Trialing for the Public Good:
Building Effective Programs from the Ground Up

Mike Fishman, Thomas Gais, Michael Wiseman*

Abstract

The President's fiscal year 2014 budget emphasizes the creation of "a culture of performance improvement." In his 2013 IBM/National Academy of Public Administration sponsored review of GPRA goals, Donald Moynihan recommended that modernization should involve, among other things, building within government a "learning culture." In this paper we argue that rigorous experimentation is an important feature of a learning culture in government, and that lack of attention to experimentation as a means of doing government better is a significant shortcoming of the performance movement. Those promoting "evidence-based policy-making" acknowledge the importance of experimentation but generally source evidence outside of normal government operations. The posture of both the performance management and evidence-based policy enthusiasts contrasts with the growing acknowledgement of experimentation as an important tool of management in the private sector, where experimentation is coming to be seen as a normal function of good management within firms in many industries. We discuss reasons for the difference in experimentation between private and public management and the potential gains from encouraging more in government. We conclude with ideas for strategy.

1. The Problem

Public managers have long struggled to identify and implement effective policies and programs for accomplishing specific social welfare goals – reducing poverty, increasing employment and earnings, supporting learning and academic success. Over the past fifty years two primary mechanisms have emerged to support the identification and sound implementation of effective programs. Rigorous research, particularly randomized control trials (RCTs), generally supported by the federal government, has sought to answer the big policy and program design questions – does a particular policy or program, when well implemented, achieve its desired impact? These studies, expensive and time-consuming, have slowly built a body of evidence about what works. Unfortunately, we know more about what hasn't worked than what has.

Public managers and program operators cannot wait for rigorous evidence before making the many policy and program decisions they face daily as they deliver programs and services. While these rigorous studies may influence the design of those programs and services over the long

* Fishman (Mike.Fishman@MEFAssociates.com) is president of MEF Associates, Inc. Gais (Thomas.Gais@rockinst.suny.edu) directs the Rockefeller Institute of Government, State University of New York. Wiseman (WisemanM@GWU.edu) is Research Professor at the George Washington University. This is a draft prepared for the 2014 Fall Research Conference of the Association for Public Policy Analysis and Management in Albuquerque, New Mexico, November 6-8, 2014. This is a work in progress that should not be cited without a co-author's permission. Comments and corrections are solicited. The co-authors thank Jim Manzi, Chairman of Applied Predictive Technologies, for several fascinating and very helpful conversations on most of the issues discussed in this paper.

Draft: 4 November 2014

term, they are not the vehicle that managers use to monitor and improve program operations. Over time, states and localities have developed approaches for gauging how well they are delivering services and how those services are perceived by their customers. Those approaches generally include developing outcome and process measures and collecting data for assessing progress. Management by Objectives, Continuous Quality Improvement, Performance Indicators, Business Process Engineering, Data Warehouses and Dashboards, have all figured prominently in public sector service delivery. It is, however, important to note that these measures capture program outcomes such as participant employment and earnings that may not be causally connected to program participation. Further efforts to maximize measured outcomes may not focus effort where they will have the largest payoff. For example, research has shown that employment programs are more likely to benefit a more disadvantaged population; however, serving harder-to-employ participants is likely to result in lower overall employment outcomes.

Despite the increasingly sophisticated efforts to identify effective public sector programs and to deliver them effectively, we are still falling short of our ultimate goals. State and local administrators and their staff work feverishly to implement a broad array of programs and improve their operations relying on limited evidence of their programs' effectiveness while responding to increasingly fervent demands for results and accountability. Researchers and program evaluators use rigorous methods to identify effective programs – and find that most programs don't work. Some don't work because they are ill conceived; however attractive the theory or promising the apparent outcomes in early pilots, they fail to show impacts when tested using more rigorous methods – particularly randomized control trials. Many fail in rigorous tests because of poor implementation; not enough people who are offered or exposed to the program choose to participate or see the program through to its completion. Or the program services are simply poorly delivered. This failure of implementation is perhaps most vexing as the policy community is left to wonder whether the program, if better delivered, might have worked. And when researchers do identify effective programs, it is often in one or two cases, limiting their ability to be confident that the program will work when broadly implemented.

In the end, we collectively struggle to provide effective programs and services to those in need. And the public grows increasingly convinced that government programs are wasteful and ineffective.

This paper aims to make the case that public administrators, program operators, and researchers can do better. RCTs offer a simple and transparent approach for identifying program impacts in-house on an ongoing iterative basis. Such impacts help us identify what would have happened in the absence of a particular intervention or which of two or three competing approaches works best. They are easy to set up, and when grounded in existing administrative data, are relatively inexpensive to implement. When focused on proximate outcomes, such as increasing program take-up rates or program completion, they can also offer rapid results and enable ongoing experimentation in order to continuously refine program implementation. Building RCTs into existing, ongoing program performance improvement efforts can accelerate learning and lead to more effective programs.

Such a shift requires moving RCTs from the domain of researchers and evaluators to the world of program administrators and managers. The private sector has been using RCTs for continuous

quality improvement for many years. Jim Manzi's examples from the private sector provide ample evidence of their utilization and value (2012).

