CHAPTER 7

EVALUATION AS A TOOL FOR IMPROVING FEDERAL PROGRAMS

Since taking office, President Obama has emphasized the need to determine what works and what does not in government, and to use those answers to inform Federal policy and budget decisions. The President's 21st Century Management Agenda, submitted to Congress with the fiscal year 2010 Budget, set bold goals for building a more efficient, more effective government that contributes to economic growth and strengthens the foundations for economic prosperity (OMB 2009a). Today, evaluating Federal programs and interventions to understand their impact, and developing the infrastructure within agencies to support a sustained level of high-quality evaluations, remains an Administration priority. By rigorously testing which programs and interventions are most effective at achieving important goals, the government can improve its programs, scaling up the approaches that work best and modifying or discontinuing those that are less effective.

This Administration has supported the use of rigorous, high-quality "impact" evaluations to measure changes in a variety of outcomes targeted by Federal programs, ranging from earnings to health to electricity usage. Many factors affect whether Federal programs achieve their goals, and identifying impacts specifically attributable to programs is challenging. An impact evaluation is a particular type of program evaluation, and aims to measure the causal effect of a program or intervention on important program outcomes. This chapter focuses on impact evaluations. "Process" evaluations (another type of program evaluation) and performance measurement also contribute to building evidence about how well programs are working, but differ in important ways from impact evaluations (Box 7-1).

Building on the efforts of previous administrations, the Obama Administration is working to reform the Federal Government's approach to improving program performance. In addition to emphasizing transparency and accountability in tracking progress toward agencies' priority goals, this new approach also aims to complement and to draw on the Administration's

Box 7-1: Impact Evaluations, Process Evaluations, and Performance Measurement

Program managers use many approaches to assess how programs operate and how well they work. Impact evaluations aim to identify the causal effects of a program or intervention on some outcome or outcomes of interest. Impact evaluations are distinct from other types of program evaluation and performance measurement. For example:

- **Process evaluations** analyze the effectiveness of how programs deliver services relative to program design, professional standards, or regulatory requirements. For instance, a process evaluation might focus on whether a program is reaching the target number of participants or whether caseworkers are consistently following a specified protocol for providing services. Process evaluations help ensure that programs are running as intended, but in general these evaluations do not directly examine whether programs are achieving their outcome goals (GAO 2011).
- **Performance measurement** is a broader category that encompasses "the ongoing monitoring and reporting of program accomplishments, particularly progress toward pre-established goals" (GAO 2011). Typically, performance measures provide a descriptive picture of how a program is functioning and how participants are faring on various "intermediate" outcomes, but do not attempt to rigorously identify the causal effects of the program. For instance, performance measures for a job training program might capture how many individuals are served, what fraction complete the training, and what fraction are employed a year later. But these measures will not answer the question of how much higher these individuals' employment rates are as a result of having completed the training. Nonetheless, performance measures serve as important indicators of program accomplishments and can help establish that a program is producing apparently promising (or troubling) outcomes.

While process evaluations and performance measurement are useful at all stages of a program's maturity, they can be particularly useful for providing evidence about how programs are working in the early years of a program's history when impacts on program outcomes may not be detectable and rigorous, high-quality impact evaluations are not possible. A logic model—a tool that depicts the intended links between program investments and outcomes and helps to ensure program activities will achieve desired outcomes—can facilitate agency efforts to develop high-quality "intermediate" indicators of impacts as well as an understanding of alternative causal channels that can affect important program outcomes.

program evaluation efforts. For example, the Administration this year is establishing strategic reviews within agencies to strengthen the use of evidence in strategic and budget decisions.

This chapter provides an overview of the implementation and use of impact evaluation in Federal programs, with a special focus on the lessons learned so far in this Administration. It begins with a discussion of some challenges inherent in conducting rigorous impact evaluations in government programs. The chapter then focuses on Administration efforts to build and to use evidence, including actions taken on lessons learned from completed evaluations, launching new evaluations in areas where not enough is known, and creating a culture of evidence-building in Federal programs, especially grant programs. The final section identifies opportunities for further progress: for example, through increasing legislative support and removing legislative barriers, embedding evaluation into routine program operations, and using existing program data to measure outcomes and impacts.

Conducting Rigorous Impact Evaluations in Federal Programs

Science, business, and government routinely confront the problem of ascertaining the effect of a program, policy, or initiative. Is a newly developed drug effective in treating the condition for which it was developed? Does a new marketing strategy boost sales? Does a preschool program improve participants' outcomes, such as success in elementary school? Despite the different settings, these questions all focus on measuring the effect of an intervention or program on one or more outcomes of interest.

One basic approach to answering questions like these is to look at outcomes before and after the "treatment"—for instance, before and after taking a drug, before and after a new marketing strategy is rolled out, or before and after participation in an education program. Another straightforward approach is to compare outcomes for program participants with outcomes for non-participants. In complex policy environments, however, these simple approaches will often give the wrong answers. Take, for example, a job training program designed to help unemployed workers get jobs. The data may show that program participants were much more likely to be employed a year after the training program than before they entered the program. But if the unemployment rate has fallen substantially over the course of the program, then the gains may be due to the improving economy, not to the training program. Similarly, a government program offering start-up assistance to new businesses may appear to boost success rates. But if capable entrepreneurs are more likely than less capable ones to participate in the program, then self-selection of program participants, not the program itself, may be driving those better outcomes.

A strong impact evaluation needs a strategy for constructing more valid comparisons—specifically, for identifying "treatment" and "control" groups for which differences in outcomes can reasonably be attributed to the program or intervention rather than to some other factor. Impact evaluations conducted using rigorous, high-quality methods provide the greatest confidence that observed changes in outcomes targeted by the program are indeed attributable to the program or intervention. It is well recognized within Congress and other branches of government (for example, GAO 2012, National Research Council 2009), in the private sector (Manzi 2012), in non-governmental research organizations (Coalition for Evidence-Based Policy 2012, Walker et al. 2006), and in academia (for example, Imbens 2010; Angrist and Krueger 1999; Burtless 1995) that evaluations measuring impacts on outcomes using random assignment provide the most definitive evidence of program effectiveness.

Although the classic impact evaluation design entails random assignment of recipients into treatment and control groups as part of the experiment, the goal of constructing valid comparisons sometimes can be achieved by taking advantage of natural variation that produces as-if randomness, an approach referred to as a quasi-experiment. Quasi-experiments can sometimes be much less expensive than traditional large-scale random assignment experiments, and are discussed further below.

Estimation of Causal Effects of a Program or Intervention

The starting point for estimating the causal effect of a program or intervention is being precise about what constitutes a causal effect. Consider a treatment delivered at the individual level: either the individual received the treatment, or did not. The difference between the potential outcome if the individual received the treatment and the potential outcome if the individual did not is the effect of the treatment on the individual.[1] The challenge of estimating this treatment effect stems from the fact that any given individual either receives the treatment or does not (for example, a child either does or does not attend preschool). Thus, for any given person only one of two potential outcomes can be observed. The fact that we cannot directly observe the counterfactual outcome (for example, the earnings a person who

[1] No two individuals are the same, so in general the effect of a program or intervention differs from one individual to the next. For example, the effect of the preschool program could depend on the child's learning opportunities at home. Impact evaluation typically focuses on estimating an average causal effect, which is the average of the individual-level causal effects.

went to preschool would have had if they had, in fact, not gone to preschool) implies that we cannot directly measure the causal effect. This problem of observing only one of the potential outcomes for any given individual is the fundamental problem of causal inference (Holland 1986).

