Draft Working Paper

How Even the Best Evidence Can Yield Bad Decisions, and What We Can Do About It

Scott Cody
Senior Vice President, Insight Policy Research
Send comments to scody@insightpolicyresearch.com
Draft September 16, 2020

Abstract: While evaluation evidence is the best evidence for identifying "what works," it is still imperfect. In evidence-based policymaking, the imperfections create a significant risk we will implement programs that do not work, and we will suffocate programs that do—or at least can—work. Numerous factors, including limitations of external validity, an overreliance on *p*-values and hypothesis testing, underpowered research, and even our intolerance for false positives can lead to incorrect conclusions. This paper discusses various ways evidence-based decisions can still be bad decisions. The paper identifies two trends that can address these shortcomings: the use of Bayesian statistical methods and continuous quality improvement. The paper concludes with six recommendations for strengthening evidence-based policymaking: (1) eliminate language that oversimplifies program effectiveness, (2) eliminate bright-line funding decisions, (3) promote Bayesian methods in evaluation and evidence-based decisionmaking, (4) invest in and reward program improvement, (5) incorporate improvement evidence in evidence ratings, and (6) create a clearinghouse of improvement findings.

Over the past 20 years, policymakers have made increasingly sophisticated use of evaluation evidence when choosing which social programs to implement. The efforts have been the cornerstone of the evidence-based policymaking movement and have represented a significant step forward for the programs, their funders, and their beneficiaries. The work has helped focus limited resources on those programs with the most potential. Indeed, a 2014 report on evidence-based policymaking suggested the practice can reduce wasteful spending, expand innovative programs, and strengthen accountability (Pew Charitable Trusts & McArthur Foundation, 2014). Because of these benefits, evidence-based policymaking has received broad support across the public, private, and philanthropic sectors.

But while evaluation evidence is the best evidence for identifying "what works," it is not perfect. Simply put, our current approach to using evaluation evidence carries significant risk we will implement programs that do not work, and we will suffocate programs that do—or at least can—work.

This paper begins with an overview of how evaluation evidence is used in evidence-based policymaking. I then describe how misuse and misinterpretation of evaluation evidence can yield bad policy decisions. I conclude with a discussion of ways to improve how we used evaluation evidence in evidence-based policymaking.

Using Evaluation in Evidence-Based Policymaking

Evidence-based policymaking refers to the use of information—including program evaluations, policy simulations, performance monitoring, and descriptive statistics—when making decisions about program policy and operations. For this paper, I am focused specifically on evidence-based policymaking initiatives that use evaluation evidence when making decisions about which programs receive funding and/or are implemented. These initiatives include evidence-preferring guidelines and explicit evidence requirements created by federal, state, and local governments and by philanthropic entities.

Most of these evaluation-focused evidence-based policy initiatives include a strong preference for *rigorous* evaluation evidence. Rigorous evaluations use methods that identify whether changes observed in target outcomes are actually caused by the social program and not other factors. Randomized controlled trials (RCTs) hold the "gold standard" reputation for rigorous research, and most evidence-based policy frameworks explicitly prefer evidence from RCTs. RCTs are not always feasible or practical, so frameworks also allow for some other types of research designs to inform policymaking.

Two broad types of evidence-based policy frameworks use evaluation evidence to guide decisions related to which programs receive funding and/or are implemented.

1. Initiatives where evidence is a preferred factor for funding and implementation

Several initiatives provide or encourage the use of research evidence as a consideration when decisionmakers select social programs. These initiatives are grounded in the assumption that making that evidence accessible to decisionmakers will generate better decisions. A common example can be observed in systematic reviews of research that summarize research evidence in a way that highlights "what works."

The U.S. Department of Education's What Works Clearinghouse (WWC) is an oft-cited example.[1] The WWC is grounded in a set of research evidence standards. WWC personnel review studies to assess whether they were designed and executed in ways consistent with these standards. Studies that meet these standards are rigorous, and their results can be trusted. For individual education programs (such as a curriculum), the WWC summarizes the findings of all studies that meet standards. If rigorous research demonstrates positive impacts, the program receives a positive rating by the WWC. In short, for each education program reviewed, the WWC answers the question: What does the most trustworthy research say about the impact of that program?

The WWC is intended as a resource for state and local education decisionmakers. For example, a school district considering purchasing Odyssey Math—an online mathematics instruction program for elementary schools—may want to know if there is evidence the product raises test scores. However, the district may not have the training or time to review the more than 20 studies evaluating the effectiveness of Odyssey Math. They can turn to the WWC, which determined that only three of the studies were rigorous enough to be trusted, and these studies showed students' math scores increased

[1] The author served as deputy director of the What Works Clearinghouse from 2009 to 2015.

as a result of Odyssey Math (U.S. Department of Education, 2017). The district can factor this evidence into their decision on whether to purchase the product.

While the WWC was one of the first such clearinghouses, it is not unique. Evidence clearinghouses support programmatic decisions across a host of social programs, including home visiting, employment and training, child welfare, health, and other initiatives (table 1). The WWC and other clearinghouses can trace their roots to the Cochrane Collaboration, which started in 1993 and has grown into a network of tens of thousands of healthcare researchers and practitioners who review research on clinical practice.

