families: First biennial report to Congress on the Moving to Opportunity for Fair Housing demonstration [electronic version]. Washington, DC: U.S. Department of Housing and Urban Development. Retrieved December 17, 2008, from http://www.huduser.org/publications/affhsg/expand/expand.html.

U.S. General Accounting Office. (1984). An evaluation of the 1981 AFDC changes: Initial analyses. Washington, DC: U.S. General Accounting Office.

U.S. General Accounting Office. (1997). Head Start: Research provides little information on impact of current program [electronic version]. Washington, DC: General Accounting Office. Retrieved December 15, 2008, from http://www.gao.gov/archive/1997/he97059.pdf.

U.S. Office of Management and Budget. (2003a). Budget of the United States government, fiscal year 2004 [electronic version]. Washington, DC: Government Printing Office. Retrieved December 15, 2008, from http://www.whitehouse.gov/omb/budget/fy2004/pdf/budget.pdf.

U.S. Office of Management and Budget. (2003b). Rating the performance of federal programs. In Budget of the United States of America: Fiscal year 2004 [electronic version]. Washington, DC: Government Printing Office. Retrieved, December 16, 2008, from http://www.gpoaccess.gov/usbudget/fy04/pdf/budget/performance.pdf.

U.S. Office of Management and Budget. (2008). ExpectMore.gov: Who we are. Retrieved December 16, 2008, from http://www.whitehouse.gov/omb/expectmore/about.html.

Walker, D. M. (2007). 21st century challenges: How performance budgeting can help. Testimony presented at Committee on the Budget, House of Representatives, Washington, DC.

Whitehurst, G. J. (2007). Acceptance remarks: 2007 Rossi Award. Retrieved December 11, 2008, from http://www.welfareacademy.org/rossi/2007_whitehurst_speech.shtml.

Wisconsin Center for Education Research. (2008). Value-Added Research Center works with TIF grantees. Retrieved December 17, 2008, from http://www.wcer.wisc.edu/index.php.