Randomization provides a solution to the problem of not observing the counterfactual outcome. If individuals are randomly assigned to treatment and control groups, then on average the individuals in the two groups are likely to be the same in terms of other characteristics that might affect outcomes. As a result, one can safely assume that ex-post differences between the groups are the result of the treatment. To take the preschool example, simply comparing test scores of all U.S. elementary school children who had attended preschool to all U.S. elementary school children who had not would not provide confidence that higher test scores for the first group were an effect of preschool. The scores might reflect differences in family background, elementary school resources, or other important factors between the two groups. On the other hand, if a group of three-year olds are randomly assigned to attend or not attend preschool, and the preschool group has higher test scores in third grade, we can attribute the test score gains to attending preschool because the two groups would not be systematically different along other dimensions that might impact learning.

In most cases, simple comparisons of treated and untreated individuals without random assignment will not produce valid comparisons because treatment status will be correlated with other important factors. For example, if potential preschool enrollees were initially screened so that those with the least learning opportunities outside school were placed in the program, then we might find that the treatment group (enrollees) has worse outcomes than the control group. However, the reason for this finding is that enrollees are more disadvantaged than non-enrollees. The variation between treatment and control groups affects ultimate outcomes both through the treatment and the differences in learning opportunities outside school. Thus, any comparison of outcomes between treatment and control groups would measure the combined effects of both the treatment and those differences in learning opportunities.

Because randomized experiments can be expensive or infeasible, researchers have also developed methods to use as-if random variation in what is known as a quasi-experiment. The necessary condition for a high-quality quasi-experimental design is that people are assigned to a treatment or control group in a way that mimics randomness. This can be done by forming treatment and control groups whose individuals have similar observable characteristics, and exploiting some rule that governs assignment to the treatment and control groups in a way that is plausibly unrelated to the outcome of interest.

One example of a quasi-experimental design that lends itself to estimating impacts of programs or interventions is when eligibility is determined based on one or more variables in a way that individuals who (just) qualify for the program are very much like those who (just) do not. If so, and if both eligible and ineligible applicants are tracked, then a method called regression discontinuity design can be used to compare the outcomes for individuals on the two sides of the threshold, controlling for other observable differences between the two groups.

Another example of quasi-experimental design is when a program varies across units for reasons unrelated to the program outcomes. Rothstein (2011), for example, exploits the fact that, due to different business cycle patterns combined with policy variation created by expirations and renewals of the Emergency Unemployment Compensation (EUC) program during the Great Recession, the number of available weeks of benefits available to jobseekers varied dramatically from month to month in differing ways across states. After controlling for local economic conditions, the haphazard nature of the changes in EUC benefit levels across states enabled estimation of EUC benefits on job-finding rates.

Describing the whole range of quasi-experimental approaches is beyond the scope of this chapter.[2] Quasi-experiments require stronger assumptions than randomized experiments and the debate around those assumptions makes it harder for quasi-experiments to be convincing, especially to non-experts. However, if the quasi-experimental variation used is plausibly unrelated to the outcomes of interest except through the treatment, quasi-experimental evidence can be convincing, with some methods and applications being nearly as compelling as randomized trials and others leaving more room for doubt.

Other Criteria for High-Quality, Successful Impact Evaluations

A strong impact evaluation also needs to address questions that are actionable and relate to outcomes that matter. In some cases, the actionable information might identify if a program is or is not effective. In other cases, the actionable information might identify which interventions are best at achieving important program outcomes, so that programs can be improved

[2] For more extensive introductions to impact evaluation (both randomized experiments and quasi-experiments), see Angrist and Pischke (2008, ch. 1) and Stock and Watson (2010, ch. 13). Shadish, Cook, and Campbell (2002) and Berk and Rossi (1998) provide more advanced textbook treatments, and Imbens and Wooldridge (2009) provide a survey of recent methodological developments in the field.

by adopting successful interventions more broadly. However, if there are legal or other impediments to expanding an evaluated small-scale intervention, then learning that the intervention works does not directly lead to an action that can improve a program at a national level. In such cases, it may be better to allocate scarce evaluation resources to testing more modest interventions or ways to run the program more effectively.

For the second of these criteria—outcomes that matter—the long-term goals of a program must be considered. For a preschool program, the number of students enrolled is an easy-to-measure intermediate outcome. However, preschool enrollment may or may not be related to ultimate outcomes, such as high school graduation rates, employment rates, or income. It is also important to consider program size and stage of development, as programs or interventions must be sufficiently mature, and treatment and control groups sufficiently large, to obtain credible estimates of impact.

Other issues must also be addressed to conduct policy-relevant impact evaluations in government programs. At the most practical level, rigorous evaluation requires adequate funding, staff expertise, and often cooperation across different parts of an agency (or across multiple agencies). Rigorous evaluation also requires support from top agency management and program managers. Further, many Federal programs have multiple goals, which can make it hard to take action on evaluation findings when the results support some goals but not others.

An important part of evaluating a program is remaining open to the findings, regardless of the outcome, to inform the best course of action to improve outcomes going forward. Findings of positive impacts provide important feedback that may indicate whether additional investment is warranted. Findings of no impact, either for all participants and program goals or for important subsets of the participants and program goals, also send valuable signals that modifications—including reallocating program funding to other strategies that could better achieve outcomes—are needed.

Lower-Cost Ways for Impact Evaluations to Facilitate Real-Time Learning

Large-scale random-assignment studies of social programs have been very influential, but also can be quite expensive, and their expense has been a major impediment to wide-scale adoption of learning and program improvement through randomization. For this reason, researchers have focused on lower-cost methods for learning about program effectiveness.

One lower-cost method is to build randomization into the design of the program, so that data on program performance can be tracked and evaluated on an ongoing basis. This strategy has been pioneered as a management tool in the private sector for ongoing product and process improvement. Indeed, some companies run thousands of randomized studies annually: by 2000, Capital One was running 60,000 studies annually using randomization methods, as they experimented with different strategies to determine what works. Google has also run randomized experiments in the tens of thousands in some years (Manzi 2012).

In the public sector, Federal agencies are also finding ways to conduct high-quality evaluation strategies at lower cost, including ways that employ the lessons learned from behavioral economics (Box 7-2). The U. S. Department of Agriculture's Food and Nutrition Service is conducting a range of rigorous demonstration projects to further develop the evidence base of effective strategies for programs that address food insecurity and improve nutrition among children; one such project implements low-cost environmental changes in lunchrooms to encourage students to make healthier food choices. One demonstration found that merely placing fruit in a colorful bowl in a convenient part of the lunch line can lead to an increase in fruit sales of up to 102 percent (Wansink, Just, and Smith 2011). Funding research for these simple, evidence-based interventions allows for the development of effective strategies to strengthen the nutrition and hunger safety net for the more than 30 million children fed by the National School Lunch Program.

