RUNNING GOVERNMENT LIKE A BUSINESS SCIENTIST

The Rise of Evidence-Based Decision Making

BY SHAYNE C. KAVANAGH AND NATE LEVENSON
Limited funding and increasing demands have led to reformed public budgeting practices, requiring governments to “do more with less.” One way to accomplish this is doing more of what works and less of what doesn’t, using evidence to tell the difference between the two. Some local governments are doing just that by bringing the power of scientific inquiry to public management. This method is known as evidence-based decision making. For early adopters, it has yielded some surprising findings and contributed to greatly improved decisions on how resources are allocated.

To illustrate, in one school district the budget battle started the usual way. The math department wanted more money for its program, and the English department wanted more for its program. Both believed strongly in the merits of their programs and cared a great deal about making sure kids were learning. To justify their requests, both departments presented anecdotes about students learning a great deal and teachers being certain the program was making a difference. In the past, charisma, persuasion, or clout would have carried the day, but the district had recently embraced evidence-based decision-making. The school therefore knew that students in the English program were making few gains, despite spending $4,000 a student. The math program, however, had produced 18 months of growth in one year, at the cost of just $1,000 a student. Presented with this evidence, even the English director no longer lobbied to continue funding, and as a result, the successful math program was expanded, and the next year, students who were struggling in English got a new, better, intervention.

**THE EXISTING REFORM BRIEF: RUN GOVERNMENT LIKE A BUSINESS**

Over past decades, public managers have sought to reform planning and budgeting by reducing the traditional focus on expenditure line items and placing more emphasis on the outputs that governments produce and the outcomes that make a difference in the lives of citizens. The increased popularity of performance measures and performance budgeting has been the most visible manifestation of this approach.

The essence of performance budgeting is captured by Exhibit 1, which illustrates a county government that has made reduced juvenile crime a priority. One of the programs used to achieve this outcome is Scared Straight, where at-risk juveniles tour a prison and hear cautionary tales from inmates. The unpleasant conditions of the prison and the unfortunate circumstances of the prisoners are thought to inspire the young people to steer clear of a life of crime. The county applies performance measures to the program, such as the number of participants and/or the satisfaction of its clients (e.g., the social workers who refer juveniles to the program).

As performance budgeting is often practiced, the ongoing funding of the program is usually secure so long as a logical conceptual link exists between the program and the desired outcome, and the measurements receive satisfactory scores. Some governments might also measure the outcome, such as juvenile crime rate, but they would not generally directly link the program and the outcome measure because of the other variables that could affect the outcome.

The existing budget reform paradigm is inspired by the practices of private corporations: Businesses use quantified “bottom line” indicators to guide initiatives in the name of creating value for its stockholders. In the place of stock prices, governments have outcomes. In place of indicators like profitability, sales volume, and market share, governments have performance measures.

**A NEW WAY? RUN GOVERNMENT LIKE A SCIENTIST**

While the existing approach to budget reform is an improvement over traditional budgeting, there is a still a nagging,
unresolved concern: cost-effectiveness. Does Scared Straight, or any other program, actually make a significant impact on the desired outcome? And does it do so for an acceptable cost?

To answer this question, some governments are beginning to employ evidence-based decision making, which is inspired by the scientific method. The essence of the scientific method is: 1) form a hypothesis; 2) do an experiment to test the hypothesis; 3) analyze the data; and 4) draw a conclusion.

Advances in technology and the greater availability of large data sets have made it feasible for governments to collect and analyze the data necessary to prove or disprove their hypotheses. Both simple data (e.g., who attended a particular program or event) and more complex data (e.g., grades, graduation rates, or criminal records) are now stored electronically. Low-cost technology makes it possible to extract meaning from millions of data points, even if they are stored in separate databases. Perhaps just as importantly, technology also makes it possible to access studies of program effectiveness performed by third parties. Just a few years ago, accessing these sources of evidence required hours in a university library. Now they can be obtained from online databases. Rather than conducting its own experiment on Scared Straight, our county could look at the results of studies done by other organizations.

The high-tech world has embraced scientific approaches to doing business. Given the cultural cache of the high-tech sector, these scientific approaches to decision making have spread to other sectors of the economy, including governments. For example, Google regularly conducts controlled experiments by releasing changes to a random sample of its users and then analyzing their responses against the responses of users still running the standard product.

