DOING EVIDENCE RIGHT

By Shelley Metzenbaum and Steven Kelman

This article originally appeared in PA TIMES, Fall 2017, Vol. 3, Issue 4. PA TIMES is a quarterly magazine produced by the American Society for Public Administration (ASPA), whose mission is to advance the art, science, teaching and practice of public administration. It is the largest and most prominent professional association for public administrators and those who study it. For more information, please visit www.aspanet.org and http://patimes.org.

For the two of us, who believe deeply that government can, should, and does improve the quality of people's lives – but who also believe, more broadly, that as individuals and as a society we should live our lives figuring out facts, learning from experience and, where relevant experience is missing, applying the best logic we can muster – the idea of using evidence to guide government decisions comes very close to a bedrock belief. In an age of anti-rationality and "alternative facts," we embrace this belief more than ever.

But we also want to do evidence right. The dominant view in the evidence-based government community is that we use evidence to figure out "what works," and then direct resources towards what works and away from what doesn't. It is also that the best method to produce evidence is randomized controlled trials (RCTs), and that we should therefore give pride of place to RCTs for the evidence in evidence-based government. Neither of these ideas is completely mistaken by any means. Both, however, have real limits that we need to take into account if we want to do evidence right.

We should not use program evaluations primarily to define programs as "effective" or "ineffective," but to help find ways to improve.

Many "What Works Clearinghouses" and other efforts to facilitate the search for evidence-based programs eligible for government funding suffer from oversimplified findings. They use average results to designate programs as effective, promising, ineffective, or inconclusive. However, a program or practice is often neither fully "effective" nor "ineffective," and paying attention to variations in performance is important. When that is the case, which it is most of the time, it is essential to look more closely at for whom, where, and when the treatment worked or not, especially when government spending is restricted to evidence-based practices.

One reason a program might not be either "effective" or "ineffective" is that, as our colleague the renowned statistician Dick Light pointed out many years ago, a "program" may, in fact, have tens or even hundreds of different design features, some of which may be effective and others ineffective. If we simply call something "job training" without attending to its many component parts and the multiple potential permutations and combinations of those parts, any conclusions about program effectiveness may well be meaningless. To learn something, we typically need to understand which design features are or are not associated with success.

A second reason a program might be neither "effective" nor "ineffective" is that some programs may show beneficial effects for only a subset of the population. On the one hand, a program might be ineffective for most people but effective for a few. Paying attention to the one person who benefitted from an otherwise failed drug trial led to the discovery of a class of patients who responded well to an "ineffective" drug, The New York Times explained in its story about pembrolizumab (now marketed as

Keytruda.) "The drug is the happy result of a failed RCT. A nearly identical drug was given to 33 colon cancer patients, and just one showed any response — but his cancer vanished altogether." Because at least one doctor took the time to follow up on the one patient for whom an otherwise ineffective treatment worked to identify detectable characteristics of the patient that might explain his health improvement, the doctor was able to find other people with tumors whom the drug might help, run a trial on a sample of those, and develop a drug likely to help 60,000 people a year in the United States alone. Conversely, a program may be effective for the majority of people, organizations, places, or situations but not for everyone. For example, as Anthony Bryk has pointed out, the first grade literacy program, Reading Recovery, is effective for most children, but not for a substantial number. By trying to decipher the characteristics of the children for whom a program does not work, policymakers can decide whether those not being helped should be a priority and, if so, follow up with additional testing to find practices that work for them.

An evaluation should be seen as a waystation on a journey to performance improvement, not the last stop.

This discussion of the fact that programs are often not merely "effective" or "ineffective" calls to our attention a crucial feature of doing evidence right. Evaluations normally try to determine whether or not a program worked and should be replicated, but typically do not discuss how the program can be improved. Too often, evaluation results are the end point of the journey, followed only by a decision to fund or defund. By contrast, we believe that a crucial part of doing evidence right is to see most evaluations, particularly for important programs, as a waystation along a journey to performance improvement. We believe evaluations should more often adopt an iterative learning approach to evidence-based government – not stopping the program if the results of an RCT are favorable for most but beginning there, tweaking the ingredients and practices a program uses to find ways to improve over time. Even when a proven set of practices is known, government should continue to test adjustments, measuring frequently, to find ways to enhance performance on multiple dimensions – not just on outcomes, but also on other important dimensions of performance such as people's experience with government and unwanted side effects.

This way of thinking about evidence moves the evaluation-oriented tradition of evidence-based government closer to the tradition of performance measurement and management, where one measures performance on objectives, sees where shortcomings might be, and then tries alternatives to current practice to see if they work better, rather than simply deciding to stop the program for being ineffective.

To take an iterative, learning approach towards evidence-based government, we must move beyond using evidence only from randomized controlled trials (RCTs).