The sections that follow discuss the shortcomings of the performance measurement schemes, why efforts to move the research community toward evidence-based policy are necessary but insufficient to the task, how the private sector has incorporated experimentation into their continuous improvement initiatives and how the public sector might move to do the same.

2. The Accomplishments and Shortcomings of Performance Management

Since the enactment of the Government Performance and Results Act of 1993, performance has been the key concept in public management reforms, along with associated concepts such as performance measurement, management, and standards (Moynihan 2008). The basic idea is to judge public sector actions by outcomes. Governments establish measures that reflect progress (or its absence) in achieving broad goals. Agencies communicate these measures to managers, public employees, service providers, and others involved in implementing programs and policies. Governments collect and report data on such measures regularly. They may then use the data to make management decisions about strategies, budget allocations, personnel decisions, contract awards, and practices or policies (U.S. Government Accountability Office 2005, 2014). Aspirations may focus on changes in performance or on satisfying certain levels, benchmarks, or standards, or both. Measures are sometimes used to reward or punish service providers, as when outcome "milestones" trigger payments in performance-based contracts (Desai, Garabedian, and Snyder 2012).

Performance management has much to offer public managers. Developing outcome measures can help public officials think more clearly about what they are trying to achieve and help agencies create greater agreement about overall goals (if, of course, agreement is feasible). Some measures, if updated frequently and analyzed well, can show the location and distribution of the problems public officials are trying to alleviate and help managers allocate resources (Smith and Bratton 2001). Performance data can indicate whether the agency is making progress in reducing the targeted problems or in achieving other goals.

Performance measures, however, do not measure the impacts of specific government actions or changes in actions, such as changes in how governments implement a program or function. They do not estimate the net effect of public interventions, though that is how many public officials interpret them (Blalock and Barnow 2001). That is particularly true when many other variables—such as individuals' backgrounds and local economic conditions—probably affect the outcomes of interest, such as job placements in welfare-to-work programs, program enrollments, or completion of training programs. If, for example, a workforce agency alters its process for servicing clients and finds that its rate of job-placements changes afterwards, it may want to attribute the shift in performance levels to the administrative change. But other factors may have played a large or even larger role. Perhaps there was a coincident change in local job openings or the number of skilled persons looking for jobs (due to large, recent lay-offs in a local industry). Or perhaps the administrative change had an indirect and unintended effect, such as winnowing out hard-to-serve clients. The change in performance may flow from a change in the characteristics of clients enrolled in the program, not because the program has become more effective in assisting its traditional target population.

The challenge of discerning the effects of specific government actions on performance measures is exacerbated by the tendency of government agencies to implement many administrative and policy changes at roughly the same time. Political cycles contribute to this pattern, as governors, legislators, and top administrators see short "windows" of time within budget and election cycles to launch initiatives. Performance measures may even encourage this pattern of bundling multiple changes together into sporadic efforts at major reform rather than establishing a more continuous effort at innovation and testing. Uncertainty about the effects of single changes and the confounding effects of other variables may lead public administrators to throw several changes together in an effort to "move the needle" on major outcomes. But if many changes are implemented simultaneously, it is more difficult to discern the effects of any one innovation.

Using performance measures to understand the real effects of public actions is also difficult because the measures may be inaccurate, not due to typical measurement error but to systematic bias. Management efforts to improve performance often incentivize outcome measures. Front-line workers, supervisors, managers, service contractors, and top executives may have strong motives to show that outcomes within their scopes of responsibility are good and improving. Most public officials and contractors probably try to achieve better outcomes simply by doing what they can to increase the effectiveness of their activities. Yet some try to improve performance levels by misrepresenting the data, or by taking actions that increase aggregate outcomes but by means that undermine the program's purpose or population target (e.g., cream-skimming in workforce investment programs). As Heckman, Heinrich, and Smith noted, "The ability of program managers and staff to manipulate the data used to monitor them poses a major challenge to the successful design of performance standards systems" (2011, 5). In short, performance management may help incentivize public bureaucracies around a few key goals, but by loading incentives onto the measures, the government may motivate its personnel and contractors to game the programs or even falsify the data—thus making it even harder to discern what public actions really work.

PM may still be appropriate for many purposes and circumstances. Performance measures may be most useful in motivating and giving guidance to managers when the nature of their work requires considerable flexibility and adaptation to local or changing circumstances. Performance measures may also be useful when logic can eliminate alternative explanations of changes in outcomes—e.g., when an internal government process is clearly insulated from the effects of other variables. Performance measures may also be useful in providing accountability to the public and higher level officials.

Performance measurement and management, however, do not clearly lead to learning or innovation. They do not produce clear conclusions about the comparative effectiveness of alternative practices or processes, nor do these management tools encourage frequent, sequenced efforts at trying and testing new practices. By adding incentives to performance information, PM may in fact diminish the credibility of the data and make it even harder for public managers to interpret the information.

3. The Evidence-Based Policy Movement: Part, but Not All, of the Solution

In recent decades interest in improving the connection between research and policy has grown across a wide range of government activity. "Evidence-based policy" has become a watchword

for serious public management. A policy, students are told, is a plan of action for accomplishing some end. Evidence-based policy is public policy selected from among alternatives on the basis of rigorously established objective evidence of efficacy and superior efficiency.