A final force in favor of a more scientific approach to decisions is that, in recent years, researchers have uncovered more information about the cognitive biases that are woven into the human psyche. A cognitive bias is a persistent flaw in logical reasoning that affects most people, such as the widespread “overconfidence bias” — our tendency to assess situations to be more favorable to us than they actually are. In other words, we are more optimistic than realistic. To illustrate, according to Tali Sharot, a neuroscientist who specializes in this subject, people hugely underestimate their chances of getting divorced and losing their jobs; they believe they will outperform their peers at work; and they overestimate their likely life spans (sometimes by 20 years or more).

Cognitive biases can actually confer important benefits. For example, overconfidence helps reduce your stress about undertaking a life-changing event. You will feel better about changing jobs, moving, or getting married if you are overconfident about how well these changes will work out. However, when it comes to clear-eyed evaluation of the cost-effectiveness of programs, cognitive biases aren’t as beneficial. For example, overconfidence bias might lead us to overestimate what we know about how a program works and its efficacy.

Beyond overconfidence, perhaps the most dangerous flaw in our analytical abilities is our well-developed ability to concoct a seemingly reasonable connection between unrelated objects or ideas, which psychologists refer to as “apophenia.” To illustrate, take a moment to think about the Scared Straight program and imagine you are being asked to explain to the county board why the program is effective in reducing juvenile crime. Could you provide a seemingly reasonable explanation for the effectiveness of this program, even if you are not a juvenile justice expert? You probably could. The problem is that the seemingly reasonable explanation is often taken at face value — we formulate the hypothesis and draw the conclusion directly from there, without collecting proof.

WHAT DOES THE EVIDENCE SAY?

So, what does the evidence on Scared Straight tell us? Not only do Scared Straight programs not work, they are actually counterproductive. One study examined nine Scared Straight programs and found that these programs were actually corre-
lated with a 28 percent increase in crime.\textsuperscript{4} Other studies have produced similarly unflattering findings. The U.S. Department of Justice has adopted an official position against Scared Straight programs.\textsuperscript{5}

Although Scared Straight is an extreme case, it is not the only time scientific evidence produced surprising findings. For instance, the popular Drug Abuse Resistance Education program (commonly known as D.A.R.E.) has been proven widely ineffective. Participants were not less likely to become a drug addict or even to refuse that first beer from their friends over the 25 years it was in existence.\textsuperscript{6} In fact, D.A.R.E. America, the organization behind D.A.R.E., has since replaced traditional program with a new one, Keepin’ It REAL, that is built around evidence-based methods for improving communications and decision-making skills in children.

In public education, special education co-teaching is a common way to help those students who need it most. The hypothesis is reasonable: A student who struggles academically and has special needs requires both a teacher who knows the content, such as math, and a teacher who knows special ways of teaching students who learn differently. Hence, schools should have two teachers teach together, bringing together the required expertise. However, the data are disappointing. Nationally, special education students in co-taught classrooms perform slightly worse than those taught by a single teacher,\textsuperscript{7} perhaps because the teacher who focuses on content spends very little time with the struggling student. There is also a financial implication: Co-teaching doubles the cost of instruction.

Though the evidence will not always produce findings as striking as those above, these examples demonstrate that “common sense” rationales for programs are not sufficient for ensuring that the public is getting value for its tax dollars. Public managers need to run local government more like a scientist by incorporating evidence-based decision making into planning and budgeting.

**HOW TO GET STARTED**

To get started with evidence-based decision making, ask for evidence that programs actually work. One of the most common justifications for any program in local government is “we’ve always done it that way.” Evidence-based decision making means going beyond this explanation to find out if a program actually produces the results the community wants for an acceptable cost.

The Pew-MacArthur Results First Initiative provides a structured way to approach this question. One of the big advantages of the Results First method is that it does not require a government to have conducted rigorous evaluations of its own programs; instead, Results First relies primarily on pre-existing studies performed by third parties. The essence of Results First is a four-step approach, which is described below and illustrated with the experiences of four California county governments: Fresno, Kern, Santa Barbara, and Santa Cruz.\textsuperscript{8}

**Step 1: Create an Inventory of Existing Programs.** Results First often recommends that governments start by conducting an inventory of programs in just one area of service and using that as a pilot of the Results First method.

