

Can Evidence-Based Policy Ameliorate the Nation's Social Problems?

Virginia Knox

Carolyn J. Hill

Gordon Berlin

Virginia Knox, MDRC, 1990 M Street NW, Suite 340, Washington DC, 20009

virginia.knox@mdrc.org

carolyn.hill@mdrc.org

gordon.berlin@mdrc.org

Virginia Knox is Director of the Families and Children policy area at MDRC. She has led numerous evaluations that provide information on both impacts and potential program improvements, particularly interventions that build parents' abilities to support their children's healthy development.

Carolyn J. Hill is a senior fellow at MDRC. Her research interests are in implementation, public management, performance measurement, and program evaluation. She is the coauthor of *Public Management: Thinking and Acting in Three Dimensions*.

Gordon Berlin is president of MDRC, a nonprofit, nonpartisan research organization dedicated to learning what works to improve the lives of low-income people. Berlin has authored numerous publications, including *Rewarding the Work of Individuals: A Counterintuitive Approach to Reducing Poverty and Strengthening Families*.

Key words: Evidence, implementation, adaptation, evaluation, counterfactual, service contrast

Note. We acknowledge our MDRC colleagues, especially Rekha Balu, William Corrin, and John Martinez. Our discussions and collaborations with them have contributed substantially to the ideas in this chapter.

This article will appear in Volume 678 of *The ANNALS of the American Academy of Political and Social Science*, edited by Ron Haskins, forthcoming July 2018.

Abstract

This chapter updates the pipeline paradigm for evidence building with a cyclical paradigm that encompasses evidence building, implementation, and adaptation. A cyclical paradigm for evidence-based policy and practice assumes that an intervention will be adapted over time, across settings, and across populations. These innovations and adaptations are encouraged and tested, with periodic review of the need for impact assessment. The cyclical paradigm also emphasizes service contrast at every stage, not just in the evidence-building stage where it has traditionally been a focus. A continuous cycle of evidence building, implementation, and adaptation — looping back to further evidence building — can help ensure that the impacts of evidence-based policies and programs are sustained and grow in new settings.

Introduction

In recent decades, policymakers have grown more likely to use research evidence to guide their attempts to meet social and educational goals. Researchers have identified many reforms worthy of broad expansion, including changes in the welfare system and programs that help low-income parents foster their children’s early development. Yet despite these successes, on the whole it remains hard to implement large-scale interventions supported by evidence. Even when interventions are grounded in current knowledge and show positive effects in early tests, those effects are often modest, and they often aren’t repeated when the programs expand to other settings.

Generally speaking, the prevailing paradigm for building evidence about programs or policies is a linear one (see the pipeline model shown in figure 1): A new intervention is developed in response to a social problem. It undergoes early tests. If those tests find positive results relative to what participants would have otherwise experienced then additional impact studies are conducted in new locations (replication). If the replication studies are positive, funders will support further expansion (scaling up), expecting to see similar effects as long as future versions implement the core elements of the intervention faithfully (see, for example, Institute of Education Sciences 2016). This pipeline paradigm is sometimes accompanied by a tiered funding model, in which more funding is made available to expand a program to a larger scale as it generates more — and more rigorous — evidence of its continued effects in more locations.

Figure 1: The Prevailing Evidence-Building Pipeline Paradigm



In this chapter, we suggest an updated paradigm. Compared with the pipeline model, this framework better reflects how social programs develop across locations and over time, and thus could more effectively address social problems. It moves evidence-based policy forward by explicitly acknowledging the complementary nature of rigorous research and implementation. In a cycle of continuous learning, implementation decisions are guided by evidence as much as possible, and adaptations of evidence-based models are used to generate further evidence about a program's outcomes. It brings together three processes: building rigorous evidence produced from a pipeline paradigm; implementing policies and programs that are grounded in evidence about how, for whom, and where an intervention can be effective; and learning from adaptation in a systematic way. The framework is cyclical, so the need for rigorous evidence building is continually assessed. And it emphasizes service contrast at every stage of the process instead of only during the impact evidence-building stage.

This updated paradigm can be used by funders, researchers, and practitioners who want to use evaluation to strengthen programs and their impacts. Ideally, the paradigm can encourage conversations among these different contributors to the improvement of the programs that are, after all, the foundational content of evidence-based policy.