Table 1. Evidence Clearinghouses to Support Social Program Decisionmaking

Clearinghouse	Focus	Funder
Best Evidence Encyclopedia	Education	Johns Hopkins University, U.S. Department of Education
Blueprints for Healthy Youth Development	Youth Development	Arnold Ventures
Campbell Collaboration	Multiple	Multiple
California Evidence-Based Clearinghouse for Child Welfare	Child Welfare	California Department of Social Services
Clearinghouse for Labor Evaluation and Research	Employment and Training	U.S. Department of Labor
Cochrane Collaboration	Clinical Healthcare Practice	Multiple
Crime Solutions	Justice	U.S. Department of Justice
Employment Strategies for Low Income Adults	Employment and Training	U.S. Department of Health and Human Services
Evidence Exchange	National and Community Service	Corporation for National and Community Service
Home Visiting Evidence of Effectiveness	Home Visiting Programs	U.S. Department of Health and Human Services
National Guidelines Clearinghouse	Clinical Healthcare Practice	U.S. Department of Health and Human Services
Prevention Services Clearinghouse	Child Welfare	U.S. Department of Health and Human Services
Results First	Multiple	Pew Charitable Trusts and the John D. and Catherine T. MacArthur Foundation
Social Programs That Work	Multiple	Arnold Ventures
Teen Pregnancy Prevention Evidence Review	Pregnancy Prevention	U.S. Department of Health and Human Services
What Works Clearinghouse	Education	U.S. Department of Education
What Works for Health	Multiple	Robert Wood Johnson Foundation
What Works in Reentry	Justice	U.S. Department of Justice

In addition to evidence clearinghouses, numerous legislative initiatives at the federal and state levels seek to encourage the use of evidence by decisionmakers. For example, the federal Every Student Succeeds Act (ESSA) of 2015 encourages state and local education agencies to prioritize evidence-based interventions (U.S. Department of Education, 2016). Although some funding streams in ESSA require evidence as an eligibility criterion (see next section), most of ESSA focuses on encouraging education administrators to consider evidence when selecting interventions. Likewise, the Department of Labor's 2019 guidance on Unemployment Insurance Reemployment Services and Eligibility Assessments grants encourages states to "use evidence-based strategies where they exist" (Results for America, 2019).

The scoring criteria for numerous federal grant programs allocate points if the grant proposal includes evidence-based programs or practices. While these grant programs do not require a specific level of evidence to be eligible for funding, programs with strong evidence are more likely to receive funding, all

else being equal. Examples of grant programs prioritizing evidence include the U.S. Department of Education's $1 billion TRIO program, the U.S. Department of Housing and Urban Development's $2 billion Continuum of Care program, and the Department of Labor's $85 million YouthBuild program (Results for America, 2018).

Finally, a key goal of the Foundations for Evidence-Based Policymaking Act of 2018 is to expand federal agencies' ability to generate and use evidence in decisionmaking. Specifically, the act directs agencies to develop evidence plans, designate evaluation officers to oversee evidence activities in agencies, and periodically assess their own ability to use evidence in day-to-day government operations (Hart & Shaw, 2018).

2. Initiatives where evidence is required for funding and implementation

Other initiatives do more than just prioritize evidence when making funding decisions: They require programs to have a specified evidence base as a precondition for funding. The U.S. Department of Education Innovation and Research program (funded through ESSA) ties the size of grant amounts to evidence of effectiveness. Using the same evidence standards as the Department's WWC, grantees can receive the largest grants only if they meet certain evidence thresholds. The program stipulates that education efforts supported by limited evidence can receive "early-phase" grants of up to $4 million, efforts supported by some rigorous evidence can receive "midphase" grants of up to $8 million, and efforts with strong rigorous research demonstrating positive impacts can receive "expansion" grants of up to $15 million.

The Family First Prevention Services Act of 2018 allows states to use federal child welfare funding for services designed to prevent children from entering foster care. The act stipulates that 50 percent of state expenditures must go toward "well supported" programs—those with multiple, high-quality studies showing positive impacts (McKlindon, 2019).

Numerous philanthropic organizations that invest in social programs make evidence a condition of funding eligibility. For example, Arnold Ventures has an open request for proposals to fund the implementation and evaluation of social programs with an existing rigorous research base showing positive impacts (Arnold Ventures, 2019). The Overdeck Family Foundation prioritizes grants to evidence-based education programs that "have a strong record of results over time [and] are able to present counterfactuals that prove significant benefit for participants ..." (Overdeck Family Foundation, 2020). Blue Meridian Partners make grants of $100 million or more in programs that have rigorous evidence of effectiveness. While many of these and like-minded philanthropic organizations fund and evaluate social programs that have yet to establish an evidence base, rigorous evidence of impact is nevertheless a key to unlocking some sizeable philanthropic funding streams.

Another way evidence is used in the funding of social programs is pay for success (PFS) financing. Under PFS, private investors enter into an agreement with a government agency to provide the upfront capital needed to operate a given social program. The agreement specifies the government agency will reimburse the investors if evidence demonstrates the social program achieved targeted impacts. In this way, governments pay only for demonstrated impacts (while investors bear the risk that programs are ineffective). Like other evidence-based policy initiatives, PFS initiatives often prioritize evidence from rigorous evaluation designs.

The recent Social Impact Partnerships to Pay for Results Act (SIPPRA) of 2018 is designed to expand the use of PFS. SIPPRA provides $100 million in funding to support state and local government PFS projects.

To be eligible for SIPPRA funds, PFS payments must be based on outcomes that "meet evidence standards for high quality experimental or non-experimental research" (U.S. Department of Treasury, 2019).

The federal Maternal, Infant, and Early Childhood Home Visiting (MIECHV) program now combines an evidence requirement for programs to be implemented with a PFS option. MIECHV requires grantees to implement programs deemed evidence-based by the Home Visiting Evidence of Effectiveness clearinghouse. The Balanced Budget Act of 2018 amends the MIECHV program to allow grantees to use up to 25 percent of their grant in a PFS framework, not only requiring evidence for a program to be administered, but also requiring evidence for the government to make payments for those programs (Fudge et al., 2019).

Making Mistakes Through Evidence-Based Policymaking

As a result of evidence-based policymaking initiatives like these, the stakes for evaluation evidence are high. Programs that have rigorous evidence of effectiveness can receive promotion through clearinghouses, priority in decisionmaking, and even exclusive access to large funding streams. Programs with one or more studies showing no impact find themselves at a funding disadvantage, even when their evidence base is mixed with both positive and no-impact findings. As a result of these stakes, we would hope our evidence could correctly classify programs as "it works" and "it does not work."

Unfortunately, the methods and assumptions underlying evidence-based policymaking can lead us to make the wrong conclusion about the effectiveness of a social program. In some cases, they can lead us to conclude a program works when it does not. In other cases, it can lead us to conclude a program does not work when it does—or when it *can* (table 2).