All four California counties began their work in 2013 and focused their initial efforts on adult criminal justice. They were inspired to use Results First by a California law that transferred responsibility for more than 60,000 offenders to
the 58 county governments in the state. The four counties need to handle this influx in the most cost-effective way possible.

**Step 2. Review the Evidence.** Results First has compiled a database of third-party studies that governments can use in comparing their own programs to what researchers have found about the effectiveness of that type of program. Results First suggests that governments then categorize their programs according to their theoretical effectiveness. For example, Scared Straight would be categorized as “negative effects,” while at-risk youth mentoring would be categorized as “highest rated.” The database is open to the public, and you can find it by doing an Internet search for “results first clearinghouse database.”

A review of the evidence does not necessarily lead to a definitive conclusion, however. Santa Cruz found that its correctional education programs (GED programs) lacked important components necessary to achieve the best outcomes — they were not provided frequently enough to make the impact that the county wanted.

**Step 3. Conduct a Cost-Benefit Analysis.** The cost of the programs is then compared against evidence of their benefit. For the California counties, an important part of finding the cost of their programs was to calculate recidivism rates and the cost of returning offenders to the criminal justice system. As with many public services, prevention is a far more cost-effective strategy than remediation. Hence, to be useful, a cost-benefit analysis must take into account the preventative potential of programs. Santa Barbara found that 63 percent of high-risk offenders had been reconvicted within seven years. Given the cost of a reconviction ($60,000 per offender), the need for more effective strategies to prevent recidivism became obvious.

Santa Barbara also found that although one of its substance abuse programs produced positive outcomes, the costs required to produce those outcomes were unacceptable.

**Step 4. Use Evidence to Inform Decision Making.** Funds might be shifted from programs that don’t have evidence to support their cost-effectiveness to programs that do. Kern County replaced unstructured art activities, intended primarily to keep inmates occupied, with “Aggression Replacement Training,” a program proven to reduce recidivism. Kern also reduced the number of uniformed jail positions in favor of specialists with training in mental health to administer evidence-based programs.

Evidence also changed how Kern approached contracting. Like many counties, Kern works with community-based service providers to accomplish its mission. Going forward, use of evidence-based programming will be a requirement for these partners to receive funding from the county.

**THE PROS AND CONS OF PRE-EXISTING THIRD-PARTY STUDIES**

Third-party studies offer a number of advantages. First, they limit the need for a government to perform its own studies, reducing the cost of evidence-based decision making. Second, third-party studies are a good way to identify programs that are ineffective and where resources might be better used elsewhere. Studies can also show the best opportunities for where to shift those funds. Third, third-party studies that are included in a database will have a strong methodological foundation. This means they will provide decision makers...
with a good introduction to what solid evidence looks like. A final advantage is that seeing the evidence on what works expands your understanding of what can be achieved through government intervention. When you only know what your own programs have achieved, your understanding of what is possible may be narrowed. When you can see the evidence of what others have been able to achieve, it may inspire you to think differently.

Relying on third-party studies also has its limitations. First, third-party research is not comprehensive across areas of public service. The Results First database, for example, currently addresses the following services: child welfare and juvenile justice; criminal justice; substance abuse and mental health; and education. This omits some services that local governments provide because some services are under-studied or the studies are not well-documented. Furthermore, even if services have been studied throughly, third-party studies might not have been done for some of the specific programs a government offers within that service.

The second limitation of third-party studies is that they are not necessarily the final word on program effectiveness; that’s because of the challenges of transferability and faithful implementation. “Transferability” means that the program may have been studied under conditions that differ significantly from conditions in other governments, calling into question the applicability of the study. An example is the "small schools" approach to improving outcomes for high school students living in poverty. Studies showed that small schools that emphasized knowing and building relationships with their students, tailoring instruction to their interests, and setting high expectations had far better results than traditional large, comprehensive high schools. This inspired many districts to create their own small schools. Unfortunately, the model didn’t transfer well. The schools in the study were created by charismatic principals who handpicked teachers who were committed to small schools. In other districts, sitting principals were ordered to implement the model and the existing teaching staff was assigned to the small schools regardless of whether they believed in or even understood the new model. “Faithful implementation” refers to the likelihood of an evidence-based program failing if the implementation is not performed correctly. For example, one study of health promotion programs for children and adolescents found that programs that were implemented correctly achieved two to three times the effects of programs with a flawed implementation.⁹

Given the limitations described above, public managers can’t rely solely on third-party studies to conduct evidence-based decision-making. They will need to build some of their own capacity.