Utilizing existing data and independent programmatic changes to measure outcomes is another strategy that agencies are using to minimize evaluation costs. For example, the Department of Justice's National Institute of Justice conducts impact evaluations of interventions that can help inform the approximately 18,000 local law enforcement agencies that do not individually have the resources to test interventions on their own. Hawaii's Opportunity Probation with Enforcement (HOPE) program was established as a demonstration pilot for drug-involved probationers in Hawaii. The pilot tested the efficacy of "swift and certain" sanctions against probationers who fail to meet the conditions of their probation. The randomized controlled experiment found that after one year, probationers who received very frequent drug testing (every other day) and—if they failed the drug test—an immediate court date and a modest but certain sanction (a night in jail), were 72 percent less likely to use drugs, 61 percent less likely to skip meetings with their supervisory officers, 55 percent less likely to be arrested for a new crime, and 53 percent less likely to have their probation revoked. These reductions led to HOPE participants being sentenced to an average of 48 fewer days in prison than those in the control group who received the traditional delayed but more severe sentence (National Institute of Justice). Because of the high costs associated with servicing inmates in prison,

Box 7-2: Using Behavioral Economics to Inform Potential Program Improvements

Increasingly, agencies are using insights from behavioral science to implement low-cost evaluations that can be used to improve program design. Utilizing randomized experiments or other rigorous evaluation designs, these studies examine aspects of program operations that can be re-designed to help people take better advantage of available programs and services—such as by simplifying application processes or highlighting the availability of student financial aid. Recently, the White House Office of Science and Technology Policy assembled a cross-agency team of behavioral science and evaluation experts—the U.S. Social and Behavioral Sciences Team—to help agencies identify promising opportunities for embedding behavioral insights into program designs and to provide the necessary technical tools to rigorously evaluate impact. Such low-cost, real-time experiments can help Federal programs operate more effectively and efficiently.

any intervention that reduces prison time can generate large savings. By making use of available administrative data, evaluations employing quasi-experimental and randomized controlled trial designs were implemented at a cost of only $150,000 and $230,000, respectively.[3] A follow-up analysis is examining the long-term impacts of the intervention; this model is also being piloted in four other locations.

The often-lengthy time between implementation and results of a rigorous evaluation can also discourage its use, but agencies are looking for ways to speed up the evaluation process to gain actionable insights more quickly. For example, the Center for Medicare and Medicaid Innovation (the "Innovation Center"), which was created by the Affordable Care Act of 2010, is using an innovative "Rapid Cycle" approach and high-quality evaluation methods to develop and test innovative payment and service delivery models designed to reduce expenditures while preserving or enhancing quality of care for Medicare, Medicaid and Children's Health Insurance Program beneficiaries. By giving more rapid feedback to health providers, as Box 7-3 shows, the Rapid Cycle approach provides actionable information, allows for more frequent course corrections, and supports continuous quality improvement (Shrank 2013).

[3] Cost estimates supplied by the Department of Justice's National Institute of Justice.

Impact of the Evidence-Based Agenda

From its first months, the Administration embedded a strong evaluation focus into many new initiatives to learn what strategies work best and to scale up approaches backed by strong evidence. During the formulation of the FY 2011 Budget in fall 2009, the Office of Management and Budget (OMB) invited agencies to submit new evaluation proposals for building rigorous evidence and also encouraged agencies to demonstrate that new program initiatives were based on credible evidence of success or to include plans to collect evidence where none exists. The Administration has maintained its emphasis on using and building evidence in every subsequent budget (OMB 2010, 2011, 2012, 2013a, 2013b).

Uses of Evaluation

Agencies have used impact evaluations to inform policy and program decisions in a wide variety of ways.

Making the Unemployment Insurance System More Effective. Unemployment insurance (UI) provides an important safety net for workers who become unemployed. Occasionally, concerns are raised that UI payments could reduce an unemployed worker's incentive to find employment. While the evidence suggests that any such effects are small (Council of Economic Advisers and the Department of Labor 2013), the Federal Reemployment and Eligibility Assessment (REA) initiative started providing funds in 2005 to states and sought to reduce UI duration by combining in-person UI eligibility reviews with (1) labor market information, (2) developing a reemployment plan, and (3) offering a referral to reemployment services. The Department of Labor funded research using a randomized design that showed the REA initiative was effective in reducing the duration of UI (Benus et al. 2008). However, these studies focused on measuring reduced duration on UI and associated costs and not on other outcomes, such as return to employment or increased wages. These studies were followed by another randomized controlled trial which showed that the REAs were also effective at reducing joblessness when eligibility assessments were personalized and more closely integrated with the delivery of reemployment services (Poe-Yamagata et al. 2011). Consequently, the Administration proposed in the American Jobs Act to create a requirement that all Emergency Unemployment Compensation claimants receive both an REA and reemployment services; this was enacted in the Middle Class Tax Relief and Job Creation Act of 2012.[4] Evidence from rigorous evaluations is playing a role in making the REA initiative more effective and getting

[4] Public Law 112-96.

Box 7-3: "Rapid Cycle" Evaluations in the Center for Medicare and Medicaid Innovation

The Center for Medicare and Medicaid Innovation (the "Innovation Center") has invested in building information systems and institutional capacity that permit a "Rapid Cycle" approach to testing a variety of new models. For instance, evaluators can gather real-time information and provide performance data to providers who adopted the same model, allowing these providers to understand and track their own performance, and to compare their performance with that of other providers. In the early stages of model implementation, the Innovation Center is applying this Rapid Cycle approach primarily to process evaluations (see Box 7-1), turning later to assessing the models' overall impacts (including impacts on important subgroups, where feasible) once there is reasonable assurance that the model has been in operation sufficiently long to detect impacts. When experimental conditions cannot be met due to logistical or other constraints, quasi-experimental methods are used, with multiple comparison groups for each treatment group identified where possible to provide robustness checks of the findings (Shrank 2013).

unemployed Americans back to work faster, and the Administration has sought to expand it to cover more workers. A modest increase in funding for REAs was included in the recently enacted Consolidated Appropriations Act of 2014.

Simplifying Applications for Student Aid. In many cases, actionable evidence on what works comes from field-generated, grant-funded research rather than from Federal program evaluations. In 2008, with support from the Department of Education's Federal Student Aid Office, Institute of Education Sciences, and other funders, university-based researchers worked with H&R Block to set up an experiment providing randomly selected low-income tax filers in North Carolina and Ohio with pre-populated Free Application for Federal Student Aid (FAFSA) forms and FAFSA assistance for themselves or their children, as well as with information about student aid. This relatively low-cost intervention had a surprisingly large effect on college enrollment outcomes. For example, college enrollment rates for high school seniors and recent high school graduates who received this help rose by about 25 percent—from 34 to 42 percent. Moreover, these gains persisted over time: three years after the intervention, treatment group students were 8 percentage points more likely to have been enrolled in college for at least two consecutive years (Bettinger et al. 2012).

The study's findings helped spur many important policy changes. Most notably, students and their families now have the option to pre-populate the FAFSA with the income information they have already provided the Internal Revenue Service on their tax returns, similar to the arrangement the researchers tested with H&R Block. This simplifies FAFSA completion for students, lowers the risk of errors, and as such should increase access to college among socioeconomically disadvantaged students and should lead to gains in college enrollment. As a complement, the Department of Education has also simplified the FAFSA application to make it easier to complete for all applicants, but especially for low-income students. In 2012, the Department of Education awarded a second grant to the research team to test the effects of FAFSA simplification at scale. The evaluation will use an experimental design and involve 9,000 tax-filing sites across the United States.

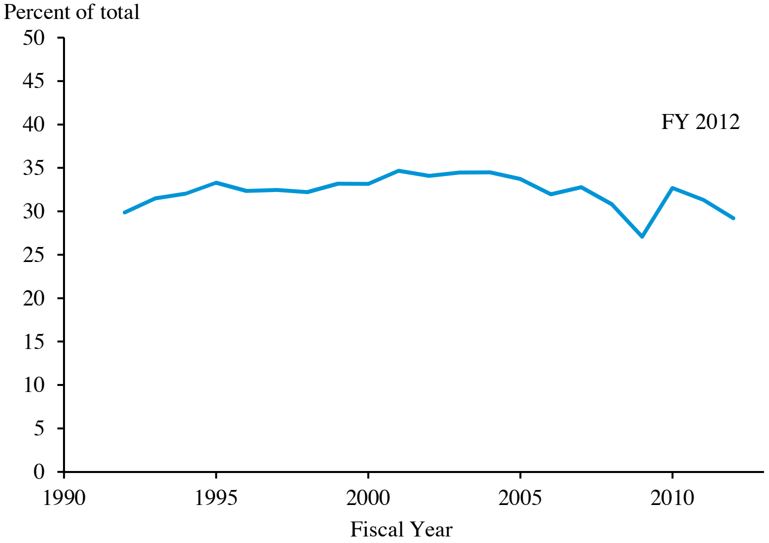
Institutionalizing Evidence-Based Decision-making in Grant Programs. In many programs, funds are distributed to states and local entities through competitive and formula grants. Grants to State and local governments have constituted roughly one-third of total outlays over the last 20 years, so increasing the use of evidence in informing policy in grant-supported programs could improve outcomes for a significant portion of outlays (Figure 7-1). Many effective program structures treat evaluation as an essential element in the decision-making framework, while also building in opportunities to scale up approaches that work and scale back or eliminate approaches that do not. As stated by then-OMB Director Peter Orszag, new initiatives should ideally have "evaluation standards built into their DNA" (OMB 2009b). The Administration has experimented with several models that embed both evidence building and evidence-based decisionmaking into the "DNA" of grant programs.