In this chapter, as in the rest of this volume, the term *rigorous evidence* refers to evidence of impacts based on research designs that allow one to conclude that a given social intervention

caused the effects observed. As discussed earlier in this volume, such studies primarily use high-quality experimental or quasi-experimental designs. Experimental designs (also called randomized controlled trials) compare the outcomes of a group of people assigned at random to receive an intervention with the outcomes of a control group whose members may often receive other services available in the community but do not receive the intervention being studied. Because they were assigned at random, the two groups' measurable and unmeasurable characteristics should be similar at the outset of the study, so researchers can be highly confident that any differences in the outcomes of the two groups are caused by the program group's participation in the intervention.

Other studies or program improvement activities that evaluate program effectiveness use methods that may incorrectly attribute improvements in participants' outcomes to their participation in the intervention (for example, designs that measure the outcomes of a group before and after it receives an intervention, or designs that compare the outcomes of program participants with the outcomes of a group of people who are similar but who were not chosen randomly). However, applying some of these methods, or "design elements," in combination can result in more rigorous quasi-experimental designs; and under certain conditions, some quasi-experimental designs can approach the rigor of a randomized controlled trial (Corrin and Cook 1998; Shadish, Cook, and Campbell 2002).

Successes and Limitations of the Pipeline Paradigm

The evidence-based policy movement is capitalizing on an unusual window of opportunity. Politicians across the ideological spectrum increasingly support using evidence to make decisions about funding, ending, or improving programs (Berlin 2016). Rigorous empirical studies give policymakers confidence that social programs can improve people's lives and at the same time give them confidence that programs can be held accountable for doing so. In recent years, early successes led to efforts by the executive branch (principally in the Office of Management and Budget) to demand and reward the production and use of evidence in policy and budgeting. They have also led to efforts by the legislative branch to write new laws with tiered evidence structures. These laws encourage states or other service providers to implement evidence-based models, while also supporting innovation by linking increasing funding to stronger evidence bases (Institute of Education Sciences 2016; Gueron and Rolston 2013; Haskins and Margolis 2014).

Research has identified many effective approaches to social problems such as poverty or poor health — and ineffective approaches as well. Early large-scale social policy experiments such as the RAND Health Insurance Experiment (Brook et al. 1984), the Abt Housing Allowance Demand Experiment (Merrill and Joseph 1980), and the Negative Income Tax Experiment (Munnell 1986), were not set up simply to test models with the idea of replicating them. Instead, these studies tested theories and provided empirical evidence of parameters underlying the design of health insurance programs, housing subsidies, and income-support concepts. The National Supported Work Demonstration (Manpower Demonstration Research Corporation 1980) and the later Welfare-to-Work experiments of the 1990s (Greenberg, Deitch, and Hamilton 2009) showed policymakers that they could generate clear lessons for policy and program design

by testing variations on a given program model over a number of years and in real-world, diverse settings (Gueron and Rolston 2013). Since then, the pace of policy-related evidence building has quickened. The past decade has seen several major federal initiatives to build evidence (including some that are reviewed in this volume), and many more have been supported by states and other funders (Berlin 2016).

Innovative policy strategies have demonstrated sizable effects that could greatly improve important social outcomes if they could be achieved on a large scale. For example, Make Work Pay strategies produced increased income and employment among low-income parents (Bloom and Michalopoulos 2001); Small Schools of Choice improved graduation rates in urban high schools (Unterman 2014); Career Academies improved post-graduation earnings (Kemple 2008); and the City University of New York Accelerated Study in Associate Programs (CUNY ASAP) substantially increased the graduation rates of low-income community college students through well-structured academic, financial, and personal support (Scrivener et al. 2015).

Of course, when subjected to rigorous tests, many promising ideas and new programs do not show results (Coalition for Evidence-Based Policy 2013). The best-known example may be the evaluation of 16 Job Training Partnership Act programs (Orr et al. 1996). The experimental evaluation showed no improvements in earnings for young people ages 16 to 22 who participated in employment and training programs funded by the U.S. Department of Labor. As a result, the Clinton administration proposed and Congress agreed to a substantial shift away from funding for employment and training programs for young people (Burtless and Greenberg 2005).

The evidence pipeline has thus identified ideas that work well, which ones need some adjustment to be successful, and which ones need to go back to the drawing board. Ideally, impact evidence about a policy or program keeps policymakers and program administrators from

continually relearning old lessons about what works and what does not, freeing them to focus on improving programs and policies.