Efforts to generate, evaluate, translate, and use rigorous evidence in government decisions have grown in breadth and strength in recent years. The U.S. Office of Management and Budget has pushed an "Evidence and Innovation Agenda," aimed at

> strengthening agencies' abilities to continually improve program performance by applying existing evidence about what works, generating new knowledge, and using experimentation and innovation to test new approaches to program delivery [2013].

OMB has promoted this agenda by encouraging federal agencies to "allocate resources to programs and practices backed by strong evidence of effectiveness while trimming activities that evidence shows are not effective." OMB will also look more favorably on new programs that promise to "yield credible evidence of program or policy impacts, for example by utilizing randomized control trials or carefully designed quasi-experimental techniques." Some federal agencies have been supporting similar goals for some years. The Institute of Education Sciences, established in 2002 within the U.S. Department of Education, has been particularly systematic in its efforts to increase the supply and dissemination of evidence regarding the effectiveness of education programs. Through major grants, IES has supported large-scale randomized controlled trials in a wide variety of areas, such as enhanced reading opportunities, middle school mathematics professional development, and mandatory random drug testing.

Private organizations have also played a big part in this movement. The Coalition for Evidence-Based Policy is the best-known recent example of an organization working both as intermediary and as advocate for both generation and use of evidence, with emphasis placed on results of randomized control trials. As stated on the CEBP website:

> The Coalition advocates many types of research to identify the most promising social interventions. However, a central theme of our advocacy . . . is that evidence of effectiveness generally cannot be considered definitive without ultimate confirmation in well-conducted randomized controlled trials.

CEBP's role as intermediary is signaled by its database of program trials and its evaluation of external validity—the reliability of forecasts of the results of introducing a similar innovation in a location other than that in which the demonstration is initially carried out. Innovations are evaluated by a panel of experts. Those considered to provide solid evidence of impact and promising evidence of transferability are divided into "Top" and "Near-Top" tiers. Top Tier interventions are those that have been tested through RCTs more than once and are thus expected to produce "sizable, sustained benefits to participants and/or society" (Coalition for Evidence-Based Policy 2014); Near Top Tier interventions just need an additional replication trial to confirm their initial, positive findings from an RCT.

The efforts of these and many other organizations have helped generate rapid growth in the number and rigor of RCTs and the dissemination of their results. This growth was built on an already rich history of social experiments in the U.S., dating back to the Negative Income Tax

experiments of the late 1960s and continuing through the AFDC waiver evaluations and other RCTs through the first years of this century (Gueron and Rolston 2013).

The value of this growing body of large-scale RCTs is undeniable. Thinking about welfare policies has changed, methodologies have been refined, and evaluation firms have developed great capacities to conduct rigorous RCTs under many circumstances and in many policy areas.

However, though we see many experiments, we find little experimentation. Sustained searches for solutions to public problems, guided by an iterative series of experiments that build on prior successes and failures, are rare in the public sector. If we want rigorous evidence to exert great influence over public actions, experimentation *within* government is needed. And for experimentation in government to happen, it must be useful to the problems public managers want to solve. Yet much of the Evidence-Based Policy movement is unlikely to nurture internal efforts of rigorous trial and error aimed at solving the problems that governments consider most important and pressing.

Large-scale RCTs do not lend themselves to sustained efforts at problem solving, innovation, and rigorous trial and error for several reasons. Such experiments take years to plan and execute; often the public officials involved in designing and authorizing the study are no longer in office when the results are in. But even if they are, changes in elected officials, economic conditions, budget constraints, federal initiatives, and other factors may have altered the relevance of the original questions and findings to current officials.

CEBP, IES, and other sophisticated efforts to promote evidence in public decision-making recognize the need to follow-up experiments with subsequent analyses—to replicate prior successes in different sites, try variations on theoretically promising yet unsuccessful interventions, or test a very different approach to achieving the same goals. But not only does a series of large-scale experiments take an even longer period of time, it is often not at all clear how one can interpret logically connected experimental results conducted at different times and in different places. Rigorous results for a program in one setting often fail to forecast findings for the same program in different settings (Cartwright and Hardie 2012), and it is often unclear why. External validity of interventions likely depends on many things, including sensitivity of outcomes to variation in the control environment, variations in intervention as actually delivered, and the moderating effect of the peculiar circumstances of the location of evidence-providing trials. Unless the results of each experiment in a long series of evaluations of a program or policy are each unambiguous and consistent with each other—an uncommon occurrence—it is no simple task to forecast the effects of future implementations of a intervention based on findings from several prior evaluations, even if they were well-conducted RCTs.

True, some rigorous evidence is better than none, but the responsible policy maker needs to think about how local factors might cause outcomes to differ. In this regard, it is important to keep in mind that advocates in medicine (the original inspiration for the EBP movement) commonly define evidence-based *practice* to mean "integrating individual clinical expertise with the best available external clinical evidence from systematic research" (Sackett 1996). Outside of medicine, this means the plans that are the core of evidence-based policy must be developed with an eye both to evidence and the circumstances of implementation. Here management must be engaged.