DIY EVIDENCE-BASED DECISION MAKING

The first step in developing your own capacity for evidence-based decision making is understanding where you might obtain evidence.

The gold standard of evidence is the randomized controlled trial (known as an RCT), which is the same process used to test new medications before they are made available to the public. The two key features of an RCT are expressed within its name. First, participants are randomly selected, which helps to avoid selection bias. Selection bias occurs when volunteers or hand-picked participants are different from the general population in some important way and are therefore not a good representation of how the intervention (e.g., medication, a government program) will perform in the general population. The second feature of an RCT is that there are two groups: a control group that receives no changes to its conditions and a group that receives the intervention.

Morality of Experiments in Government

Some might find experiments objectionable because only those in the intervention group will receive a potentially valuable program. However, despite the best intentions of public managers, not all new interventions work out as their inventors hope. Morality therefore suggests testing out the intervention first, thereby limiting the risk of a failed intervention.
The experimenters can then see if the intervention group is any better off than the control group.

If the RCT is the gold standard, then the quasi-experimental design takes the silver. When RCT isn’t possible, the power of statistics and thoughtful program evaluation design can be an excellent runner up. For example, you can search for experiments hiding in existing data. To illustrate, many schools provide a reading program for struggling students (i.e., those scoring below a given threshold, say 300, on a standardized test). Students who score 299 would get extra help, but students who score 301 wouldn’t. Since the margin of error on the test is far greater than a few points, students scoring 299 and 301 are actually very similar. An analysis comparing reading growth of students who were just above and with those students who were just below the cut-off score can provide great insight into the impact of the reading intervention. Quasi-experimental designs can also be intentionally created. (Later, we’ll show you an example from a school district in Michigan.)

After the gold and silver standards comes lead — relying on anecdotes and or personal observation. Eyewitness accounts are not valid sources of evidence because they are highly vulnerable to cognitive biases.

Now that we’ve reviewed where the evidence base comes from, let’s examine how to bring this evidence into decision making. The first step is deciding where to apply it. Applying the approach to every line of the budget, every program funded, or every question raised is too much work. Realistically, you can only analyze a handful of programs each year. For any given program, if you can answer yes to many of the questions below, then it might be a good candidate for an evidence-based review. If you have already implemented performance budgeting you have a good foundation for answering these questions:

- Does the program consume a lot of staff time or money?
- Are the necessary data readily available?
- Are there plans to substantially expand the program?

An often overlooked step in creating your own evidence-based decision capacity is having a set definition of what constitutes an “effective” program. Just like in science, evidence-based decision making works best if you have a clear standard for success going into the analysis. For example, when evaluating a reading program for struggling students, one district learned that students in the program made eight months of progress in a year. No one debated the figure, but raging disagreements ensued just the same. Some felt that making less than one year’s growth in one year was proof of failure — the students ended further behind their classmates than when they started. Others felt it showed progress because some gain is better than no gain at all. Ultimately, they decided that interventions are only successful if the students made much more than a year’s gain in a year’s time, but a lot of time was wasted and hard feelings were created in the process.

Knowing what works is important but not sufficient in an era of tight budgets. Imagine the joy of learning that a particular reading intervention is very effective. Say students who receive this service make 30 percent more gains than students who didn’t. It’s too soon to declare success. This program cost more than $6,000 per student, which means that very few students can be helped, given limited resources. A third-party study showed that an alternative program produced gains 80 percent as great but at a cost of $2,000 per student. This example makes it clear that data on cost-effectiveness frames an entirely different decision than if cost where not part of the conversation.

Finally, when designing your DIY evidence-based decision-making system, be forward looking. Evidence-based decision making might inspire a desire to look back at past performance. However, historical data are often not of the quality needed to make a valid evidence-based decision; there might not be random assignment or control groups, for example.
Instead, design the systems to collect the right data over time and make better decisions in the future.

**A TALE OF TWO SCHOOL DISTRICTS**

We’ll conclude this article with two very different stories of successful evidence-based decision making and then draw the lessons from each.