During this Administration, several initiatives have adopted a "tiered evidence" approach that embeds evidence-based decision making into program structure. Tiered evidence programs tie grant funding to the evidence base behind proposed interventions. In these programs, interventions that provide better evidence of success move to higher tiers and become eligible to receive more funding for expanded implementation and evaluation. The built-in mechanism for scaling up interventions that work helps prevent the troubling problem of not investing in programs with proven high returns.

A successful example of a three-tier approach is the Investing in Innovation program at the Department of Education. This program provides seed development grants of up to $3 million for high-potential and relatively untested interventions, validation grants of up to $12 million for interventions based on only a moderate amount of evidence, and scale-up grants of up to $20 million for potentially high-impact, transformative education interventions. Evidence of effectiveness is an "entry requirement"

Figure 7-1

Outlays for Grants to State and Local Governments, 1992–2012



Note: Total excludes outlays for defense, interest and social security.
Source: Office of Management and Budget, FY 2014 Budget, Tables 3.1 and 12.2.

for validation and scale-up grants, and all grantees are expected to conduct an evaluation that will add to the evidence base on effectiveness. For the scale-up and validation grants, the grantee must make the data from their evaluations available to third-party researchers, consistent with applicable privacy requirements (Department of Education 2013a).

Similarly, the Maternal, Infant, and Early Childhood Home Visiting Program in the Department of Health and Human Services (HHS) was an early Administration initiative that uses a two-tiered evidence structure. Implemented in 2010 as part of the Affordable Care Act, this voluntary home visiting program uses trained professionals and paraprofessionals to provide support to vulnerable pregnant women and parents of young children to improve health, development, and well-being outcomes for at-risk children and their families. The Act required that at least 75 percent of the home visiting program funds be spent on proven, evidence-based approaches and allowed for the remainder to be spent on promising approaches as long as they are rigorously evaluated. Currently, 14 home visiting models meet the HHS criteria of "evidence-based approaches," and have been evaluated with a mix of randomized experiments and quasi-experiments using multiple measures of key outcomes (Paulsell et al. 2013). While the Act funded the home visiting program through 2014, the Administration has proposed to continue funding and expand the availability of voluntary evidence-based home visiting programs to reach additional families in need as part of a continuum of early childhood interventions.

In addition to tiered evidence structures, agencies have begun using other designs in competitive grant programs that encourage the use of evidence-based practices. One such design is the “Pay for Success” approach. In this performance-based model, philanthropic and private funding is leveraged and the government provides payment only after targeted outcomes are achieved. In 2012 and 2013, the Administration started supporting programs that use a Pay for Success model to fund preventive services, and which had outcomes that could be measured with credible evaluation methodologies. The first Pay for Success awards were for projects to prevent prison recidivism.[5] The Consolidated Appropriations Act of 2014 authorized up to $21.5 million for Pay for Success projects.

Even in more traditionally structured grant programs where funding is provided upfront, agencies are embedding more rigorous evaluation requirements into funding requirements. For example, upfront grants in the Department of Labor’s Workforce Innovation Fund, first issued in 2011, fund promising but untested employment and training service and administrative strategies. These grants also fund well-tested ideas being adapted to new contexts as a way to significantly increase evidence about interventions that generate long-term improvements in public workforce system performance, such as reduced duration of unemployment. Grantees are required to conduct rigorous evaluations, and a national evaluation coordinator works with grantee evaluators to ensure consistent and high-quality evaluations (Department of Labor 2011).

Ending or Reducing Funding for Interventions or Programs. The Administration’s commitment to evidence-based evaluation means terminating or reducing funding for a program when a body of evidence consistently shows that the program is not achieving its stated goals, helping to reduce the use of taxpayer dollars on ineffective programs. The FY 2012 Budget took this approach with the Mentoring Children of Prisoners (MCP) program run by the Department of Health and Human Services. Rigorous evaluations show that high-quality mentoring relationships lasting for at least 12 months can have positive impacts on youth, while relationships that do not last more than three months can actually have harmful effects on youth (Grossman and Rhodes 2002). According to the MCP program

[5] For example, the Department of Labor allocated nearly $24 million in Workforce Innovation Fund grants to pilot Pay for Success grants to increase employment and reduce recidivism among formerly incarcerated individuals (United States Interagency Council on Homelessness, 2013a). DOL required the grantees to employ rigorous evaluation methods in gauging impacts on outcomes, which was defined in the grant solicitation as an experimental or credible quasi-experimental evaluation design.

performance data, fewer than half of program participants each year were in matches that lasted at least 12 months and, in 2008 alone, as many as 27 percent of matches that ended prematurely ended within three months. An evaluation of one MCP-funded program suggested that premature terminations were the result of program performance and were independent of the demographics of the participants (Schlafer et al. 2009).

Interpreting the MCP performance data in light of the evidence from impact evaluations of other mentoring programs, the Administration concluded that the MCP was not as effective as it should be. As a result, the Administration proposed to reduce funding for the MCP, noting that other competitive grant programs could serve the youth targeted by the MCP, and that some of those programs, such as Promise Neighborhoods, utilize evidence-based practices. Congress ultimately eliminated funding for the program in the Continuing Appropriations Act of 2011.

Even Start, originally designed to improve family literacy in disadvantaged populations, was another program not meeting its stated goals that the Administration took steps to replace. While the literacy levels of Even Start children and parents improved, multiple national randomized experiments showed that parents and children in control groups who did not participate in Even Start (one-third of whom received other early childhood education or adult education services) had comparable improvements (see for example St. Pierre et al. 2003). The President's FY 2012 Budget proposed, and Congress approved, the elimination of separate funding for Even Start. The Administration has proposed incorporating it and other narrowly focused literacy programs into the newly created literacy component of the Effective Teaching and Learning program that would support competitive grants to states for high-quality, evidence-based literacy programs.

Building Evidence when Existing Evidence is Limited

In many of the examples highlighted above, evidence existed on what programs or interventions were most effective, and the key challenge facing policymakers was to act on that evidence. However, not enough is known about what works in many other important areas, and so the first step in evidence-based policymaking is to invest in developing evidence.