Yet the evidence pipeline paradigm has not identified many programs or their essential elements that have population-level effects (Supplee and Metz 2015). Even social interventions that are shown to be effective in early tests tend to have smaller effects (or, worse, no measureable effect) when conducted in a wider range of real-world settings (Bloom 2016). In the next section, we discuss how the research and practice communities have tried to overcome these shortcomings.

Attempts to Do More Within the Pipeline Paradigm

On the whole, social policy researchers have focused more on building evidence about individual program models than they have on building evidence about effective strategies for replicating or expanding programs. Similarly, tiered funding systems support the expansion of an effective program while also encouraging further evaluation to test whether the expanded model still produces effects. The underlying assumption is that a policy or program that works in one location will continue to work when it is replicated elsewhere. The process often ends in disappointment, in part because that assumption is naïve. It ignores or underestimates the challenges of replicating a program at a larger scale.

It's Complicated

Why do interventions that are effective on a small scale often produce more modest effects when they expand? Research methods aside, interventions are typically embedded in organizations or systems, so the approaches being tested are buffeted by external influences as

they are adopted and implemented (McCarthy 2014; Supplee and Metz 2015). Moreover, impacts are usually measured by comparing a group that receives an intervention with another group that does not. Even if the intervention is implemented well in another context, it may not bring greater benefits to participants than the programs already available.

Further, it is complicated to address social problems — to try to achieve aims such as “closing the educational achievement gap” or “helping displaced adult workers gain a foothold in a rapidly changing economy.” The ideal is that if educators or local program staff members are trained in new practices and given help to carry them out, and if program participants respond to these new practices, then educational, social, or economic outcomes will improve. But each step in this process involves numerous influences and interactions, many perhaps unanticipated.

What Works? How? For Whom? Under What Conditions?

To gain insight into the steps and mechanisms involved in addressing social problems, some impact studies are accompanied by *implementation analyses* that use both quantitative and qualitative information. These analyses aim to document the contexts in which the programs were implemented, to convey some of the factors that future program operators should consider, to predict whether the intervention is likely to have similar effects in new contexts, and to strengthen implementation. Within the pipeline paradigm, implementation research at each stage of the process can inform future expansion efforts (Manno and Miller-Gaubert 2016). Though they vary in emphasis and comprehensiveness, implementation analyses often focus on describing “what,” “who,” and “where”: what services were provided, who participated, and where the intervention took place.

Researchers also want to understand *how* a program achieves its effects, so they can guide future practitioners who want to implement a similar model. More broadly, learning about the effects of particular program practices allows other practitioners to apply those practices in different settings (Weisz et al. 2015; Bates and Glennerster 2017). For example, program approaches guided by behavioral science and motivational interviewing are being used in multiple fields to encourage clients to use services and to benefit from them (Higgins 2015; Richburg-Hayes, Anzelone, and Dechausay 2017). Strong implementation research and informative measures about the services available to the control group also provide insight into how a program achieves its effects (Hamilton and Scrivener, in progress).

Impact studies may examine *for whom* a program works, so that future practitioners can recruit participants among populations that might benefit most. Early studies of a model often try to answer this question by examining differences in effects among subgroups, by testing variations of the program model, or by describing program implementation. It can be hard to draw strong conclusions from such analyses, though, because subgroup sample sizes may be small, and because implementation settings may not vary much (meaning there is not much meaningful subgroup variation to analyze). Though meta-analyses (systematic, quantitative analyses of effects from multiple studies) can help with these issues, they may also mask important differences in individual studies' weaknesses or designs and can have methodological problems of their own.

If they are designed explicitly to learn from variation, studies conducted at multiple sites can be especially helpful at shedding light on what works for whom, and under what conditions (Michalopoulos et al. 2015; Peck et al. 2014; Miller et al. 2016). Bloom, Hill, and Riccio (2003) illustrate the kinds of analyses that are possible with foresight about implementation data

collection for both the group that receives an intervention and the control group that does not. However, multisite studies that collect rich implementation data can be expensive and difficult to conduct and may not be feasible in many circumstances.

Continuing to Learn After Evidence-Based Practices Expand

The evidence pipeline encourages large-scale studies of practices that have demonstrated effectiveness on a small scale and have then expanded widely, so that funders can learn about their effects in broader contexts. Examples of such large-scale studies include the National Job Training Partnership Act Study (Doolittle et al. 1993), the National Evaluation of Welfare-to-Work Strategies (Hamilton 2002), the Early Head Start Research and Evaluation Study (Administration for Children and Families 2002), the Head Start Impact Study (Administration for Children and Families 2010), and the Mother and Infant Home Visiting Program Evaluation (Michalopoulos et al. 2015). But these kinds of national evaluations are relatively rare. As a result, we do not have very much information about how evidence-based practices can maintain or even increase their effects as they are widely implemented.