**Traverse City Area Public Schools Runs an Experiment.** Traverse City Area Public Schools (TCAPS), Michigan, serves 10,000 students. TCAPS had a problem: Its elementary school math scores had been consistently below state averages. There were no shortage of ideas for interventions, but TCAPS needed to choose the most cost-effective one.

The first step was to investigate the problem in order to develop a solid hypothesis. TCAPS noted that it was above the state average for elementary school reading scores. The same teachers and students who were producing the below-average math scores were producing these above-average reading scores, so TCAPS concluded that the problem was specific to how math was being taught and learned. TCAPS also noted widespread dissatisfaction with the current math curriculum. These were two good clues, so TCAPS examined third-party studies on curriculum effectiveness and found that the existing math curriculum compared unfavorably to many of the available alternatives.

Hence, it seemed reasonable to hypothesize that procuring a new math curriculum could address TCAPS’ math problem. But, in the words of the TCAPS school board vice president, implementing a new curriculum in the traditional way was like “buying a house based on the blueprint” — full of promise, but also rife with uncertainty as to how it will ultimately turn out.

The solution was to run an experiment to see the effect of a new curriculum before committing to full purchase and implementation. TCAPS decided to include three new curricula in the experiment, each supported by third-party studies of effectiveness. Besides providing TCAPS leadership with more information on which curriculum would be the best fit for TCAPS, testing three curricula sent a signal that this was a real experiment and not just a formality on the way to a predetermined conclusion. The experiment was scheduled over a full year to make sure it ran long enough to produce valid results.

Several schools volunteered to test one of the new curricula. TCAPS recognized that the non-random participation was a weakness of its experimental design, but decided that the enthusiasm it generated for evidence-based decision making was worth the risk. The remaining schools continued with the old curriculum, serving as a control group. TCAPS measured the results from multiple perspectives, including growth in test scores, teacher and parent satisfaction, and cost-effectiveness.

After a year, TCAPS found that two of the three curricula they tested produced a statistically significant improvement in test scores over the old curricula — that is, the differences in performance were large enough that they were highly unlikely to be due to chance. They also found that teachers greatly preferred any of the new curricula to the old one.
An unexpected result of the experiment was the great enthusiasm for the process. Staff felt that TCAPS was making the best possible decision for the future of the students, with the associate superintendent calling it the “best experience of his career.” The former chief financial officer and now-current superintendent said: “The process brought a great deal of credibility to this decision due to its transparency and use of objective data. Perhaps most importantly, it established a framework for future decision-making.” The school board vice-president likened using an experiment to select a curriculum to doing a walk-through of a house before buying it — both processes lend far more certainty to the final decision.

TCAPS plans to extend evidence-based decision making to other areas and is developing a strategic financial plan to guide where evidence-based decision making should be applied.

**Going with Your Gut Could Give You Indigestion.** In Arlington (Massachusetts) Public Schools, the English department presented the superintendent with a program to help middle-school students who were struggling with writing in their English classes. The program would feature small group learning to provide focused help for students who were a year or two behind their peers in their writing skills, but who otherwise seemed to have the potential to catch up. The superintendent loved the plan, which appeared to fill a pressing need to improve students’ writing skills in alignment with the superintendent’s own theory on how the school district could best help children: highly skilled teachers providing intensive, targeted support for struggling students. Based on this, the district made a substantial commitment to the program, which received a dedicated room, complete with new computers, new carpeting, and a new paint job, at a cost of $40,000 as well as the equivalent of four full-time teachers dedicated to running the program. The superintendent also made a point of stopping by this program regularly to check in. He liked what he saw — everything appeared to going well, with students engaged in orderly and concentrated study, and teachers working with the students. The teachers liked the program, and the English department also supported it.

At the same time, a much smaller investment was made to help students who were struggling in math. The program was a hybrid of multiple teaching and learning styles. At any one time, one-third of the class would participate in group lecture with the teacher, a third would work independently, and a third would work on computer-aided lessons. The program was only offered as a concession to the math department, which loudly voiced its displeasure with the disproportionate amount of resources going to English. The superintendent was not personally invested in this program and when he did go to observe it, he was not favorably impressed. The class was chaotic, noisy, and did not seem to present a productive learning environment. Further, the teacher of the program seemed stressed.