Two other, and inter-connected, problems with large-scale evaluations are their costs and provenance. The costs of the experiments are rarely borne exclusively by the implementing agency; they are typically provided by a higher level of government or philanthropic sources, or both. Moreover, in most cases the initial motivation for the initiative came from outside the agency, though successful implementation clearly required active agency participation. This locus of motivation is hardly surprising. Despite the challenges of generalizing results, funders usually want to use the results to build a "body of evidence" on outcomes they care about—outcomes they hope will benefit governments and citizens beyond the immediate site of the evaluation. In most cases the benefits to the implementing agency of conducting the experiment likely do not justify the costs of doing it "in a well-designed and implemented" way. Good experiments on major programs, in other words, provide "public goods," benefits that extend beyond the implementing authority. The focus is on big policy choices.

But fostering the experimental evidence base for major policy choices should not be the only public concern. One issue as well is how the public operations that deliver policy can best be conducted. As we move from asking *what* to do to achieve a particular policy end, we become engaged in questions of *how* to do it. Fifty years ago Harvey Leibenstein published an article on what he termed "X-efficiency" (Leibenstein 1966). He contrasted two sorts of efficiency. One, which he termed allocative efficiency, was the traditional focus of economics—how markets allocate production resources across goods and how market "failures," such as the under-production of public goods, might be best addressed. X-efficiency, on the other hand, involved failure to produce any good, with or without spillovers, at least cost. X-inefficiency arose when incentives were lacking for finding or utilizing the best techniques.

Leibenstein's focus was on the private sector, and critics argued that X-inefficiencies, should they exist, would be eliminated in a market economy by competition (Perelman 2011). The counter-argument turned on the behavior of managers in less than competitive markets and on the time required for relatively inefficient producers to be driven from the market. As tools improved for discovering opportunities for improving technique, competitive pressure on X-inefficiency would presumably grow.

But what of the public sector? Critics have long argued that however weak the incentives for production ("X" in Leibenstein's terms) efficiency in private sector markets, they are weaker in the public sector. The performance management movement can be interpreted in part as an effort to enhance such incentives. Surprisingly, in recent years the private sector has increasingly turned to experimentation—randomized control trials—to discover opportunities for increased efficiency and profitability. In contrast to experiments in the public sector, these experiments tend to work from the bottom-up, emphasizing internal management. Given the possibility that X-inefficiency is also a public sector issue, there may be lessons for public managers in private sector experience. After reviewing the randomized trial revolution in the private sector, we will argue that imitation in public management opens up the prospect of real experimentation within government—of constantly searching for and testing potential solutions to the problems that public agencies most care about.

4. The Experiment Revolution in the Private Sector

In his book *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*, Jim Manzi argues that experiments are central to a dynamic economy—indeed, a dynamic society—and that identification of a counterfactual against which experimental accomplishments can be assessed is the fundamental issue in experimentation. Where possible the best mode of evaluation is by randomized control trials in which units of observation are assigned at random between treatment and control groups. Randomization, if done successfully, ensures that "all other things equal" is made plausible and observed differences in outcomes may be interpreted as estimates of impact of the trialed innovation.

Uncontrolled emphasizes the shortcomings of other evaluation techniques applied in context of "high causal density" (i.e. multiple confounding influences on the outcomes of interest). Formulation of the control forecasts against which innovation outcomes (think of the indicators on performance management dashboards) can be measured takes analyst time and substantial operations history for extrapolation. Randomized trials provide the counterfactual simultaneously with performance; they provide the clearest picture of both the successes and the errors that in trial-and-error lead to further experimentation.

Further experimentation is important. In recent years business has experienced a major revolution in the use of randomized control trials in everything from advertising to employee incentives to production technique. The revolution is driven by success: businesses that make experiments central to decision making "earn a lot of money." It is facilitated by computational capacity and large data. And it is iterative. Success in one or two well-designed experiments in one or two sites leads to forecasts—presumptions about external validity—that then themselves become the objects of testing by experiment elsewhere.

This iterative process is a tactic growing out of, and interacting with, both business structure and business strategy. Structure determines whether or not there are enough units of analysis (customers, employees, stores, etc.) to support random assignment and provide statistical power. Testing takes resources; it is strategy that determines how such resources are best allocated. In context of strategy, experiments provide information about tactics. Experiments provide feedback on strategy impact. And experiments sometimes add degrees of freedom, opportunities to trial possibilities outside a firm's prevailing operating strategy. Iterative experiments—lots of them—are essential in a world in which many ostensibly promising ideas fail. Oddly, many failures can contribute to the learning environment essential to overall success. Innovators begin to know from experience when to shrug and move on. Large firms fully engaged in learning through experiment can literally engage in hundreds of randomized trials a year.

Such firms have commonalities. One is senior "political" sponsorship. A CEO or persons with similar responsibility to shareholders who genuinely believes in experiments as a means to enhanced profitability. The second is an independent unit within the firm responsible for design of experiments and the interpretation of results. The independent unit must be led by a testing enthusiast who is evangelical in promotion of experiments and commitment to the ideal. Finally, experimentation must be routinized. This means few if any big tests and lots of small ones, conducted in rapid succession, generally exploiting the firm's basic information technology systems. Continuous embedding of experiments in standard operations attenuates any

exceptional effects produced by the experimental environment itself and reduces disparity between how the treatment is delivered and what might be accomplished in a standard operations environment. But a particularly interesting features is the *dynamic*: the process of experimentation is iterative and continuous. One doesn't ask "what works," but rather "what is working," and "what should we trial next."