Table 2. Sources of Mistaken Conclusions in Evidence-Based Policy

Concluding "It Works" When It Does Not	Concluding "It Does Not Work" When It Does (or Can)
▪ Just because it worked somewhere else does not mean it will work here ▪ Just because it worked in the past does not mean it will work now ▪ Relying solely on statistical significance to identify impacts can generate large numbers of false discoveries	▪ Just because it is not working now does not mean it cannot work when it is improved ▪ Rigorous evaluations cannot prove "it does not work" ▪ Underpowered research makes "no impact" findings a fait accompli ▪ Our intolerance for false positives leads to false negatives ▪ Relying on measures of the average impact masks where programs actually do work

Mistake 1: Concluding it works when it does not

Nothing makes those of us involved in designing, administering, and evaluating social programs happier than the discovery of statistically significant positive impacts. It works! Even better is replicated evidence compiled from multiple, rigorous evaluations. Yet even in the optimal-yet-rare situation where we have strong evidence of effectiveness, we run the risk of falsely concluding a program shown to work in research will work in the next place we implement it. What is worse, using traditional statistical

methods, we run a surprisingly high risk of falsely identifying a program is effective, even when using gold standard RCTs.

What Works Where?

One reason we falsely conclude a program will work is the assumption that the conditions reflected in the research are similar to conditions where we will implement the program. This is more than just a problem with taking research out of the lab and into the real world. Much of the research we draw on in social policy evaluations was conducted in real-world settings. However, just because something works in one real-world setting does not mean it will work in another (Orr et al., 2019).

This problem—known as external validity—is well known. External validity has a location component to it. A program effective with students in 1 city—or even in 10 cities—may not be effective with students in *your* city. This could be because the students in your city are meaningfully different than the students in the study cities. They may be more or less diverse, with different cultural experiences and possibly different primary languages. Another reason could be the social and economic conditions in your city are different from the conditions in the study cities. Students in your city may have more or less access to opportunity and may have grown up in economic conditions that foster or impede educational attainment. In other words, because your population and setting are different, the program that worked elsewhere may not work for you.

It is also useful to keep in mind that research identifies "what works" by making comparisons. The effectiveness of a new job training program may be determined by comparing earnings for individuals in the new program with earnings of those receiving "business-as-usual" services. Two cities with identical populations may experience different results from this training program because their "business-as-usual" services differ. In one location, the business-as-usual services may be poorly administered, making the new program relatively more effective. However, in another location, well-executed business-as-usual services may mean participants in the new program would achieve the same levels of earnings as they would have if no change were made. Policymakers in this second location may think the new training program will improve earnings for their clients, but it will not.

What Works When?

External validity also has a time component. A program shown to be effective 10 or 15 years ago may not be effective today. One example is the oft-cited HighScope Perry Preschool intervention, an RCT that concluded children enrolled in preschool fare better on a host of near-term and long-term outcomes than children not enrolled in preschool.[2] This study, conducted in the early 1960s, is still cited today as evidence of the power of preschool. However, the study was conducted at a time when early childhood care was much less common. Students not in preschool were more likely to be in unstructured home environments.

Today, with higher rates of working mothers and expansion of childcare and preschool opportunities, many children are in some form of structured early childhood program. It is unrealistic to expect the effects of preschool relative to nothing in the 1960s are the same as the effects of preschool relative to something in 2020. Yet policymakers still cite this research assuming more investments in preschool will yield comparable improvements in outcomes for today's children. Indeed, the HighScope website

[2] There is some debate that postrandom assignment changes made by researchers in the HighScope Perry Preschool study undermine the causal validity of this research (Social Programs That Work, 2017).

presents return-on-investment numbers derived from this 1960s study as if similar returns would be achieved today (HighScope Educational Research Foundation, 2020). As much as I hope such returns are attainable, I question the reliability of data from research more than 50 years old.

Other factors are also changing constantly. In education, new technology is introduced regularly. A math program shown to work 10 years ago may be less effective now when classrooms have access to a host of online tools and activities. Just because an expensive mathematics curriculum worked 10 years ago does not mean it is a worthy investment today: Perhaps the school district could get the same outcomes leveraging new technology. We do not know what the evaluation would show if it were conducted today.

This issue extends beyond education. We can think of the limited number of social program interventions shown to improve outcomes for the most disadvantaged populations, including employment, wages, nutrition, family stability, etc. Much of this research has been conducted over the past 30 years. However, over the past 10 years, many of the communities that are home to disadvantaged populations have been devastated by an unprecedented opioid epidemic. It is entirely reasonable to think services that were effective at promoting positive outcomes 20 years ago cannot be effective in this context. Employment, nutrition, and family services needed to help someone coping with opioid addiction—their own or that of loved one they are caring for—are likely different from what may have been effective 20 years ago.

In other words, the real world is diverse and constantly changing, but our evidence-based policy approach assumes what works in one location works anywhere, and what worked 10 or 15 years ago will work today. These assumptions are shaky at best and can lead policymakers to invest in ineffective programs for their community.

Misinterpreting p

In 2016, the American Statistical Association (ASA) published a widely read paper castigating the research community for its widespread misinterpretation of *p*-values (Wasserstein & Lazar, 2016). The *p*-value is a hallmark of the frequentist approach to statistics. While *p*-values drive most research conclusions in practice, the ASA made clear that "[b]y itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis." Stated differently, **"a label of statistical significance does not mean…[an] effect is highly probable, real, true or important"** (Wasserstein, Schirm & Lazar, 2019).

The *p*-value is the probability that an impact greater than your estimate would occur by chance when the true impact is zero. If you conduct 100 studies of the same *ineffective* program (that is, the program has zero impact), then 5 of those studies would yield an impact estimate so large it would be statistically significant at the $p < 0.05$ level. In each individual study, the *p*-value tells you nothing about whether the program is effective; it tells you only how likely it is you would obtain your estimate if the program were ineffective.