Initiatives under the umbrella of “implementation science” have focused on how evidence-based policy interventions are disseminated, implemented, and adapted over time (Fixsen et al. 2005; Supplee and Metz 2015; Chambers, Glasgow, and Stange 2013). Frameworks such as Getting to Outcomes (RAND Health 2017) are designed to guide a local organization as it adopts, adapts, and implements an evidence-based intervention. On the whole, however, systematic, ongoing learning about evidence-based practices is relatively rare.

An Updated Paradigm:

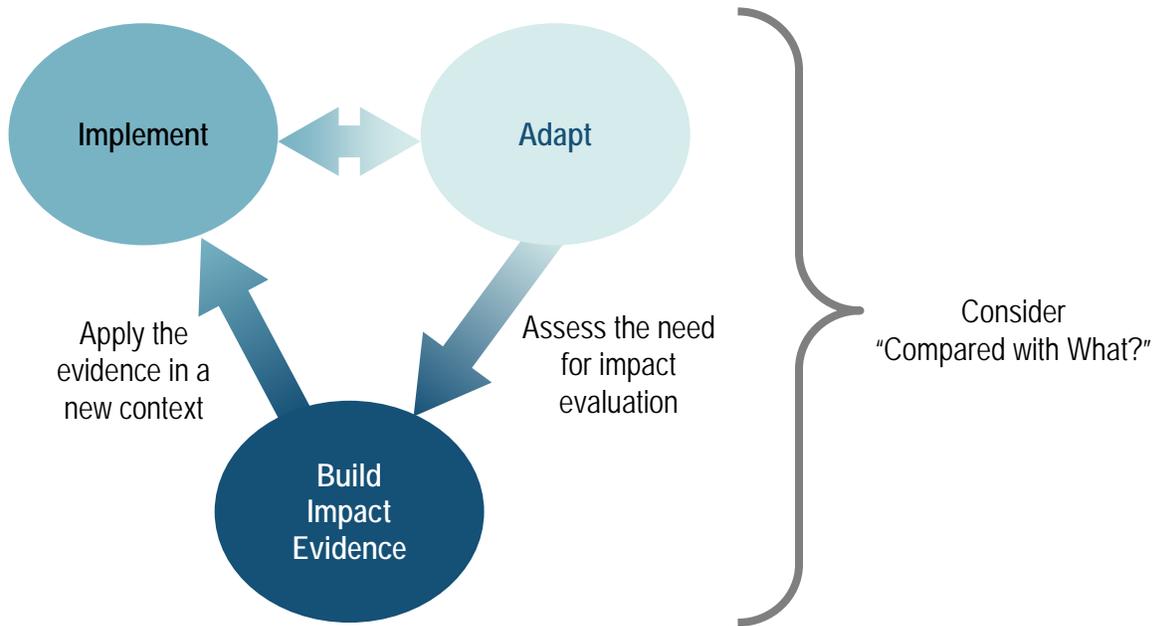
The Evidence-Building, Implementation, and Adaptation Cycle

If we take seriously the goal of using evidence-based policy to ameliorate social problems, evidence-based policymaking needs a paradigm shift. An updated paradigm, described below, reflects how connections between research and practice can potentially improve an intervention over time and across settings. It requires stretching our concept of evidence-based policy beyond individual program packages to envision robust systems that can evolve and improve over time (Coburn 2003). It requires reimagining rigorous evidence building and program innovation as part of a continuous cycle that sustains and increases effects. As shown in figure 2, this cycle involves:

- (1) building impact evidence about innovations and new approaches to social programs;
- (2) implementing evidence-based programs and supporting the expansion of these programs in ways that recognize the complexity inherent in replication, to maintain or increase their effectiveness as they spread;
- (3) encouraging adaptations and improvements; and
- (4) assessing the need for further impact evidence.

In addition to being informed by evidence-based policies and programs, implementation and adaptation in the field are informed by a theory of change that links the program's activities to its desired outcomes, as well as by performance measurement of indicators that are consistent with the theory of change. Periodically, assessments are made regarding the need for updated impact evidence.