Reducing Electricity Use. Experts have long suggested time-varying pricing (more costly at times of peak demand) as a way of increasing the efficiency of electricity use, including reducing electricity demand. Such time-varying pricing could increase efficiency, defer investments in expensive new power plants, and reduce pollution. However, most electricity delivery systems have not invested in the in-home technologies necessary to allow residential consumers to respond to time-varying prices. In addition, regulators have been hesitant to approve varying rates, and private companies have been reluctant to invest in modernizing their systems without knowing whether time-varying pricing will significantly impact consumer behavior. In recent years, the Federal Government, in partnership with states and utilities, invested in evaluating the impact of time-varying pricing on consumer behavior so that this information would be available to utilities, regulators, and states. These consumer behavior studies were implemented with American Recovery and Reinvestment Act funds and use randomized controlled experimental methods. Deciding which type of pricing strategy to use falls within State jurisdiction, rather than Federal, so these studies will allow State and local public utilities to make more informed decisions on pricing models (Cappers et al 2013). While these studies are still ongoing, two utilities and their regulators have decided to implement time-varying rates across their service territories based on the results observed to date. Such efforts can serve as an impetus to get more public utilities to adopt time-varying pricing.

Improving Health Care Delivery. In another example, the Affordable Care Act made a number of major investments in understanding how to improve quality and reduce cost in health care delivery, in addition to expanding access to affordable health insurance coverage. As described earlier in this chapter, the Center for Medicare and Medicaid Innovation (the "Innovation Center"), created by the Act, is using high-quality evaluation approaches to test innovative payment and service delivery models designed to reduce expenditures while preserving or enhancing quality of care for Medicare, Medicaid and Children's Health Insurance Program beneficiaries. Several ongoing Innovation Center payment reform initiatives—and early results from those initiatives—are discussed in Chapter 4. The Innovation Center will use the results of such model evaluations and actuarial data to identify best practices and determine which successful models could be implemented more broadly.

Better Outcomes for Youth with Disabilities. The Administration is also testing many different approaches aimed at youth with disabilities. The Promoting Readiness of Minors in Supplemental Security Income (PROMISE) is a joint initiative of the departments of Education, Health and Human Services, Labor, and the Social Security Administration. PROMISE aims to improve the education and employment outcomes for youth with disabilities who receive Supplemental Security Income (SSI) and their families, by improving coordination of services such as those available through the Individuals with Disabilities Education Act, the Vocational Rehabilitation State Grants program, Medicaid health and home and community based services, Job Corps, Temporary Assistance for Needy Families, and Workforce

Investment Act programs. The PROMISE program allows grantees (states or consortia of states) to design their own intervention models to serve youth and their families for three years with a two-year extension option, provided they include a minimum set of services. Grantees may also apply for waivers of funding restrictions or rules in individual programs that they believe will constrain their ability to achieve outcomes. Grantees agree to enroll a large number of youth (around 2,000) who are eligible to be served by a PROMISE intervention, and to allow random assignment to be used to assign half of eligible youth to the treatment group and the remaining youth to a control group that receives the services that child SSI recipients normally receive. The first grants were awarded in September 2013. To evaluate whether PROMISE can help child SSI recipients achieve better outcomes, a national evaluation will be conducted of all grantees to analyze intervention impacts on educational attainment, employment credentials and outcomes, and whether the interventions reduce long-term reliance on public benefits, and SSI payments in particular (Social Security Administration 2013).

Improving Outcomes For At-Risk Youth. The Administration also is working to identify approaches that help at-risk youth. The National Guard Youth Challenge (ChalleNGe) program, which has been rigorously evaluated, is designed to provide opportunities for adolescents who have dropped out of school but demonstrate a willingness to turn their lives around. Using random assignment, Millensky et al. (2011) found significant benefits to program participation in addition to higher earnings, as ChalleNGe graduates were more likely than the control group to have obtained a high school diploma or GED, to have earned college credits, and to be working three years after completing the program. Participation was projected to increase discounted lifetime earnings by over $40,000 (in 2010 dollars) (Perez-Arce et al. 2012). After considering education costs to the student and other non-earnings benefits, the ChalleNGe program was estimated to generate $2.66 for every dollar of program cost (Perez-Arce et al. 2012). The Administration now plans to test the application of the ChalleNGe model to adjudicated youth, through the Department of Labor's Reintegration of Ex-Offenders program.

Reducing Homelessness. Sharply reducing homelessness is a key focus of the Administration.[6] Although once considered an intractable problem, a broad body of research (including rigorous evaluations) documented that

[6] Spurred in part by the Homeless Emergency Assistance and Rapid Transition to Housing Act of 2009, the Obama Administration released *Opening Doors*: *The Federal Strategic Plan to End Homelessness* in 2010. The plan establishes ambitious goals to end veterans' and chronic homelessness as well as homelessness among youth and families. The U.S. Interagency Council on Homelessness serves to coordinate action by 19 member agencies (United States Interagency Council on Homelessness, 2013b).

there are models that effectively serve individuals experiencing chronic homelessness. The Department of Housing and Urban Development (HUD) has invested heavily in promoting these evidence-based approaches, and has re-oriented the Homelessness Assistance Grant Program away from such traditional approaches as transitional housing and toward more-effective permanent supportive housing (Figure 7-2). Because research on interventions that are effective for homeless families does not yet exist at the same level of rigor as for homeless individuals (Culhane et al. 2007), HUD has undertaken an experimental study of family homelessness called the Family Options Study. This study will compare several combinations of housing assistance and services in a multi-site experiment to determine which interventions work best to promote housing stability, family preservation, child well-being, adult well-being, and self-sufficiency. In addition to usual care, defined as remaining in emergency shelter and accessing whatever resources that would normally be available to families in shelter, three interventions are being studied: 1) subsidy only (a voucher primarily), 2) transitional housing, 3) rapid re-housing.[7]

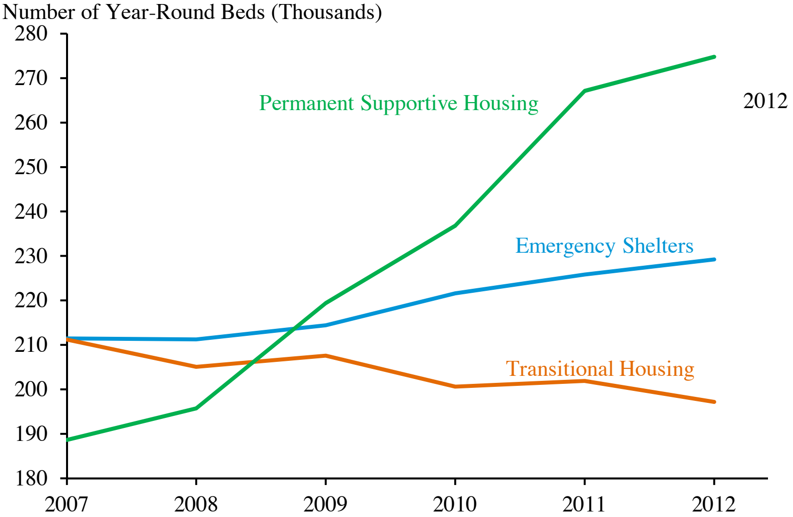
Furthering the Evidence Agenda

Relative to when President Obama first took office in January 2009, agencies are doing more to build actionable evidence to answer important program and policy questions. These efforts span a wide range of agencies and programs. While largely focused on improving the performance of programs that provide direct services to individuals and account for roughly 65 percent of total Federal outlays (OMB 2014), many agencies, including the Department of Commerce, the Small Business Administration, the Department of Agriculture, and the Treasury Department are also pursuing ways to incorporate impact evaluations in programs that provide assistance to businesses.

Instilling a culture of evidence-based decision making within agencies, and building the foundations that enable rigorous evaluations to guide new investments and drive policy, is neither quick nor easy. Evaluations of particular interventions or entire programs should not be isolated exercises that occur on an ad hoc basis, but rather planned in advance. Challenges always accompany efforts to enact significant change, but addressing several key elements can greatly facilitate agency efforts to improve the collection and use of evidence. While not a comprehensive list, these issues represent

[7] A summary of the study, as well as the Interim Report (which documents the study design process of randomization, and characteristics of the study population) can be found here: http://www.huduser.org/portal/family_options_study.html

Figure 7-2
Inventory of Beds for Homeless and Formerly Homeless People, 2007–2012



Source: Department of Housing and Urban Development, Homeless Data Exchange, Housing Inventory Count.

several major areas that provide either useful opportunities, or serve as barriers, in agency efforts to build and advance an evidence-based agenda.