We call this experimental management in business. While RCTs are common in social policy, we know of no public organization that resembles either in leadership or operation the experiment-oriented firm that *Uncontrolled* describes. Why not? How do we get experimental management in government? Would we want it if we could?

5. Why Does the Public Sector Lag?

It is important to appreciate the distinction between the episodic, mostly large-scale program evaluations promoted by the Evidence-Based Policy movement to date and the activities central to *Uncontrolled.* Both involve randomized trials. Both are engaged in the accumulation of information pertinent to strategy. But while the major players in the EBP movement seek to build a well-tested body of knowledge about the effectiveness of various program models, experimentation in the private sector is a largely internal process for reducing uncertainty in many aspects of a firm's activities. It is also a flexible tool for managers, one that can be applied to near-term tactical problems as well as long-term questions. The *Uncontrolled* approach also diminishes the costs of failure when compared to the typical RCT evaluations, since the technique can be applied to questions less politically sensitive than entire programs or policies, and since one experimental result is often just a jumping-off point for the next experiment.

There are, of course, many possible reasons why this iterative, rigorous, trial-and-error approach to management has probably not spread among governments. It's a fairly new development, and few public managers are trained in its techniques and potential applications. It also involves a different dynamic than most public sector initiatives follow. Rather than piloting and testing each management or policy change separately, political cycles (e.g., budgets, elections, legislative sessions) often push many changes at the same time. Staff resources are also scarce in many cities and states, particularly after the unprecedented cuts in state and local government workers in the aftermath of the Great Recession.

Some of these challenges might be chipped away a bit if, to follow the lead of *Uncontrolled*, governments were to establish a public sector analog to the experimenting manager in business. In general terms, this would mean a chief executive officer in an organization that carries out government business of sufficient scale and character to provide enough units of analysis (again, customers, products, what-have-you) to support randomized trials. The organization would have clear objectives and metrics for assessing their attainment. The CEO would have competence for conducting the agency's external, generally political, relations and providing cover for experimental operations. The agency would include an analytic unit responsible for identifying opportunities of experimentation, designing the trials, assisting related division heads in implementation, and interpreting results. This unit would report directly to the CEO. Overtime, the agency would develop a culture of innovation in which experiments embedded within day-to-day operation were a central feature. Looking for such an agency is hardly a new idea. Over a quarter of a century ago Aaron Wildavsky identified as "the central problem of public

management" the development of an agency culture that continuously fosters effort at improvement and rewards those personnel who contribute to achieving this end. "Evaluation," he said, "is an obsession of such a culture" (Wildavsky 1987). While randomized trials were not part of the Wildavsky *zeitgeist*, the stretch to include them as a functional obsession for the kind of culture he was promoting seems small.

The problem is finding government organizations like this. There are some approximations; we mention them below. But they are decidedly rare. Reasons include the difference between what government does and what occurs in the private sector, difference in motivation of managers, structural differences in organization, and simple lack of capacity. It is possible that some processes for carrying out public business are simply not conducive to randomized trials of alternatives. Government is generally expected to provide uniform service to its constituents; randomized trials of alternatives could violate this norm. For business, profit is the ultimate performance indicator and capital markets create intense pressure for managers to find means of improvement. Motivators in the public sector can be complicated, and experimentation, if not routinized and politically appreciated, can be very risky. Compared to most government agencies, firms have very flat administrative structures. The larger number of steps from top to bottom complicates martialing enthusiasm for the challenge of randomized trials or for successful implementation of an innovation unit operating across agency divisions to promote experimentation. Creation of an innovation unit in most government agencies would be complicated by the difficulty of finding the needed leader/ technical enthusiast/ evangelist in an era in which resources for new appoints are in short supply and most positions are filled from within. Nevertheless, there seems little reason to believe that movement in the direction of experimental management is not feasible. As we discuss in the next section, we believe the benefits would justify the effort.

6. Potential Gains from Experimental Public Management

We anticipate the potential for reaping benefits on several levels as experimental management becomes broadly implemented. First, program implementation at the delivery level will improve incrementally as managers continuously test the impacts of specific program and policy strategies. Second, these incremental improvements will accelerate learning, spurring increased openness to testing new ideas and broad sharing of effective strategies and approaches both within and across programs and communities. Third, resource allocation decisions will begin to follow, moving resources to more effective programs and more importantly rewarding managers that are actively experimenting. Finally, public confidence and trust in the effectiveness of public management will grow as public services become more effective and the story of experimental management becomes part of the political conversation.

Incremental improvements at the delivery level may emerge in several ways. Testing side-by-side strategies for program implementation will yield definitive results. A particular participant recruitment or retention approach will either produce better, worse or the same results as a competing approach. Program managers will discard less effective approaches and most importantly continue to try to do better. A focus on optimization, rather than on hitting some arbitrary performance target, will yield efforts to better understand the dynamic behind the measured outcome.

The incorporation of behavioral economics into the design of public programs is providing a significant beachhead. Recognizing that people are not purely rational decision makers has led program developers to test approaches based on behavioral economics principals in real world situations. Can different messaging increase the response of imprisoned non-custodial parents to seek modification of their child support orders? Do different billing mechanisms increase child support payments? Do differently structured letters increase the number of welfare recipients who come in for meetings with their caseworkers?