Figure 2: An Updated Paradigm for Evidence Building, Implementation, and Adaptation



The updated paradigm can serve both funders (who benefit from the accountability provided by impact studies in the evidence-building stage) and organizations that run programs (who benefit from evidence that can help them improve). This cyclic, iterative paradigm promotes innovation in a changing environment, and continues to build credible evidence about whether and how innovations bring down costs or have greater effects than the status quo.

As shown in figure 2 and discussed further below, the cycle builds on the pipeline paradigm but departs from it in two important ways. First, it emphasizes *service contrast* at every stage, not just in the evidence-building stage with impact studies. Second, the implementation of an evidence-based program is not the end point, as it is in the pipeline model. Instead, the updated paradigm assumes the intervention will be adapted over time and across settings, to sustain or increase impacts under changing circumstances. The cycle encourages programs to

innovate and test new approaches — and to learn from adaptations, regularly assessing the need for impact evaluation.

Compared with What?

A central component of impact studies is service contrast, which refers to the differences in content, quality, quantity, and modality of services experienced by participants in the tested intervention, compared with the services experienced by members of a control group (the counterfactual or “business as usual”; Weiss, Bloom, and Brock 2013; also see Hamilton and Scrivener in progress). Researchers in impact studies will typically document the services received by control group members.

Yet often service contrast is downplayed when policymakers or others interpret the findings from a study about an intervention. When the intervention is expanded or adapted and then tested again, a meaningful service contrast may no longer be apparent. For example, if the intervention expands into a setting where participants have access to substantially similar services, we may see no difference between the experiences of program and control group members, leading to a finding of “no effect” for the intervention.

Continuing attention to service contrast during the implementation and adaptation of an intervention — even if service contrast cannot be precisely measured — can help to identify settings that might be appropriate for the intervention, to isolate the intervention’s unique features, and to focus implementing organizations on what may be the intervention’s most important elements for improving outcomes.

Strengthening Impacts as Programs Expand to New Settings

In the linear pipeline paradigm depicted in figure 1, the goal is to implement evidence-based policies and programs on a large scale. To realize the full potential of evidence-based policy for improving social problems, however, we need to link evidence-based programs with new settings systematically, and we need to incorporate and test adaptations.

Implement an evidence-based program. First, to increase the likelihood that a program's previously measured impacts will translate into a new setting, potential implementers should consider the service contrast that existed when the program was found to be effective in the evidence-building phase.¹ For example, as the first step in planning to implement evidence-based early childhood home visiting programs, the federal home visiting program (Maternal, Infant, and Early Childhood Home Visiting) required states to scan their existing services and their populations at risk of poor child development who might benefit from home visiting. This step is important because operating an evidence-based program with fidelity is not the only factor that determines whether its impacts in a new setting will be comparable to those it achieved before. Another important factor is whether alternatives are available. An evidence-based program can probably benefit participants the most in places without existing, similar services. In such settings, even local programs without high fidelity to a program model potentially can produce effects (Corrin and Martinez 2017). By making information about service contrast from their impact studies more accessible, researchers can help organizations that implement the programs involved. Examples of clear descriptions of service contrast include

¹ A few implementation science frameworks emphasize this kind of fit. For example, see RAND Health (2017).

the evaluation of CUNY ASAP (Scrivener et al. 2015) and of Diplomas Now (Sepanik et al. 2015)

Second, innovative program approaches are typically designed with the intention of benefiting a target group. Evidence-based practices are appropriate when applied to the populations they were tested with (a principle acknowledged by many frameworks for program expansion). Understanding those populations means understanding both who was eligible for the intervention and who was actually served.

Third, new services are embedded in local contexts that can affect how the innovations are implemented and how participants respond. Factors that can influence implementation and program effects include the implementing organization's capabilities, culture, or other characteristics; the community's social, political, and economic characteristics; and the funding available for the program, the laws and regulations governing it, professional networks, other interest groups, and so on. Just as an implementing agency must understand whether its intended target population differs from the population served in the intervention's prior impact studies, it also needs to think through how the operational context differs from earlier settings.