Legislative Support for Evaluation

Authorizing legislation and appropriations bills can direct how agencies should use program funds for a wide variety of activities. Legislation can encourage stronger and more cost-effective evaluation in many ways. One is through language that recognizes the importance of conducting rigorous evaluations. Another is by making sure already-collected program data are made available for such statistical and analytical purposes.

Legislative Support for Rigorous Evaluations. Two ways that legislation can support rigorous evaluation is through set-asides and support for evaluation of demonstration programs. In recent years, with support of top management within agencies and the Administration, several agencies have had set-asides for evaluation specified in program legislation and appropriations. For example, the Consolidated Appropriations Act of 2012 first enabled the Secretary of Labor to reserve up to 0.5 percent of specific Department of Labor (DOL) appropriations for evaluations. Also, a set-aside of 5 percent of competitive grant funds in the Teacher Incentive Fund allows the Department of Education to conduct a rigorous national evaluation of the program and to share with grantees the results of current, rigorous research to help facilitate ongoing improvement. Additionally, authority to set-aside a percentage of program funds for evaluation is specified in some HHS and HUD programs, including several that received additional funding through the Affordable Care Act of 2010 and the American Recovery and Reinvestment Act of 2009.[8]

Many programs are funded through annual appropriations, which generally require obligation of those funds within a given fiscal year. However, the DOL set-aside for evaluation in the Consolidated Appropriations Act extends the deadline by which the DOL must obligate transferred evaluation funds to two years. Because designing rigorous evaluations takes time, a window beyond the standard one-year for obligating evaluation funds can in some cases enable agencies to plan and execute more thorough, higher-quality evaluations.

Legislation that specifies funding for demonstration pilots also provides important support for developing an evidence base. The legislatively authorized demonstrations being conducted by the Innovation Center are recent examples that illustrate the value of legislative support for evidence building.[9] As another example, the Department of Health and Human Services' child welfare waiver authority allows states to design and demonstrate a wide range of approaches for reforming child welfare and improving outcomes, including decreased first-time entries and re-entries into foster care, and improvements in various aspects of child developmental, behavioral, and social functioning. States are required to conduct rigorous impact evaluations as well as process evaluations as part of their waiver agreements.[10] In addition, the Administration is proposing to restore demonstration authority for the Disability Insurance program, while also providing new authority for the Social Security Administration and partner agencies to test early-intervention strategies that would help people with disabilities remain in the workforce.

Legislation can also encourage stronger evaluations through explicit language requiring grantees to participate in evaluations and by requiring use of proven interventions. The Healthy, Hunger-Free Kids Act, for

[8] While set-asides within programs are useful, some have noted that department-wide set-asides may have advantages over program-level set-asides by providing agencies with more flexibility over maximizing the return to evaluation investments. Also, set-asides will be used most effectively when agencies have a demonstrated capacity to manage evaluation funds.

[9] Prior to passage of the ACA, existing demonstration payment waiver authority allowed HHS to conduct Medicare demonstrations of the impacts of new service delivery methods and new payment approaches. However, due to statutory restrictions these demonstrations tended to be relatively small. The ACA provided the Secretary with more flexible authority for testing payment and delivery system innovations, and expanding them based on evidence. This work is conducted under the auspices of the CMMI.

[10] Child and Family Services Improvement and Innovation Act, Title II, Sec. 201, P.L. 112-34.

example, included a nondiscretionary provision that requires State and local grant recipients in a number of nutrition assistance programs, including the National School Lunch Program, the Special Supplemental Nutrition Program for Women, Infants and Children, and other programs authorized in the National School Lunch Act and the Child Nutrition Act to cooperate in evaluations conducted by or on behalf of the Department of Agriculture.[11] This Act also reformed the structure of the nutrition education provided through the Supplemental Nutrition Assistance Program, one of the Nation's main anchors of the social safety net that provides nutrition assistance to eligible low-income individuals and families. It established a new and improved Nutrition Education and Obesity Prevention Grant Program that requires a greater emphasis on evidence-based, outcome-driven interventions, with a focus on preventing obesity and coordinating with other programs for maximum impact and cost-effectiveness.

Legislative Support for Access to Data for Statistical Purposes, Including Evaluations. Existing laws can be explicit or implicit regarding whether information collected as part of administering programs can be used for statistical purposes integral to evaluation. Explicit and supportive laws can save significant time and effort in negotiating agreements to provide data for evaluations and can facilitate more and better analysis. For example, the Social Security Act explicitly states that one of the agency's datasets can be used for statistical and research activities conducted by Federal and State agencies.

Some legislation provides the agency head with broad authority to determine appropriate uses of program data. Given that the statistical uses of data in program evaluation often inform the context, policies, and operations of the same programs authorized by a given statute, agencies sometimes determine that their general statutory authority can grant sufficient authorization to provide administrative data to other Federal agencies for statistical purposes. For example, the Social Security Administration provides certain datasets for statistical and research purposes as described in its implementing regulations.

Multiple legitimate goals must be balanced when determining appropriate access to data, including reducing the burden of data collection on individuals and institutions and protecting personal privacy. Even so, careful crafting of legislative language can achieve those aims while still making data available for Federal researchers to rigorously evaluate and to improve

[11] U.S. Department of Agriculture, Food and Nutrition Service Final Rule, Cooperation in USDA Studies and Evaluations, and Full Use of Federal Funds in Nutrition Assistance Programs Nondiscretionary Provisions of the Healthy, Hunger-Free Kids Act of 2010, Public Law 111–296. Federal Register Vol 76, No. 125, June 29, 2011.

government programs. Key considerations include: avoiding vague or unclear authority to determine appropriate uses of program data; avoiding narrowly written statutory language that only allows access to data for narrowly defined programmatic reasons; or restricting a Federal agency's ability to collect data from grantees.

The information needs in programs managed at the State level could theoretically be addressed through non-Federal data systems, but this is not always possible. States or other grantees may not voluntarily develop comprehensive data systems in ways that are comparable across states, or have the capacity or incentive to make data available to researchers. When no feasible solutions exist to alleviate these issues, legislation may be warranted to authorize creation of Federal datasets accessible to researchers, or to establish requirements for State-held datasets that enable data exchange and comparability across states and to ensure access by researchers.

Building Evaluation into the Design of Programs

Many of the examples described earlier demonstrate the ways in which agencies are designing programs to facilitate evaluation. But agencies can still do more to embed rigorous evaluation designs into both new programs and existing programs.

New programs. The benefits of adopting evidence-based program designs, like the tiered evidence structure in the Investing in Innovation program and in HHS' home visiting program, include the ability to guide competitive grant funds to the strategies with a strong evidence base, while also requiring grantees to conduct evaluations where no evidence is yet available. Even without such a program structure, agencies implementing new programs over the past five years have increasingly required grantees to collect data and develop administrative data systems that can improve comparability and facilitate evaluation in addition to meeting program operating needs. For example, the Department of Education's Promise Neighborhoods initiative implemented in 2010 requires grantees to collect and track outcome data in an individual-level longitudinal data system to facilitate rigorous evaluation. This initiative aims to improve educational and developmental outcomes of children and youth in distressed communities.[12] To assist grantees in collecting high-quality and comparable data, the Education Department is providing grantees with extensive guidance on data collection and reporting (Comey et al. 2013).