When it was time to build next year’s budget, the superintendent expected the district to cut the math program, assuming it would show poor results, which would help build support among the rest of district’s management for cutting the program. However, examining data on program effectiveness was an important principle of the budget process, so the data for both programs were analyzed. The district discovered that the scores of the math students greatly exceeded expectations: on average, students made 18 months of progress in a year. Meanwhile, there was virtually no detectable improvement among the participants in the English writing program, considering both grades and the quality of writing samples. It turned out that the noise the superintendent had observed in the math program was actually the natural byproduct of middle schoolers getting excited about something (in this case, math) and the chaos was partially a result of students actually sneaking into the class because they had heard that this was the place where they’d finally conquer math. Conversely, the apparent order in the English class was actually because, as attendance data showed, about half the students were cutting the class (so they weren’t there to cause disorder) and the concentrated work between students and teachers turned out to be not much more than a glorified study hall where students would get tutoring on their regular classwork, rather than systematic instruction on how to improve their lagging writing abilities. In addition, the English program cost almost four times as much.

“Common sense” rationales for programs are not sufficient for ensuring that the public is getting value for its tax dollars.
The decision then was clear — the entire district management team agreed to cancel the English program, and a new English program was modeled on the successful math program.

**LESSONS LEARNED**

The following lessons can be learned from the two case studies, illuminating the path forward for other early adopters of evidence-based decision making.

**Enshrine Evidence-Based Decision Making as an Explicit Principle that Defines Your Approach to Budgeting and Planning.** The superintendent’s first inclination was to cancel the math program based on his personal observation of how the program was performing. However, because the district’s management team had agreed that it would use data to inform its decision making, he was required to look at the data first.

**Beware the Pitfalls of Personal Observation.** Personal observation has an important role in decision making, but because observations are vulnerable to our cognitive biases, data are needed to keep observations grounded in objective reality. In the second case study, the superintendent fell victim to “confirmation bias,” which means that we interpret events in ways that support our existing opinions.

**Where Possible, Use Multiple Sources of Evidence.** Though objective, data are an abstraction of reality. As such, looking at just one kind of data gives us an incomplete perspective on reality. For example, looking at both grades and writing samples provided a much better assessment of the reading program than just one or the other. TCAPS looked at test scores as well as stakeholder satisfaction to assess the curricula.

**Examine Cost-Effectiveness, not Just Effectiveness.** In the second case study, the difference in the academic impact of the two programs was striking, but when the cost differential factored in, the decision that needed to be made was undeniable. Similarly, considering the cost of the different math curricula made for a much richer and financially savvy decision for TCAPS.

**Run Government Like a Scientist, but Not Exactly Like a Scientist.** TCAPS compromised on the randomness of its participants because volunteers would give their full energy to the experiment. While TCAPS did not quite reach the level of the RCT “gold standard,” it was still far better than a conventional planning and budgeting approach. Accommodations can be made to the practical realities of managing a political organization while still honoring the spirit of the scientific method and making vast improvements to decision-making.

**Notes**

1. Witness the major Hollywood productions that have been made about the genesis of technology companies, (e.g., “The Social Network,” “Steve Jobs”) and the cultural fascination with being an “entrepreneur.”
3. Ibid.
5. “Justice Department Discourages the Use of ‘Scared Straight’ Program,” OJJDP News @ A Glance, United States Office of Juvenile Justice and Delinquency Prevention.
8. Information on the counties’ experiences are adapted from “Results First at the Local Level: Evidence-based Policymaking in Four California Counties,” September 2015, The Pew Charitable Trusts and the MacArthur Foundation. Results First is used by more than 20 additional jurisdictions, including a number of state governments.

SHAYNE KAVANAGH is the senior manager of research for GFOA and can be reached at skavanagh@gfoa.org. NATE LEVENSON is the president of the District Management Council. Before that, he was the superintendent of Arlington Public Schools in Massachusetts.

The authors would like to acknowledge the contributions of Mike Bailey, finance director, City of Redmond, Washington; Sara Dube, director, Pew-MacArthur Results First Initiative; Peter Hutchinson, managing director, Accenture Public Services Strategy; and Andrew Klein, budget director, City of Baltimore, Maryland.