Adapt practices, informed by the program's theory of change and by measures of performance. Organizations that implement evidence-based models may focus on achieving fidelity to a program model, on aligning their activities with a theory of change, or on achieving certain levels of participation and service delivery. Their approaches to improving short-term organizational performance sometimes include "rapid-cycle" tests or "plan-do-study-act" cycles. They may involve rigorous tests of new implementation strategies (Manzi 2012). Performance-improvement systems, however, typically focus on helping organizations measure and reach milestones such as fidelity to specific aspects of the evidence-based model, or certain levels of

high-priority outcomes. These measures may provide some information about program operations, but they seldom reflect service contrast, which can be difficult to assess regularly.²

It is unrealistic to think that evidence-based practices will be implemented on a large scale and over time without being adapted to local circumstances and in response to information from performance-measurement systems. Instead of assuming that fidelity to an evidence-based model is the ultimate goal, many researchers are seeking to understand how adaptations can be developed and measured systematically (Chambers, Glasgow, and Stange 2013; Cannata and Rutledge forthcoming; Center on the Developing Child at Harvard University 2016).

Adaptations may be evident through components of the program model, through who is served, or through context-driven aspects. First, the *components* of a program model may be deliberately adapted. Fidelity to an evidence-based model can help staff members to implement a program in ways intended by the program developer or the funder, and maintaining fidelity saves programs from reinventing the wheel. But, for example, an agency's leaders may theorize that strict fidelity will be less effective than some carefully structured modifications that widen the service contrast. There is a fundamental tension between such adaptations and the assurances of accountability that evidence-based models provided in the first place. If it is possible to resolve those tensions, well-conceived adaptations may expand program impacts beyond those of the original model.

Programs may also adapt *who* is served. Programs sometimes face pressure to extend interventions to other groups who might benefit, but for whom the program is untested. To do so, program operators need to consider how this new population is different from the population

² Gugerty and Karlan (forthcoming) provide guidance on performance measurement for organizations at this stage, when programs are not in a position to be evaluated with a rigorous evaluation design. They also emphasize the importance of a theory of change in program implementation and performance measurement.

already studied. What is the most compelling explanation for the benefits the original target population experienced, given the service contrast they experienced? Does the same reasoning apply to the new target population? If there is a good rationale for believing that a modified population would benefit as well, there could be ways to adapt the program for that new target population while building evidence about the program's effects for them (Jacob et al. 2014).

Finally, programs may be adapted to new *contexts*. A program's context — aspects of the organization, its surrounding system, the community, and the legal and political landscapes — will change over time. The question is how these changes influence the program's ability to deliver benefits to its participants, and how the program responds (Chambers, Glasgow, and Stange 2013).

Assess the need and readiness for impact evaluation. Though adaptations can be beneficial, organizations implementing evidence-based practices should also be aware of the risks of adapting practices without a strong rationale or applying them to populations or in contexts where they are unlikely to be effective. Administrators are less likely to expose themselves to such risks if they periodically reassess their services, clients, and contexts — that is, what, who, and where — while asking, “Compared with what?” At some point, an evidence-based model may reflect substantial modification because of new technologies, new constraints, new opportunities, or lessons learned from previous implementations. Such changes should be used as opportunities for learning (Balu 2017; Chambers and Norton 2016; Durlak and DePre 2008). Organizations and funders could agree on guidelines for assessing whether updated impact evidence is needed.

Conclusion

This chapter describes an updated paradigm for evidence-based policy that views program expansion as a means to an end — alleviating social problems — rather than as an end in itself. As Sabelli and Harris (2015, p. 14) put it in the context of education: “the ultimate goal of scaling up is sustainable educational improvement rather than to merely expand the use of a given educational innovation.” Under a cyclical paradigm of evidence building, implementation, and adaptation, researchers and funders support evidence that is better attuned to the real-world needs of people who will implement and adapt evidence-based programs. Moreover, implementers systematically consider how, for whom, and where they operate programs, to decide what adaptations they should make, and to continue building evidence about the effectiveness of those adaptations.

In medicine, advances in treating illnesses lead to changes in standards of care. Similarly, in social and educational policy, achieving effective and efficient results requires both openness to innovation and a willingness to subject new approaches to rigorous testing. To that end, evidence-based policy has the most potential to ameliorate social problems if it is viewed as a cycle, one that helps public agencies and service systems work more effectively over time.