[12] This program is based on the Harlem Children's Zone model, which was found to increase earnings for students, decrease the probability of committing crimes and decrease health disability probabilities—with the potential for providing large public benefits (Dobbie and Fryer 2011).

Other benefits of considering evaluation needs in the design of new programs include creating opportunities to save time and money by identifying evaluation data up front, minimizing burden on program respondents, and avoiding the loss of information all together that cannot be created too long after the fact. When not considered in the earliest stages of program design, a typical alternative to collecting the information needed for evaluation is to conduct surveys, which requires identifying expertise to design and test the surveys, gaining approval for their use, and then administering them to collect data, often long after the fact. Surveys add to the time and cost to build evidence, due to the time and skill involved in developing survey instruments that will yield high-quality data, the requirements for obtaining needed approvals, and the actual implementation of the survey.[13] Careful planning can help limit the need for evaluation-related surveys to data that cannot be obtained in any other way, such as information on post-program choices, earnings, or jobs necessary for identifying longer-term impacts of a program or intervention.

One of the most important ways the design of a program can facilitate evidence building is through careful consideration of how treatment and control groups can be established to facilitate impact evaluation. As discussed earlier, randomly assigning potential program participants to treatment and control groups enables the most credible impact evaluations. Several mechanisms exist for creating good comparison groups that allow for experimental or quasi-experimental techniques to be employed to produce high-quality estimates of program effectiveness.

Several options for enrolling potential participants in a program or intervention, presented in order of the rigor of evaluation they might support, are as follows:

1. Random assignment by lottery when capacity is limited. In many instances, due to limited funds or other constraints, a program or intervention cannot serve every person or entity that is eligible to apply. In such cases, rather than "first-come, first-serve" or other nonrandom devices, implementing a lottery to select which applicants may participate in a program or intervention generates a low-cost randomized experiment. This

[13] For example, the Paperwork Reduction Act (PRA), first enacted in 1980 and amended in 1995 (44 U.S.C., Chapter 35), requires Federal agencies to obtain OMB approval when an agency plans to collect information from ten or more persons using identical reporting, recordkeeping, or disclosure requirements. Among the PRA's goals are ensuring the greatest possible public benefit from and maximizing the utility of information created, collected, maintained, used, shared and disseminated by or for the Federal government and minimizing the burden for persons resulting from the collection of information by or for the Federal government. As a further example, some data collections are subject to review by Institutional Review Boards, in order to protect the rights of the human subjects of such research, a requirement under (42 USC 289) under 45 CFR 46.

strategy has been used recently to determine the impact of Medicaid access (Baicker et al. 2013), charter school attendance (Abdulkadiroglu et al. 2011), and small business entrepreneurship training (Benus et al. 2009). Note that the losers of the lottery need to be followed to track their outcome data.

2. Assignment based on a continuous "need score." A common objection to random assignment is that resources should be targeted to those with the greatest need, or those most likely to benefit. In this situation, program assignment might lend itself to a strong evaluation if it incorporates some sort of explicit, continuous, ranking of applicant need (or likely benefit), and bases program eligibility on some cutoff in need. For example, Ludwig and Miller (2007) study the effect of participating in Head Start on mortality rates for children by exploiting the fact that the Office of Economic Opportunity provided technical assistance to the 300 poorest counties in 1965. This created lasting differences in Head Start funding rates for counties with poverty levels just below and just above the poverty rate of the 300th poorest county. With this type of assignment rule, a regression discontinuity design can be used to study the impact of the program. The logic of the design is that individuals with "scores" just above and just below the threshold—in Ludwig and Miller, living in a county with a poverty rate just above or below the poverty rate of the 300th poorest county's rate—are likely to be similar to each other in ways that affect their outcomes, except that those just below receive the treatment (in this case, participation in Head Start). This design can deliver estimates of the effect of the program that are similar to randomized experiments.[14]

3. Staging the rollout of a large program. If a program will be introduced that will ultimately serve many participants spread across different geographic areas, or schools, or other natural groupings, staggering the rollout across time and space, with the rollout sequence chosen randomly, makes it much easier to evaluate. For example, suppose a mentoring program aimed at increasing college attendance will be introduced in a group of schools and the government hopes to learn about the effect of the program by estimating the change in college enrollment among students at the school before and after the program is introduced. If the program is introduced in only one school district, then any other changes that the school district introduced around the same time might affect the change in outcomes and bias the conclusion. Similarly, if the program is introduced in many different schools but all in the same year, then any other changes in policy, the economic climate, or other macro-economic conditions may be confounded

[14] On the other hand, the estimates from an RDD strictly pertain only to the types of participants "near" the cutoff. To the extent the impact varies across participants with different levels of need, this can be a limitation.

with the treatment effect and thus may "bias" the estimate of the treatment effect. Staggering the rollout of the program over time and space, using randomization and possibly further matching treatment and control units based on observable characteristics, helps to control for these potential biases, and thus allows for better estimates of a program's impact. This strategy has been used by Rothstein (2010) to study the effect of extended unemployment benefits.

The three strategies above create experiments or quasi-experiments that lend themselves to high-quality impact evaluations. In the absence of such devices, evaluators need to acknowledge the differences that do exist between program (or intervention) participants and non-participants and to use statistical techniques like multivariate regression and matching to control for these differences. Since these strategies all attempt to compare the outcomes of program participants and non-participants with similar characteristics, the success of the evaluation will be determined by the availability of good information on the characteristics of the population that are most predictive of the outcome under study, as well as the reasons why individuals choose to participate in a program. However, for these strategies to work, the variation between the treatment and control groups after using statistical or matching techniques to control for differences between these groups must be plausibly unrelated to the outcomes of interest, except through the effect of the treatment. There needs to be some part of that variation between the treatment and control groups that operates like randomization.

Existing Programs. Designing programs to facilitate evaluation may be relatively simpler in new programs than in existing programs, due to program manager reluctance in the latter to trying new strategies, concerns about equity among participants if the control group receives no services, and other reasons. But experiences at several agencies demonstrate these barriers can be overcome. Lotteries for oversubscribed programs are as applicable in longstanding programs as in new ones (see for example the Jacob and Ludwig (2012) study on impacts of housing vouchers). However, increasing opportunities for evidence-based decisionmaking in programs that allocate funds to states on the basis of formulas remain especially challenging, because evaluations and evidence-based funding allocations are not a requirement of States receiving the funds. Waiver authorities or other mechanisms to incentivize evaluations in these programs are only available in a few instances. A control that could prompt State and local grant recipients to do evaluations in these types of programs is a legislated requirement that a certain portion of funds be set aside for evidence-based grants or models of delivery. For example, in the Senate Appropriations Bill for FY 2014, the Substance Abuse and Mental Health Services Administration mental health block grant programs included language defining a 5 percent set-aside for evidence-based grants.[15]

There is still work to be done to embed evaluation and evidence-based decision making into more programs. Agencies can focus evaluation efforts in those programs that can help ensure that the agency's most critical program and policy questions are addressed.[16]

Developing the Capacity to Link to Other Administrative and Survey Data Sources

Increasingly, agencies are seeking opportunities to improve their evaluation approaches by supplementing their administrative program data with other available government data, where appropriate and while ensuring strong privacy protections. Using pre-existing data collected for other reasons, while maintaining strong privacy protections, provides a number of benefits. Several challenges arise when doing so, and the Administration is taking steps to address these challenges.