In the performance measurement and management tradition, one typically uses data to inform goalsetting and then tests alternative ways to deal with a problem using methods that are considerably less rigorous than RCTs, sometimes, for example, using convenience samples more readily available to real life organizations and with considerably smaller sample sizes. These methods have a real virtue of providing fast feedback. Furthermore, if the alternative to the less rigorous approach in the performance measurement tradition is not an RCT but no testing at all, we should be careful not to let the perfect be the enemy of the good. Evidence from analyzing performance measures is a good way to start the performance improvement journey without a full-blown RCT.

Analytics applied to performance and other data to show trends, find variations in patterns across different subsets of those being affected as well as positive and negative outliers, anomalies, and relationships are valuable non-RCT sources of evidence. This sort of analysis helps detect problems needing attention, find promising practices worth trying to replicate, inform priorities, detect root causes to try to influence, and refine program design based on the analytics. We have become increasingly aware of how the private sector analyzes "big data" to understand individuals' purchasing patterns, leading to increases in sales thanks to more targeted marketing. Parts of government have begun analyzing big data to look for anomalous patterns that might point to fraud. There is also an increasing amount of discussion of and experience with predictive analytics in government – analyzing past trends to predict the results of various interventions and using other statistical methods to draw conclusions about which approaches to a problem are worth focusing on and likely to work.

Useful evidence comes in multiple shapes and sizes. As previously noted, RCT's are typically treated as the "gold standard" of evidence by evidence-based advocates. RCTs are typically relatively large, costly, lengthy and, as a consequence, rare. If we want to expand the scope of evidence-based government, and not the least if we want to adopt an iterative, learning approach that sees evaluations as a waystation in a journey of continuous improvement, we need to use not only performance analytics but also other forms of evidence that are quicker and less-expensive to gather. Many programs are learning to employ "rapid-cycle evaluations," essentially small-scale, quickly executed, iterative RCT's that test the impact of discrete changes in policy, management, or practice rather than evaluating an entire program to point to ways to improve. This approach to RCT's may be on the rise because of growing familiarity with agile IT and web-design practices, such as A/B testing, which uses random assignment principles to examine the impact of alternative web design features. Another form of a small-scale RCT is "nudge" interventions to test alternative ways to achieve a variety of objectives, such as reducing hiring bias, increasing taxpayer compliance, or boosting school attendance rates.

In addition to RCTs and performance analytics, role-playing exercises such as FEMA's tabletop exercises are another useful source of evidence, predicting how people are likely to act in different situations and problems that are likely to arise when responding to actual events while also providing those involved in the exercises the opportunity to practice, learn from experience, and sort out future roles.

When using non-RCT information, it remains important to interpret the information through the lens of causal thinking, asking what the performance trends might have been under different scenarios, including without the government action that is under consideration for continuation, expansion, adjustment, or elimination.

Defunding what does not work does not always make sense.

It is appealing, and sometimes justified, to defund programs that don't work. Indeed, this is a strong reason to use evidence, especially to counteract political forces encouraging a failing status quo. Nonetheless, we should understand that if a government program does not work, the problem the program was intended to address likely still exists. If the problem itself is serious, we should be cautious about prematurely removing funding from programs that don't work before going through a performance improvement journey to try to locate new approaches, or evidence about positive outliers, that might produce improvement.

Evidenced-based government is a good thing – and clearly better than the alternative. But we should try to do evidence right to achieve the most learning and on-the ground benefits with the fewest costs.

About the Authors:

Shelley Metzenbaum is currently creating The BETTER Project to encourage a better world through better government and a Senior Fellow at the Volcker Alliance. She led federal efforts to improve government outcomes, efficiency, and understandability as Office of Management and Budget (OMB) Associate Director for Performance and Personnel Management during the first term of the Obama Administration, and subsequently served as founding president of The Volcker Alliance. Before that, she was U.S. Environmental Protection Agency (EPA) Associate Administrator for Regional Operations and State/Local Relations; Undersecretary of Environmental Affairs and Capital Budget Director in Massachusetts; founding director of UMass Boston's Collins Center for Public Management; and director of the Kennedy School Executive Session on Public Sector Performance Management.

Dr. Steven Kelman is the Weatherhead Professor of Public Management at Harvard University's John F. Kennedy School of Government. A *summa cum laude* graduate of Harvard College, with a Ph.D. in government from Harvard University, he is the author of many books and articles on the policymaking process and on improving the management of government organizations. From 1993 through 1997, Dr. Kelman served as Administrator of the Office of Federal Procurement Policy in the Office of Management and Budget. During his tenure as Administrator, he played a lead role in the "reinventing government" effort, including the Federal Acquisition Streamlining Act of 1994. He is a Fellow of the National Academy of Public Administration. In 2001, he received the Herbert Roback Memorial Award, the highest achievement award of the National Contract Management Association. In 2003 he was elected as a Director of The Procurement Roundtable, and he was inducted in 2007 into the *Government Computer News* Hall of Fame. He writes a weekly blog called The Lectern, which is regularly one of the most downloaded links on the FCW.com site. (https://fcw.com/blogs/lectern/list/blog-list.aspx)