References

- Administration for Children and Families (ACF). 2002. *Making a difference in the lives of children and families: The impacts of Early Head Start programs on infants and toddlers and their families*. Washington, DC: U.S. Department of Health and Human Services.
- Administration for Children and Families (ACF). 2010. *Head Start impact study final report*. Washington, DC: U.S. Department of Health and Human Services.
- Balu, Rekha. 2017. *Anticipating variation in implementation: The multistate evaluation of Response to Intervention practices*. New York, NY: MDRC.
- Bates, Mary Ann, and Rachel Glennerster. 2017. The Generalizability Puzzle. *Stanford Social Innovation Review* 15 (3).
- Berlin, Gordon. 2016. Using evidence as the driver of policy change: The next steps in supporting innovation, continuous improvement, and accountability. Testimony before the Senate Finance Committee, Washington, DC, May 10.
- Bloom, Dan. 2016. Should the government subsidize jobs for the unemployed? *Spotlight on Poverty and Opportunity*. September 7. <https://spotlightonpoverty.org/spotlight-exclusives/government-subsidize-jobs-unemployed>.
- Bloom, Dan, and Charles Michalopoulos. 2001. *How welfare and work policies affect employment and income*. New York, NY. MDRC.
- Bloom, Howard S., Carolyn J. Hill, and James A. Riccio. 2003. Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management* 22 (4): 551-575.
- Brook, Robert H., John E. Ware, William H. Rogers, Emmett B. Keeler, Allyson Ross Davies, Cathy Donald Sherbourne, George A. Goldberg, Kathleen N. Lohr, Patti Camp, and

- Joseph P. Newhouse. 1984. *The effect of coinsurance on the health of adults: Results from the RAND Health Insurance Experiment*. Santa Monica, CA: RAND Corporation, R-3055-HHS.
- Burtless, Gary, and David Greenberg. 2005. *Use of experimental methods in workforce evaluations*. Occasional Paper 2005-08. Washington, DC: U.S. Department of Labor, Employment and Training Administration.
- Cannata, Marissa, and Stacey Rutledge. Forthcoming. New frontiers in scaling up research. *Peabody Journal of Education*.
- Center on the Developing Child at Harvard University. 2016. From Best Practices to Breakthrough Impacts: A Science-Based Approach to Building a More Promising Future for Young Children and Families. www.developingchild.harvard.edu.
- Chambers, David A., Russell E. Glasgow, and Kurt C. Stange. 2013. The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. *Implementation Science* 8 (117).
- Chambers, David A., and Wynne E. Norton. 2016. The adaptome: Advancing the science of intervention adaptation. *American Journal of Preventive Medicine* 51 (4): 124-131.
- Coalition for Evidence-Based Policy. 2013. Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects. <http://coalition4evidence.org/wp-content/uploads/2013/06/IES-Commissioned-RCTs-positive-vs-weak-or-null-findings-7-2013.pdf>.
- Coburn, Cynthia. 2003. Rethinking scale: moving beyond numbers to deep and lasting change. *Educational Researcher* 32 (6): 3-12.

- Corrin, William J., and Thomas D. Cook. 1998. "Design elements of quasi-experiments." In A.J. Reynolds and H.J. Walberg (eds.), *Advances in Educational Productivity*.
- Corrin, William J., and John Martinez. 2017. "Social Progress's True Metric: Compared to What?" *Stanford Social Innovation Review*, April 28.
https://ssir.org/articles/entry/social_progresss_true_metric_compared_to_what#
- Doolittle, Fred, Steve Bell, Howard Bloom, George Cave, James Kemple, Larry Orr, Linda Traeger, and John Wallace. 1993. *A summary of the design and implementation of the National JTPA Study*. New York, NY: Manpower Demonstration Research Corporation.
- Durlak, Joseph A., and Emily P. DuPre. 2008. Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology* 41 (23).
- Fixsen, Dean, Sandra Naoom, Karen Blasé, Robert Friedman, and Frances Wallace. 2005. *Implementation research: A synthesis of the literature*. Chapel Hill, NC: National Implementation Research Network.
- Greenberg, David, Victoria Deitch, and Gayle Hamilton. 2009. *Welfare-to-Work program benefits and costs: A synthesis of research*. New York, NY: MDRC.
- Gueron, Judith M., and Howard Rolston. 2013. *Fighting for reliable evidence*. New York, NY: Russell Sage Foundation.
- Gugerty, Mary Kay, and Dean Karlan. Forthcoming. *The Goldilocks challenge: Right-fit evidence for the social sector*. Oxford: Oxford University Press.
- Hamilton, Gayle. 2002. *Moving people from welfare to work: Lessons from the National Evaluation of Welfare-to-Work Strategies*. New York, NY: Manpower Demonstration Research Corporation.