Benefits of using existing data resources. Using pre-existing administrative data collected for other reasons, while maintaining strong privacy protections, can help agencies answer important policy questions that could not otherwise easily be addressed with a single program database or survey. Administrative data provides the most complete and accurate source of information on program participation and can provide more accurate data on earnings, test scores, and other outcomes of interest. Indeed, the benefits of using pre-existing administrative data for evaluation and other statistical purposes have been widely acknowledged for some time. Data from multiple sources have been used in a number of impact evaluations, primarily to identify the characteristics of treatment and control groups, identify outcome variables which indicate the impacts of treatments, reduce study costs and reduce the burden on study participants by avoiding the need to collect the data via another survey (Coalition for Evidence-Based Policy 2012; Finkelstein et al. 2012; Bettinger et al. 2009; Jacob and Ludwig 2012). Linked datasets are also facilitating current evidence-building efforts in various agencies, such as in the Department of Health and Human Service's Office of Child Support Enforcement, which is currently implementing a child support-led employment services demonstration project with a random assignment impact evaluation (where treatment consists of extra services under the program, and the control group receives regular services that are available) and a cost-benefit analysis. The planned evaluation will draw on

[15] S. 1284, Report No. 113–71.

[16] Recognizing that agencies operate with scarce resources, OMB has encouraged agencies to adopt such a focus (OMB 2013b).

unemployment insurance wage and benefit records, as well as State administrative data on benefits in the Supplemental Nutrition Assistance Program and other public assistance programs, criminal justice system data and other data to more cost-effectively and accurately determine the effectiveness and the true costs and benefits of the program. As another example, HUD and HHS are pilot-testing links between HUD administrative data and HHS Medicare/Medicaid data, to build evidence on opportunities to improve the health of Medicare/Medicaid beneficiaries in HUD-assisted housing as well as the impact of housing assistance on health.

Challenges and Solutions. Nevertheless, accessing administrative data for these statistical uses is challenging. These data are collected to facilitate day-to-day program operations, including developing performance measures. Unless evaluation needs are considered in the database design stage, however, the meaningfulness of administrative data for conducting rigorous evaluation may be limited. Also, data definitions can vary dramatically across datasets; especially with State-level data, the definitions often vary across states and even counties. Aside from definitional differences, the quality of programmatic data—its completeness and accuracy—can vary dramatically across datasets. Significant data-quality gaps or errors can compromise analysis. It can also be costly to negotiate access to data on a state-by-state basis.

One key practical challenge is that agencies, in an attempt to be privacy sensitive, may not include in program databases unique identifiers for program applicants and participants. Such unique identifiers facilitate linking to data provided by subjects through other programs or even for the same program over time. Linking datasets through name and address matching or matching on other less unique variables can introduce bias and render the linked data unusable for rigorous analysis. While some agencies have an established history of allowing use of data (including identifying information) for statistical purposes, in many cases access to such data is not readily available due to real or perceived legal, policy, or operational barriers.[17] In some cases, extensive negotiations with the agency responsible for the data are needed to gain access to the data for use in evaluation studies; sometimes the efforts are not successful even after months or years of negotiations.

[17] One legal barrier is that when a program's authorizing statute is silent about whether access to data can be provided for statistical purposes (which includes evaluation), agencies need to make a determination about allowable uses. In such cases, agencies may conservatively interpret the lack of an express authority as a prohibition on providing data to another agency. However, as discussed in OMB memorandum M-14-06, agencies may be able to provide the data under their general statutory authority (OMB 2014).

To help address these barriers, the OMB recently issued guidance to assist both program and statistical agencies (and statistical components within agencies) in increasing the opportunities to use administrative data for statistical purposes, which includes evaluation.[18] In part, this guidance requires government departments to engage both program and statistical agencies in identifying administrative datasets of potential value for statistical purposes; communicating the importance to staff of promoting the use of administrative data for statistical purposes; and identifying several datasets with the most value for statistical purposes but which are not currently being provided, along with descriptions of critical barriers that appear to preclude providing access for statistical purposes. The guidance also offers tools to help agencies in these tasks, including guidance in understanding relevant legal requirements, a tool to facilitate more efficient interagency agreements, and a tool to assess administrative data quality developed under the auspices of the Federal Committee on Statistical Methodology. Departments must also report to the OMB on their efforts to foster collaboration and increase access to administrative data for statistical purposes.

Facilitating Researcher Access to Federal Data while Protecting Privacy

Some agencies have developed ways for researchers to access Federal data for statistical purposes in secure research environments that preserve the confidentiality of individual records. The Census Bureau and National Center for Health Statistics operate secure research data centers, in which qualified researchers with approved projects can use micro-data files for statistical research. The Retirement Research Consortium is a key tool that the Social Security Administration uses to facilitate policy-relevant research on retirement and Social Security. The consortium comprises three competitively selected research centers based at the University of Michigan, Boston College, and the National Bureau of Economic Research. The centers perform valuable research and evaluation of retirement policy, disseminate results, provide training and education awards, and facilitate the use of SSA's administrative data by outside researchers. Nonetheless, due to confidentiality restrictions, uneven interpretations of laws governing privacy of data provided to the government, and other reasons, many data sets remain

[18] Statistical purposes is defined in footnote 2 of the OMB memorandum M-14-06 (OMB 2014): [it] refers to "the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups," (PL-107-347, Title V—Confidential Information Protection and Statistical Efficiency Act (CIPSEA), Section 502 (9)(A)). Statistical purposes exclude "any administrative, regulatory, law enforcement, adjudicatory, or other purpose that affects the rights, privileges, or benefits of a particular identifiable respondent" (PL-107-347, Title V—CIPSEA, Section 502 (5)(A)).

hard for researchers to access for statistical uses, and opportunities to link to researcher-collected data remain limited.

This Administration is committed to improving opportunities for researcher access in ways that fully maintain privacy protections of Federal program participants. HHS' Centers for Medicare and Medicaid Services' (CMS) Virtual Research Data Center is an innovative example of ways agencies are working to improve access to Federal agencies for their own use and for their grantees carrying out federally sponsored research activities. In late 2013, the Virtual Research Data Center began providing users with a dedicated workspace where they can upload external files and use them with CMS data to run analyses and download aggregate statistical files to their workstations. This model is a more-efficient, less-expensive, more-flexible and more-secure way for researchers to access a variety of Medicare and Medicaid program data, relative to the existing approach that entails cutting, encrypting, and shipping large quantities of information.

Conclusion

Whatever the findings, rigorous evaluations provide critical and credible feedback about whether the current design of a program is effective or whether program modifications are needed so that important program goals are met. Indeed, in some fields—including business and medicine—the vast majority of randomized controlled trials used to evaluate the efficacy of interventions and strategies find no positive effects of interventions (Coalition for Evidence-Based Policy 2013). Rigorous impact evaluations serve as important learning tools to guide management decisions about program investments. The Administration continues to support the use of these tools, broadly and often, to facilitate continuous improvement in government programs as well as to identify best practices and effective new approaches that can be shared with organizations delivering services funded with Federal dollars.

Over the last five years, Federal agencies have increasingly used rigorous impact evaluations to inform program decisions, including how to improve programs. Agencies are trying new approaches when the evidence indicates existing strategies are not yielding sufficiently positive impacts on important outcomes. They are restructuring programs to increase their effectiveness when evidence shows new strategies produce better results, and are developing evidence where an insufficient evidence base exists. And they are scaling up approaches that work, improving public policy and people's lives. As part of this effort, agencies are improving the collection and comparability of data to provide new opportunities for evaluation. They are also using cutting-edge technology to improve data access to other Federal agencies and to outside researchers while protecting privacy—strategies that can enable evaluations to be done more rapidly and at lower cost. The Administration continues to support these efforts to affect change. By using rigorous evaluation strategies to identify what works, and by taking steps to make needed modifications, agencies and taxpayers will have the greatest confidence that scarce resources are being used as efficiently as possible in meeting priority goals.