Such questions are being explored, for example, via RCTs in the Administration for Children and Families (ACF) BIAS project. They test operational alternatives using administrative data to obtain rapid answers about at low cost. The behavioral economics approach starts by mapping the process targeted for improvement. Next they identify bottlenecks and breakdowns in the process and look for alternatives that take account of people's inherently flawed decision-making processes. Then, because it isn't always clear how people will respond to alternative options, they test alternatives using RTCs.

An example of an experimental outcome can be seen in the recent BIAS experiment in Texas:

> The state tested an approach for increasing the proportion of prisoners who apply for a modification of their child support orders. The BIAS project diagnosed bottlenecks in the application process, hypothesized behavioral reasons for the bottlenecks, and designed behaviorally informed changes to the mailing sent to incarcerated noncustodial parents. It revised the letter to make it more readable, printed it on blue paper so that it would stand out, pre-populated a section of the application, and sent a postcard before the letter was sent and another postcard following the letter to those who had not responded. While this was a low-cost effort (less than $2 per person), the revised outreach increased the application response rate to 39 percent, an 11 percentage point increase over the control group's response rate of roughly 28 percent. Program administrators hope that this is an important first step in a causal chain hypothesized to reduce child support arrears owed, leading, in turn, to an increase in the likelihood that, on release, parents will support their children [Farrell, et al. 2014].

As in the BIAS experience, one would hope that successful experiences with experimentation would increase their utilization spurring learning and innovation. This type of approach worked in Texas. Can it work in Washington State? BIAS is currently testing a similar strategy there. Texas might move on to examine whether an increase in order modification applications lead to an increase in order modifications. If not, how might one use the same systematic approach to test ways to achieve those increases? Or they might say a 39 percent application rate is great but how can we get to a 50 percent rate? Mechanisms would evolve to catalogue and share results to help managers build upon what they have learned in their own experiments as well as those done in other program domains or geographic areas.

Active experimentation and learning will capture the attention of budgeters. It makes sense to shift money to support more effective programs. As documented in the recent report on New York City's Center for Economic Opportunity, a culture emerged in that organization to pilot and rigorously evaluate new program ideas; eliminating funding for approaches that did not work and finding ways to more broadly implement those that showed promise (Gais, et al. 2014). One

would hope that a similar pattern might emerge as experimental management becomes more commonplace. While many factors do, and should influence, the allocation of public resources, placing a premium on funding what works and funding those in search of continuous improvement would be a welcome change.

The biggest payoff may be in shifting public perceptions about the motives and competence of public managers. There is widespread cynicism about government and its ability to spend public money effectively. Unfortunately, there are an abundance of examples to support that perception. While the growth in public experimentation cannot single-handedly reverse both the reality and perception of government, it may help to turn the tide. As a sage policy analyst at DHHS noted when discussing the messy and complex conundrum of welfare reform in the 1980's, there are no silver bullets for solving the problems of poverty and dependency among low-income single parent families; just lots of small answers that can help make a difference. The same is true for changing public perceptions with regard to the effectiveness of public management. Experimental management offers a method for continuously identifying and improving upon the effectiveness of public investments.

7. How Do We Increase Experimentation?

Increasing the role of experimentation in public management is a very different challenge than growing, classifying, summarizing, and disseminating a body of knowledge generated by RCTs and other rigorous research designs. While the EBP movement has made substantial progress with respect to the latter tasks, the former task—as it involves a transformation in governance—will be much harder. Fortunately, we have some clues from the private sector and even a few relevant experiences from public institutions.

As already noted, two common characteristics of private firms that use experimentation for continuous improvement are senior political sponsorship and an independent unit within the firm responsible for assisting managers with designing, implementing, and interpreting experiments. We found a rough approximation of these two elements in a recent study that two of us participated in of the New York City Center for Economic Opportunity (Gais, et al. 2014). Mayor Michael Bloomberg established the CEO in 2006 to find new and effective ways to lift New Yorkers out of poverty. The CEO worked with a wide range of city agencies and combinations of agencies to elicit new ideas for programs or modifications in programs, and it assisted with implementation and evaluation of the initiatives, which all began as small pilots. Although most of the evaluations were not RCTs, several were. But perhaps more important, nearly all of the initiatives were evaluated in some way, usually in an iterative fashion, with relatively simple program reviews early in the implementation cycle and later, more rigorous evaluations of pilots that survive the initial analyses. If the pilots appeared in the end to be effective, they might have been scaled up, depending of course on many other factors; if they were found to be ineffective and not fixable, they were terminated.

Over the years, the CEO became more of a general provider of technical assistance in evaluating programs across many policy areas. CEO's talented staff also helped promote an iterative, problem-solving process in agencies by meeting with city agency managers and clarifying what they knew and didn't know, and how they might answer the latter questions. CEO's efforts were greatly assisted by working directly under a deputy mayor and getting highly visible support

from the mayor—two formidable political sponsors. Just as important, both the mayor and deputy mayor repeatedly stressed that they wanted constant innovation and testing—and they not only tolerated failures but expected and wanted to see them as indicators that managers were really trying to find new approaches. Other cities are said to be adopting similar centers, though it's unclear whether they will serve the same function in promoting cycles of eliciting and testing ideas and building on those results. We do know, however, that the UK government is now discussing the establishment of a "Trialing and Experimenting Strategy Board," one that is intended to function within government much in the way that Manzi's experiments office is supposed to function in business.