The ASA paper sought to remind us that a *p*-value is *not* the probability that the impact of a program is the result of random chance. If you want to know the probability that a significant impact is due to randomness alone, you need to know the false discover rate (FDR). The FDR represents the percentage of all potentially significant findings that would be false if you studied all programs. Unfortunately, the FDR cannot be calculated directly because much of the information needed is unknown. However, we

know from arithmetic of the FDR calculation the false discovery rate will be high in situations where a program is effective in a relatively small proportion of locations.

My favorite example of this is explained clearly by Deke and Finucane (2019), who assume omniscient knowledge of a hypothetical social program. This program is run in 100 locations, but (for reasons like those discussed earlier), it is effective in only 10. One program location (of the 100) is randomly selected to be evaluated, and that evaluation is large enough to detect an impact with a power of 80 percent (that is, of the 10 program locations that are, in reality, effective, the study is large enough to detect 8 of those as effective). Statistical significance is determined at the $p < 0.05$ level, meaning 5 percent of the 90 ineffective program locations, if selected, will have an impact estimate with $p < 0.05$, and researchers will falsely conclude the program works. Under these assumptions, 13 of the 100 programs, if sampled, would be classified as effective. Among these 13 are the 8 programs that actually are effective (and can be detected with the study) plus the 5 programs that are not effective but have a $p < 0.05$ (technically 4.5 programs, since 90 * 0.05 is 4.5). The FDR—the percentage of times researchers would conclude "it works" when it does not—is surprisingly high. Of the 13 programs that could be detected as effective, 5—a whopping 38 percent—would be false discoveries. In other words, if you conduct a study in this reasonable scenario and find a positive impact, **there is a 38 percent chance that positive finding is false** (not less than a 5 percent chance, as misinterpretations of the *p*-value assume).

It does not get better from here. Our estimates are more likely to overstate the magnitude of the impact of a program when the true impact is small, and other factors can affect the outcome (that is, the outcome is noisy). In such cases, we are likely to overstate the impact of a given intervention by a surprisingly large amount (Gelman & Carlin, 2014; Gelman, 2017). If the outcome is noisy, estimates of the impact on that outcome will be error prone, and some erroneously large estimates will be statistically significant. As a result, statistically significant estimates will have a large bias when the true impact is small (because estimates close to the true impact are less likely to be statistically significant under these conditions). This means that **any statistically significant impact is likely to substantially overstate the true impact.** Unfortunately, this situation—a program has a small impact on an otherwise noisy outcome—is far too common for social programs. As a result, this bias further increases the risk we falsely conclude a program worked.

Issues like these led the ASA panel to conclude that using *p*-values results in "bad science" and bad decisions. ASA was particularly concerned about the dominance that this individual, misleading statistic carries in our evidence-based decisionmaking process. The believability and substantive importance of an impact estimate depends on multiple factors—not just *p*-values—yet we repeatedly ignore those factors and still claim our studies are rigorous.

Mistake 2: Concluding it does not work when it does (or can)

When research fails to show a program works, we often conclude the program does not work. Such a determination can result in decreased funding from Congress, an inability to access philanthropic funding, and/or local communities opting for a different, "proven" program model. But most research methods are not actually designed to tell us if something does not work. In another frustrating nuance of the commonly used statistical methods, the methods can tell us only that we fail to reject the hypothesis that something does not work. But even setting that confusing technicality aside, our statements that programs do not work can often be too broad, ignoring how they can work in certain circumstances, with certain individuals, and/or with certain processes.

Timing Is Everything

Discussions of social policy research typically include a tacit assumption that a social program is a clearly defined, fixed, immutable thing. This, of course, is not the case. Programs are constantly developing and evolving. Those designing and administering programs are all too aware of changing context (whether changing counterfactuals, changing priorities among funders, changing funding streams, changing staff, etc.). As a result, most social programs are works in progress. Changing circumstances can mean what works now may not work tomorrow; the same is true in reverse: What does not work today may work tomorrow.

This problem is most acute for new programs. Programs early in their development may be anxious for a rigorous program evaluation: positive findings from independent researchers can create buzz and open funding opportunities. However, it can take programs years to flesh out their theory of change, work out the kinks of their service delivery, and refine and adapt program operations to different contexts. Without this work, the programs are much less likely to be effective. Evaluating a program too soon in its development can suggest the program does not work, when it is possible the program does not work *yet*.

Mature programs may be required to adapt program services to changing contexts and/or new environments. As stated previously, programs that were effective prior to the opioid crisis may require adaptation to be effective in communities hard hit by opioid addiction. Evaluating a program as it seeks to adjust services may similarly result in concluding a program that is not working at the moment cannot work at all.

Similarly, programs with fully developed models may struggle with common operational challenges such as recruiting and retaining participants, improving staff skills, and leveraging technology. These challenges may be location specific and may ebb and flow over time. Rigorous evaluations may conclude programs struggling with these operational challenges have no impact, but it is entirely possible these programs could be effective if administrators could solve the operational challenges.

Proving a Negative

In frequentist null hypothesis testing, the null hypothesis is something we assume is false ("this program does not work"). We then use evidence to reject that null hypothesis (and therefore conclude "this program works"). Importantly, failing to reject a null hypothesis is not the same as proving the null hypothesis. If we *cannot* reject the null hypothesis, there is not much we *can* say.

For example, say my null hypothesis is that a species—let's call it *Programus Effectus*—is extinct. I can reject that null hypothesis if I find a single surviving *Programus Effectus*. That living, breathing *Programus Effectus* is evidence my null hypothesis is false. But failing to find a *Programus Effectus* does not prove *Programus Effectus* is extinct. It only means I have not found one. This same logic applies to drawing conclusions from null hypothesis testing in social policy.

Despite this, researchers and policymakers are apt to use a single study showing no impact as evidence a program does not work, when all we can really say is the study did not disprove the hypothesis that the program does not work. It is safer to conclude a program does not work if there are multiple well-designed, well-executed studies, all of which fail to reject the null hypothesis. Even then, though, we

need to keep in mind no study has proved the program does not work. The studies just failed to disprove it.