There are, to be sure, several other conditions that would help make public management by experimentation a reality. The quality and accessibility of administrative data are essential. So is flexibility in procedures and policies, so that innovations can be developed and tried by managers without requiring lengthy negotiations with federal agencies. Here we can note a difference between the EBP movement and the "experimenting manager" approach. While federal waivers are viewed as useful by EBP advocates because they can require states to conduct major evaluations, to promote real continuous experimentation, it is best to have waivers that provide administrative and even policy flexibility, perhaps with regular reporting about what is being tried, learned, and finally adopted as a result.

We need to raise the level of understanding and awareness among public managers that RCTs are a useful tool for solving the typical problems of program management, not an arcane research method reserved for the use of evaluators. Several efforts are already making progress in that regard by promoting small-scale RCTs, particularly those that focus on measuring more proximate outcomes using administrative data to reduce the cost and duration of such experiments. CEBP sponsored a competition to fund three lower cost RCTs. The Administration for Children and Families (ACF) has funded an evaluation support contract, the development of administrative data sources, and a pool of academics to serve as a resource to state and local human service agencies. The White House, along with several Cabinet Departments, is sponsoring and supporting tests of interventions that target small changes in staff or participant behavior that can captured quickly with administrative data. The National Governors Association is sponsoring an initiative to promote the use of data and evidence for policy development and program management. The National Association of Welfare Research and Statistics (NAWRS) has sponsored a Research Academy at its annual workshop for the last three years to build knowledge and skills among state and local researchers. And the U.S. Department of Education has sponsored the development of a toolkit for conducting opportunistic evaluation, while the ACF is adapting it for use among human service programs.

Drawing on these public sector efforts as well as private sector experiences is an essential first step for building and disseminating a compelling argument for experimentation for public managers, including guidance regarding where and when experimentation is worth the effort. What questions can I answer with RCTs that cannot be answered by measuring outcomes and processes? How can this help my organization increase program effectiveness and efficiency? What are the costs and benefits? While making a general case for experimentation is a start, we need to make the case concretely in specific organizations. This requires not only summarizing the questions answered by past small-scale experiments but also learning what problems high-level public managers are trying to answer. For example, they may be much less interested in

external impacts of their programs than on finding ways to cut costs in the implementation of programs while not reducing performance levels. Can intake procedures be simplified and use fewer staff without increasing error rates or dropouts? Can electronic messages ensure that more clients show up at certain venues or satisfy other program requirements? What are the effects of subtracting a particular assessment or service on performance outcomes? Based on interviews with public managers and reviews of past and current efforts at small-scale experiments, a guide should be developed for disseminating the potential of experimental public administration among practitioners.

Another step is to demonstrate on the ground that RCTs add value. This will require finding a small group of governments or government agencies willing to try this approach long enough (perhaps three years) to assess its benefits and costs. Selecting the right managers and governments is essential in this step. The managers should be committed to the effort and have a challenging, measurable problem they want to solve. Top administrators—those who oversee the managers—should also show a commitment to letting the managers use RCTs to address the problem. The participating governments should have a strong data infrastructure and some flexibility in modifying administrative processes, allocating staff and other resources, communicating with clients, and other management tools—as well as in establishing new procedures for deciding when and where to conduct RCTs, for implementing the analyses, and for acting on their results. Intensive technical assistance and support from experienced evaluators will be key to this incubation process. It is also essential to be opportunistic and, above all, to be responsive to the problems *as the public managers see them,* even if that means promoting experimentation to find ways to cut program costs while ensuring an acceptable level of service.

Public honors and prizes as well as financial support for building capacity and adding staff, including a "champion for experimentation" within the government, may also help. But it is critical that outside funding is directed toward working with the executives in the organization and building capacity within core line functions; not focusing on research and evaluation shops (if they exist). State and local research staff can be a valuable resource, but the locus of the effort must be on operations – creating real time value for line staff.

It would then be important to build on these incubation sites. Once a few sites are engaged, the technical assistance team might facilitate workshops/meetings among the managers engaged in these initial efforts—to share ideas and experiences and begin to build a cadre of public managers who can help disseminate this approach to public management. If the incubation sites include the evaluation units discussed above, they too should share what they've learned about how to work across agencies with public managers in promoting the process of trial and error.

In all three steps, it is essential that efforts to promote experimentation take heed of ethical concerns. Many assert that it is unethical to conduct experiments on people, but there are many answers to this argument. If we don't know if a particular strategy works, why is it unethical to deny it to a randomly selected group to find out if it works? We don't have enough resources to provide service x to everyone – random assignment is an ethical way to distribute a scarce resource. We rarely introduce a no-service control group in current experiments but rather test alternative program approaches to see what works better. In the research world the use of Institutional Review Boards (IRBs) provide for an independent review of research proposals to

ensure that human subjects are adequately protected. In spurring the use of RCTs in non-traditional research settings, it will be important to create appropriate ethical standards for their use and oversight mechanisms.

8. Summary

Rigorous experiments can make great contributions to government decisions. However, much of the effort to promote them to date has not focused on the need to connect them in a process of experimentation, of constant trial and error within real governments.