Power Outage

Even if we agree that multiple, well-designed, well-executed studies showing no impact are enough to conclude a program does not work, other limitations of these statistical methods can still lead us to the wrong conclusion. First, these studies must be large enough to identify an impact. This requirement—known as statistical power—is not defined in absolute terms but in terms relative to the outcome we are trying to impact. I am often asked how many participants are needed in a study to tell if a program works, and my response—"it depends"—is never satisfying to me or the person asking. But it depends. I need more individuals in an evaluation if the outcome is highly variable from individual to individual than I do if the outcome is relatively consistent from individual to individual. Basically, if the outcome is noisy, it will require more observations to determine if outliers represent impact or just more noise.

Unfortunately, while many researchers understand the principles of statistical power, many nevertheless conduct underpowered studies. Often this is because real-world research constraints (limited funding or limited numbers of program participants) present the researcher with a choice of conducting an underpowered study or no study at all. The researcher may conclude there is at least some value in conducting the underpowered study. But in reality, the findings of such underpowered studies—that the program has no detectible impact—are evident before the first participant is enrolled. This does not mean the program does not work. It just means the researcher spent a lot of resources even though they would never be able to distinguish impact from noise. Unfortunately, underpowered studies get included in the dialog of "what works" all too often.

Are False Positives as Negative as We Think?

Because hypothesis testing cannot answer the question we want to ask ("Did this program work?"), the standard $p < 0.05$ threshold for statistical significance is intentionally set as a high bar to minimize false positives. A higher threshold (say, $p < 0.1$) will lead us to reject the null hypothesis more frequently (and conclude programs work), but it will also generate more false positives.

Minimizing the risk of false positives makes sense when the consequences of a false positive are substantial. For example, in clinical drug trials, a false positive could lead patients to take ineffective drugs that are expensive and carry side effect risks. In that context, $p < 0.05$ reduces those risks. However, raising the bar on false positives opens the door to false negatives. When we set the bar high, by definition we accept increased risk of "failing to reject the null hypothesis" when programs actually do work.

Unfortunately, in social policy, researchers have relied on the default 5 percent threshold without a constructive dialog on the consequences of false positives. In some cases, the consequences could be substantial—exposing individuals to expensive services that may have both opportunity costs and unintended consequences. When we consider that many social policy evaluations are conducted in the context of "we're going to implement some program," then the consequences of a false positive may be less. If policymakers decide they are going to invest in some program and wish to select the most promising (as is often the case), then a series of studies failing to reject their respective null hypotheses provides them with little information to go on.

In short, our methods are designed to minimize false positives, and that increases the risk of false negatives, but this tradeoff was decided without the decisionmaker's input. Some situations may warrant higher tolerance of false positives, or better still, some situations may warrant knowing the *probability* something will work.

The Mean Mean

Much of the evidence currently supporting evidence-based policymaking reflects the "average treatment effect" (ATE). For example, when we evaluate a supplemental instruction program for elementary school students who struggle with reading, we compare the average reading performance of students with supplemental instruction to the average performance of students without. The average is a statistic designed to convey just one characteristic of a distribution of data, but the average cannot tell you all you would want to know about any distribution of data: It's just the average. As a result, when we compute the ATE, we are simplifying what could be a complex distribution of program impacts into a single number.

Using ATE to identify "what works" would make sense if we believed all individuals participating in a program respond in a consistent, uniform way. This assumption is more than unrealistic. In practice, some individuals may benefit significantly, while other individuals not at all. The benefit of a supplemental reading program may depend on mediating factors, such as the nature of a student's learning disability, their home environment, or even their diet. The benefit of an employment and training program may depend on an individual's work history, education, or sadly, the color of their skin.

In *The End of Average*, Todd Rose details numerous limitations of this ubiquitous statistic (Rose, 2016). Rose explains an outcome "average" loses meaning when the factors influencing that outcome are loosely correlated. According to Rose, an outcome such as the average reading score from a standardized test (commonly used in education evaluations) provides little insight into student performance when students differ widely on underlying factors such as memory, cognition, vocabulary, curiosity, interest, primary language, etc. The average masks meaningful variation across different types of students. Rose is particularly concerned with education programs that attempt to design interventions for the average student: He argues there is no such thing as an average student. The principle holds when we try to evaluate programs as if the average program participant represents all participants, but in reality, it likely represents none.

As a result of our overreliance on the ATE, we are prone to conclude a program does not work even if it does work for some participants. Tools like the What Works Clearinghouse may identify a curriculum as ineffective when, in reality, it is effective for some types of students. In this way, telling a local school district that the supplemental reading program they are considering does not work, when in fact it works very well for students with a certain combination of interest, vocabulary, and cognitive profile, denies an effective resource to some students and may make regression to a mean impact of zero a fait accompli.

Improving Evidence-Based Policymaking

If we can make the wrong conclusion so easily, is there value in evidence-based policymaking? Yes, there absolutely is. Just because our evidence-based policymaking framework may lead to bad decisions does not mean we should abandon it. Rather—just like the social programs we examine—we should find ways to take what is promising and improve it.

Embracing Change

Two emerging trends in evidence hold promise for addressing the shortcomings of our current evidence-based policy framework. By embracing these trends, we can increase the reliability of evidence we generate and provide actionable intelligence for decisionmakers. The first trend is the use of Bayesian statistical methods in program evaluation. While Bayes' theorem is more than 250 years old, Bayesian methods have rarely been used in program evaluation. The second is the expansion of efforts to promote continuous program improvement. By both promoting and learning from continuous improvement, we can simultaneously improve the effectiveness of social programs and prevent ourselves from suffocating program innovation.

Embracing Bayesian Methods

Bayesian methods are an alternative to the frequentist *p*-value framework. Instead of conducting null hypothesis tests and classifying findings as "reject the null hypothesis" or "fail to reject the null hypothesis," Bayesian methods provide impact estimates on a continuum, estimating the probability associated with any impact estimate (Gelman, 2011). This avoids the bright-line (good–bad) framework inherent in more frequentist statistical methods.

For example, Bayesian methods support statements such as, "There is a 70 percent probability this program increases earnings by at least $100, a 90 percent probability it increases earnings by at least $50, and a 95 percent probability it increases earnings by at least $25." This framing takes the decision on whether to adopt a program out of the hands of the bewildering *p*-value and puts it into the hands of those making cost-benefit decisions. With results presented probabilistically, decisionmakers can weigh that probability against other contextual factors—such as cost and adverse consequences—to make an optimal decision for their context. For some decisionmakers, 70 percent confidence of a $100 impact may be too low; for others, it is an opportunity for improvement they cannot pass up. Researchers cannot—and certainly *p*-values cannot—make these judgments on behalf of decisionmakers.

Bayesian methods can support sophisticated analyses in which experimentation adapts "on the fly" as findings are generated (Finucane et al., 2017). They also support efforts to identify the "active ingredients" of a complicated social intervention. Such analyses are necessary if we wish to answer "what works for whom, under what circumstances, and how?"

Bayesian methods have only recently been adopted in program evaluations (see, for example, Kimmey et al., 2019). The methods are late to the scene in part because the underlying models are computationally intensive and require high-end processing capacity that only recently became widely available. But the methods also carry controversy. Their probabilistic estimates are generated by combining results from the current study with a set of prior assumptions. Some argue that prior assumptions introduce subjectivity into a process that should be objective, but the absence of prior assumptions is still a subjective assumption. By having no explicit prior assumption, traditional statistical methods implicitly assume an infinite range of impacts is possible. This is an unrealistic assumption in social policy evaluation that creates its own bias on impact estimates (Gelman & Weakliem, 2009).

The process of generating prior assumptions provides the opportunity to address the problems stemming from using older, potentially less relevant research in evidence-based decisionmaking. In generating priors, Bayesian statisticians can assign more weight to the existing evidence that is most relevant to the current context. Studies from 40 or 50 years ago have useful information, but we can

appropriately give them less influence in our conclusions than studies from the past 5 years. We can also vary this weight to see just how sensitive our conclusions are to differences between older and more recent research.

The methods are, of course, not perfect, but steps can be taken to ensure any biases resulting from the inclusion of prior assumptions are minimized or at least documented in the open. And when contrasted with the significant problems stemming from *p*-value based hypothesis testing, Bayesian methods reduce, rather than increase, the likelihood we will make a bad decision based on evidence.

Embracing Continuous Quality Improvement

Efforts to incorporate continuous quality improvement in programs' operations can ensure programs are enhanced to work for different populations at different times and in different locations. Continuous improvement starts with the assumption programs are neither immutable nor perfect, and with an intentional analytic approach, they can be made more effective. These efforts can be combined with rigorous program evaluations to help us understand how continuous improvement can generate better outcomes and give us greater insight into "what works for whom, under what circumstances, and how?"

Recently, there has been an increasing emphasis on continuous improvement for social programs. For example, the MIECHV program now includes a requirement that grantees demonstrate continuous improvement on key benchmark outcomes (Health Resources and Services Administration, 2018). Some federal agencies, such as the U.S. Department of Health and Human Services' Office of Planning, Research and Evaluation (OPRE) are funding efforts to improve program operations prior to program evaluation (including two recent efforts to address enrollment and retention for healthy marriage and responsible fatherhood programs). Other efforts have incorporated process improvement directly into program evaluation. For example, Year Up is a 1-year training program for underserved youth aimed at building skills and improving long-term employment outcomes. Year Up has been in operation since 2000 and has undergone numerous evaluations of effectiveness (Fein & Hamadyk, 2018). Yet one recent innovative study of Year Up evaluated the combination of Year Up's traditional services bundled with a continuous improvement framework, answering the question, "What is the impact of Year Up services combined with continuous improvement?" (Fein et al., 2018).

This trend is consistent with the attitude that something that does not work now may work in the future (and something that does work now may not in the future). Incorporating continuous improvement approaches into program delivery can increase programs' ability to have an impact and reduce the chances we conclude a program does not work when in fact it could.

Of course, this line of thinking—taken to the extreme—suggests we should never conclude a program does not work. We should be careful to avoid that extreme line of thinking. While we need to be cautious about when we conclude a program does not work, I am confident there are some programs that do not work and never will. I am even more confident some programs are a better investment (because they have greater impact for a given level of funding) than other programs. As a result, we need a way to distinguish those programs that can work (or are good investments) from those programs that likely will not work (or are bad investments).

Six Steps to Better Evidence-Based Policymaking

Below are six steps we as a policy and research community can take to strengthen the evidence-based policymaking framework.[3] None of these steps will be easy; each requires a shift in how we think about evidence and how we think about "what works." However, in making these shifts, we can increase the value that evidence brings to policy decisions.

1. **Eliminate language that oversimplifies program effectiveness.** It is appealing to think we can classify programs into clear "it works" and "it does not work" categories. Unfortunately, this is not consistent with the real world. In the real world, programs work in some places, not others. They work for some participants, not others. To classify a program based on its average effect is like classifying an ocean based on the average type fish. It simply does not make sense.

   The truth is, social programs address challenging and complex social issues. Indeed, if improving social outcomes were easy, we would have done it by now! Because it is complex, we must anticipate complex, nuanced solutions. Any attempt to oversimplify program effectiveness will lead to bad decisionmaking.

   At a minimum, we should establish an evidence-based policymaking framework based on more than just the answer to "what works"; our framework should routinely answer *what works for whom, under what circumstances, and how?* (OPRE, 2016). And we should stop referring to programs as "proven" to work; such language will overpromise and invariably create bad decisions.

2. **Eliminate bright-line funding decisions.** Even after shifting our language away from just "what works," we will still need to resist the temptation to make bright-line funding decisions, particularly on just one study, regardless of how well designed it is. Program effectiveness is not black-and-white. Treating funding decisions as if it were will lead to bad decisions.

   What decisionmakers really need to know:

   - What is the likelihood this program will work as intended in my community?
   - How big an impact am I likely to obtain in my community?
   - What are the monetary and nonmonetary costs I will incur implementing this program?