The time is ripe for incorporating experimentation into mainstream public management. There is a solid base of research experience on which to build. The private sector has shown the way toward using experiments for performance management and continuous improvement. By the end of the decade we can take giant steps toward making the use of experiments in public management the normal way of doing business.

References

Baron, Jon. 2013a. "Demonstrating How Low-Cost Randomized Controlled Trials can Drive Effective Social Spending." Washington: Coalition for Evidence-Based Policy. URL: http://coalition4evidence.org/wp-content/uploads/2013/09/Low-cost-RCT-demonstration-concept-paper-7-10-13.pdf. Accessed 4 November 2013.

Baron, Jon. 2013b. Statement of Jon Baron before House Budget Committee Hearing on Progress in the War on Poverty. Committee on the Budget. House of Representatives. http://budget.house.gov/uploadedfiles/_jon_baron_testimony_.pdf.

Blalock, Ann B., and Burt S. Barnow. 2001. "Is the New Obsession with Performance Management Masking the Truth About Social Programs?" In *Quicker, Better, Cheaper: Managing Performance in American Government*. Edited by Dall W. Forsythe. Rockefeller Institute Press.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide for Doing It Better.* Oxford.

Coleman, Stephen. 1996. *The Minnesota Income Tax Compliance Experiment State Tax Results.* Minnesota Department of Revenue. URL: http://ssrn.com/abstract=1585242. Accessed November 4, 2013.

Decker, Paul. 2013. *False Choices, Policy Framing, and the Promise of "Big Data".* Presidential address, 2013 Research Conference, Association for Public Policy Analysis and Management, Washington DC, November 8. Forthcoming, Journal of Policy Analysis.

Desai, Swati; Lisa Garabedian; and Karl Snyder. 2012. "Performance-Based Contracts in New York City: Lessons Learned from Welfare to Work." *Rockefeller Institute Brief.* Albany: Rockefeller Institute of Government, SUNY.

Farrell, Mary; Caitlin Anzelone; Dan Cullinan; and Jessica Wille. 2014. *Taking the First Step: Using Behavioral Economics to Help Incarcerated Parents Apply for Child Support Order Modifications.* OPRE Report 2014-37. Washington: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Gais, Thomas; Patricia Strach; Katie Zuber; Mike Fishman; and Asaph Glosser. 2014. "Poverty and Evidence-Based Governance: The New York City Center for Economic Opportunity." *Rockefeller Report*. Albany: Rockefeller Institute of Government, SUNY.

Gueron, Judith M., and Howard Rolston. 2013. *Fighting for Reliable Evidence.* Russell Sage.

Haynes, Laura, Owain Service, Ben Goldacre, David Torgerson. 2012. *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*. London: UK Cabinet Office Behavioural Insights Team.

Heckman, James J., Carolyn J. Heinrich, and Jeffrey Smith. 2011. "Performance Standards and the Potential to Improve Government Performance." In Heckman, et al., *The Performance of Performance Standards*. Upjohn Institute.

Jakobsen, Morten and Simon Calmar Andersen. 2013. *Intensifying Social Exchange Relationships in Public Organizations: Evidence from a Randomized Field Experiment. Journal of Policy Analysis and Management*, 32 (1), 60–82.

John, Peter. 2014. "Changing bureaucrats and politicians: the transformative potential of an experimental public administration." Manuscript.

Leibenstein, Harvey. 1966. "Allocative Efficiency vs. X-Efficiency." *American Economic Review*, 56(3): 392–415.

Manzi, Jim. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. New York: Basic Books.

McKinsey & Company. 2009. *"And the winner is …": Capturing the promise of philanthropic prizes*. New York: The Company. URL: http://www.mckinseyonsociety.com/downloads/reports/Social-Innovation/And_the_winner_is.pdf. Accessed March 3, 2013.

Moynihan, Donald P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Georgetown.

Perelman, Michael. 2011. "Retrospectives: X-Efficiency." *Journal of Economic Perspectives*, 25(4), 211-222.

Sackett, David. 1996. "Evidence-based Medicine - What it is and what it isn't." *BMJ*, 312, 71-72.

Silver, Nate. 2012. *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't.* Penguin Press.

Smith, Dennis C., and William J. Bratton. 2001. "Performance Management in New York City: Compstat and the Revolution in Police Management." In *Quicker, Better, Cheaper: Managing Performance in American Government*. Edited by Dall W. Forsythe. Rockefeller Institute Press.

U.S. Government Accountability Office (GAO). 2005. *Managing for Results: Enhancing Agency Use of Performance Information for Management Decision Making.* Report GAO-05-927. Washington: GAO.

U.S. Government Accountability Office (GAO). 2014. *Managing for Results: Agencies' Trends in the Use of Performance Information to Make Decisions*. Report GAO-14-747. Washington: GAO

U.S. Office of Management and Budget (OMB). 2013. "Next Steps in the Evidence and Innovation Agenda." Memorandum M-13-17. Washington: The Agency. URL: http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-17.pdf. Accessed November 11, 2013.

Wildavsky, Aaron. 1987. *Speaking Truth to Power: The Art and Craft of Policy Analysis*. Piscataway, New Jersey, USA: Transaction Publishers.