   In short, just because there is evidence something works elsewhere should not be enough to lead a decisionmaker to implement a program. Decisionmakers need to know whether the program will be cost-effective given the expected return in their community.

   Although there are challenges, it is possible to generate localized estimates of a program's likelihood of impact, the anticipated magnitude of impact, and the anticipated cost. This will require more data, refined analytic methods, and more sophisticated computer processing power. If we aspire to such a framework, we can make much better decisions.

3. **Be Bayesian.** We have extracted all the value we can from frequentist approaches to program evaluation. To take evidence-based policymaking further, we need to embrace a Bayesian approach to evaluating programs. Using Bayesian methods will help researchers avoid the various pitfalls associated with frequentist *p*-value-based hypothesis testing—pitfalls that can lead us to conclude an effective program does not work, and an ineffective program does. Bayesian methods will present results in a way that helps decisionmakers weigh costs and

[3] In addition to these steps, I recommend researchers consider recommendations by Wasserstein, Schirm & Lazar (2019) about how to generate constructive evidence in the face of inherent uncertainty.

benefits and evaluate risks. The methods can also generate greater insights into what works for whom, under what circumstances, and how.

Becoming more Bayesian will require major changes to how we conduct program evaluations. Bayesian approaches require complex, computationally intensive modelling. To expand Bayesian applications in social policy, we need to revise how we train future policy analysts and provide professional development to current policy researchers. We will need to ensure program evaluations have the robust computer processing infrastructure needed to support Bayesian modelling. We will also need to revise our standards for documenting, reviewing, and interpreting research findings.

Becoming more Bayesian will also require consumers of evidence to change how they make decisions. Instead of making decisions based on a simple—though possibly false—statement about "what works," decisionmakers will need to weigh likely benefits against likely costs, and make a decision about what is best for their context. This may seem like more work, and it is, but it is work that can lead to better decisions, and it is the type of analysis that decisionmakers bring to bear on other aspects of public management.

4. **Invest in and reward improvement.** Program improvement and program adaptation should be foundational elements of our evidence-based policymaking framework. We should acknowledge that any program we seek to evaluate and fund will not be successful if it is one-size-fits-all. Instead, investment in process improvement and program adaptation should be standard.

   The evidence-based policymaking funding structure can both offer incentive and reward improvement. For example, instead of just funding programs according to whether they meet an evidence threshold, some funding could be allocated to programs that do not yet demonstrate an impact but have a *high-quality process improvement framework, potentially certified by an accrediting authority*. Alternatively, funding could be allocated to programs that may not yet have demonstrated an impact, but they *show improvement on relevant program outputs and intermediary outcomes*. In this way, our funding incentives would encourage programs to strive for improvement and would not starve those programs that fail to be perfect out of the gate.

   If done right, such a funding structure would reward programs with demonstrated effectiveness and encourage continuous improvement for all programs. This would help ensure those programs that do not work yet have the best chance of working in the future, and those programs that do work now continue to work as context changes.

5. **Incorporate improvement evidence in evidence ratings.** Evidence ratings such as those from the What Works Clearinghouse are a critical component of the existing evidence-based policymaking infrastructure. However, these ratings focus only on effectiveness studies and not other types of evidence. If we agree our evidence-based policymaking framework must offer incentive for improvement, it stands to reason our evidence ratings should also incorporate evidence of improvement efforts.

   Of course, effectiveness studies are answering a different question than improvement efforts. Effectiveness studies answer, "Does this program work relative to other programs?" and improvement efforts answer, "Can this program get better over time?" The result is a potential for combining apples and oranges into one rating. However, we can revise our evidence-based framework to have more than one evidence rating. Imagine making a decision about a program where you know it has (1) no evidence of overall impact but (2) evidence of positive improvement over time. You may be more inclined to adopt such a program than one with no

evidence of impact and no evidence of improvement, or even evidence of some impact but no evidence of improvement.

I would encourage those engaged in developing evidence ratings to construct a framework for systematically reviewing and rating evidence from improvement efforts. This will not be easy. It will require standards and best practices for monitoring and evaluating improvement efforts. It will also require complex meta-analytic methods designed to combine findings from across a variety of methodologies, settings, and outcomes. Finally, because some researchers employ effectiveness evaluation methods to evaluate improvement efforts, such an effort would benefit from expanded use of research registries. Specifically, researchers engaged in overall effectiveness evaluations should declare as much in public registries prior to conducting the evaluation; ratings of overall effectiveness should include only evidence of effectiveness gleaned from registered studies. This will help avoid a situation where research used to evaluate improvements (many of which will fail) could hold back an otherwise promising program and/or diminish the incentive for much-needed improvement research.

The final reason to include improvement evidence in evidence ratings is to help decisionmakers make better decisions about when to keep investing in a program versus when to cut bait. While it is true a program that is not effective today could be effective in the future, it is also true some programs will never work. With a standardized process for tracking improvement evidence, we can begin to distinguish ineffective programs that may work in the future from ineffective programs that demonstrate they cannot improve.

6. **Create a clearinghouse of improvement findings.** Social programs of all shapes and sizes face similar operational challenges—including increasing enrollment, increasing engagement, increasing staff efficiency, managing technology changes, streamlining decisionmaking, tracking outputs, and acting effectively on data. All these factors can affect a program's ability to be successful. Yet even if there is a collective effort to promote improvement, we run the risk of helping thousands of programs reinvent the wheel.

   The federal government should invest in an improvement clearinghouse that catalogs various strategies used to address operational problems. This clearinghouse should be searchable to enable a given program to learn what programs in a similar setting tried, and how those trials fared. Such a clearinghouse should be careful not to suggest improvement strategies that worked for some programs will work for another program (indeed, I have often referred to this scenario as the "Your Mileage May Vary" clearinghouse). But it should be a way to spur ideas that programs can test and a platform for programs to share learning across domains.

Evidence-based policymaking is imperfect, but it can be improved. We can take concrete steps now to ensure policymakers 5 and 10 years from now are better equipped to provide effective services to diverse populations. Improved evidence-based policymaking will not happen until we admit we have a problem. And we definitely have a problem.

References

Arnold Ventures. (2019). *Demonstrating the power of evidence-based programs to "move the needle" on major U.S. social problems: Funding announcement and request for proposals*. https://craftmediabucket.s3.amazonaws.com/uploads/Moving-the-Needle-RFP.pdf

Deke, J., & Finucane, M. (2019). *Moving beyond statistical significance: the BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations* (OPRE Report 201935). Administration for Children and Families, Office of Planning, Research, and Evaluation.

Fein, D., & Hamadyk, J. (2018). *Bridging the opportunity divide for low-income youth: Implementation and early impacts of the Year Up program* (OPRE Report 2018-65). Administration for Children and Families, Office of Planning, Research, and Evaluation.

Fein, D., Maynard, R., & Warfield, G. (2018, November). *Building evidence across generations of a promising youth development program: Year Up*. Presentation to Association of Public Policy Analysis and Management Annual Research Conference, Washington, DC.

Finucane, M. M., Martinez, I., & Cody, S. (2017, November). What works for whom? A Bayesian approach to channeling big data streams for public program evaluation. *American Journal of Evaluation, 39*(1), SAGE Journals.

Fudge, K., Ballard, K., & Brown, M. (2019). *Funding home visiting with a pay for outcomes approach* (OPRE Report 2019-70). Administration for Children and Families, Office of Planning, Research, and Evaluation.

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, 2, 67–78.

Gelman, A. (2017, July). *The failure of null hypothesis significance testing when studying incremental changes, and what to do about it* (Working paper). http://www.stat.columbia.edu/~gelman/research/published/incrementalism_3.pdf

Gelman, A., & Carlin, J. B. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641–651.

Gelman, A., Hill, J., & Yajima, M. (2012, April). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power. *American Scientist*, *97*(4), 310–316.

Gopal, S., & Schorr, L. (2016, June). Getting "Moneyball" right in the social sector. *Stanford Social Innovation Review*. https://ssir.org/articles/entry/getting_moneyball_right_in_the_social_sector

Hart, N., & Shaw, T. (2019). *Congress provides new foundations for evidence based policymaking.* Bipartisan Policy Center. December 22. https://bipartisanpolicy.org/blog/congress-provides-new-foundation-for-evidence-based-policymaking/

Haskins, R., & Margolis, G. (2016). *Show me the evidence*. Brookings. September 7.

Health Resources and Services Administration. (2018). *Maternal, Infant, and Early Childhood Home Visiting Program: Guidance on meeting requirements to demonstrate improvement in benchmark areas.* https://mchb.hrsa.gov/sites/default/files/mchb/MaternalChildHealthInitiatives/HomeVisiting/MIECHV-Assessment-of-Improvement-Guidance-508.pdf

HighScope Educational Research Foundation. *Perry Preschool Project*. (2020). https://highscope.org/perry-preschool-project/

Kimmey, L., Anderson, M., Cheh, V., Li, E., McLaughlin, C., Barterian, L., Crosson, J., Stepanczuk, C., Timmins, L., Li, J., Heitkamp, S., Cheu, C., Fisher, T., Harvey, B., Johnson, H., Wu, B., Zhang, S., Finucane, M., & Eckstein, A. (2019, May). *Evaluation of the Independence at Home demonstration: An examination of the first four years*. Mathematica Policy Research. Office of the Assistant Secretary for Planning and Evaluation. https://www.mathematica.org/our-publications-and-findings/publications/evaluation-of-the-independence-at-home-demonstration-an-examination-of-the-first-four-years

McKlindon, A. (2019, October). Applying the research and evaluation provisions of the Family First Prevention Services Act. *Child Trends*. https://www.childtrends.org/publications/applying-the-research-and-evaluation-provisions-of-the-family-first-prevention-services-act

OPRE (Office of Planning, Research and Evaluation). (2016). *What works, under what circumstances, and how? Methods for unpacking the "black box" of programs and policies* (OPRE Report 2016-54). Administration for Children and Families.

Orr, L., Olsen, R., Bell, S., Schmid, I., Shivki, A., & Stuart, E. (2019, June). Using the results from rigorous multisite evaluations to inform local policy decisions. *Journal of Policy Analysis and Management. 38*(4). SAGE Journals.

Overdeck Family Foundation. (2020). *How we fund*. https://overdeck.org/grantmaking/how-we-fund/

Pew Charitable Trusts & Macarthur Foundation. (2014). *Evidence-based policymaking: A guide for effective government*. Results First Initiative.

Results for America. (2018, November). *Invest in what works federal standard of excellence*. https://results4america.org/wp-content/uploads/2018/11/2018-Invest-of-What-Works-Federal-Index_Interactive.pdf

Results for America. (2019, September). *The promise of the Foundations for Evidence-Based Policymaking Act and proposed next steps* (Results for America Evidence Act brief). https://results4america.org/wp-content/uploads/2019/09/Evidence-Act-Proposed-Next-Steps-FINAL.pdf

Rose, T. (2017). *The end of average*. Penguin Books.

Social Programs That Work. (2017, November). *Social Programs That Work review: Evidence summary for the Perry Preschool Project*. https://evidencebasedprograms.org/document/perry-preschool-project-evidence-summary/

U.S. Department of Education. (2016). *Non-regulator guidance: Using evidence to strengthen education investments*. September 16.

U.S. Department of Education, Institute of Education Sciences. (2017, January). *Odyssey® Math. What Works Clearinghouse intervention report*. http://whatworks.ed.gov

U.S. Department of the Treasury. (2019). *Social impact partnerships to pay for Results Act demonstration projects: Notice of funding availability*. https://home.treasury.gov/system/files/226/SIPPRA-NOFA-FINAL-FY2019.pdf

Wasserstein, R., & Lazar, N. (2016, March). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.

Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a World Beyond "$p < 0.05$". *The American Statistician*, 73(1), 1-19.