- Hamilton, Gayle, and Sue Scrivener. In progress. "Measuring treatment contrast." MDRC working paper.
- Haskins, Ron, and Greg Margolis. 2014. *Show me the evidence: Obama's fight for rigor and results in social policy*. Washington, DC: Brookings Institution Press.
- Higgins, Margaret M. 2015. *Application of motivational interviewing techniques in child welfare practice*. Minneapolis, MN: Center for Advanced Studies in Child Welfare, University of Minnesota.
- Institute of Education Sciences (IES). 2016. *Non-regulatory guidance: Using evidence to strengthen education investment*. Washington, DC: U.S. Department of Education.
- Jacob, Robin Tepper, Thomas J. Smith, Jacklyn A. Willard, and Rachel E. Rifkin. 2014. *Reading Partners: The implementation and effectiveness of a one-on-one tutoring program delivered by community volunteers*. New York, NY: MDRC.
- Kemple, James J. 2008. *Career Academies: Long-term impacts on work, education, and transitions to adulthood*. New York, NY: MDRC.
- Manno, Michelle, and Jennifer Miller-Gaubert. 2016. A conceptual framework for supporting scale-up through implementation research. Internal guidance document, MDRC.
- Manpower Demonstration Research Corporation (MDRC). 1980. *Summary and findings of the National Supported Work Demonstration*. New York, NY: Manpower Demonstration Research Corporation.
- Manzi, Jim. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. Basic Books.
- McCarthy, Patrick T. 2014. The road to scale runs through public systems. *Stanford Social Innovation Review*. 12 (2).

- Merrill, Sally R., and Catherine A. Joseph. 1980. *Housing Allowance Demand Experiment: Housing improvements and upgrading in the Housing Allowance Demand Experiment*. Cambridge, MA: Abt Associates Inc.
- Michalopoulos, Charles, Helen Lee, Anne Duggan, Erika Lundquist, Ada Tso, Sarah Shea Crowne, Lori Burrell, Jennifer Somers, Jill H. Filene, and Virginia Knox. 2015. *The Mother and Infant Home Visiting Program Evaluation: Early findings on the Maternal, Infant, and Early Childhood Home Visiting Program: A report to Congress*. New York, NY: MDRC.
- Miller, Cynthia, Megan Millenky, Lisa Schwartz, Lisbeth Goble, and Jillian Stein. 2016. *Building a future: Interim impact findings from the YouthBuild evaluation*. New York, NY: MDRC.
- Munnell, Alicia H., ed. 1986. *Lessons from the Income Maintenance Experiment: Proceedings of a conference held in September 1986*. Boston, MA: Federal Reserve Bank of Boston.
- Orr, Larry L. Howard Bloom, Stephen Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. *Does training for the disadvantaged work? Evidence from the national JTPA study*. Washington, DC: Urban Institute.
- Peck, Laura R., Alan Werner, Alyssa Rulf Fountain, Jennifer Lewis Buell, Stephen H. Bell, Eleanor Harvill, Hiren Nisar, David Judkins, Gretchen Locke. 2014. *Health Profession Opportunity Grants (HPOG) impact study design report*. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
- RAND Health. 2017. "Getting to Outcomes: Improving Community-Based Prevention." Website: www.rand.org/health/projects/getting-to-outcomes.html.

- Richburg-Hayes, Lashawn, Caitlin Anzelone, and Nadine Dechausay. 2017. *Nudging change in human services: Final report of the Behavioral Interventions to Advance Self-Sufficiency (BIAS) project*. New York, NY: MDRC.
- Sabelli, Nora H., and Christopher J. Harris. 2015. The role of innovation in scaling up education innovations. In Chee-Kit Looi and Laik Woon Teh (eds.), *Scaling Education Innovations*, Springer Singapore.
- Scrivener, Susan, Michael J. Weiss, Alyssa Ratledge, Timothy Rudd, Colleen Sommo, and Hannah Fresques. 2015. *Doubling graduation rates: Three-year effects of CUNY's Accelerated Study in Associate Programs (ASAP) for developmental education students*. New York, NY: MDRC.
- Sepanik, Susan, William Corrin, David Roy, Aracelis Gray, Felix Fernandez, Ashley Briggs, and Kathleen K. Wang. 2015. *Moving down the track: Changing school practices during the second year of Diplomas Now*. New York, NY: MDRC.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage Learning.
- Supplee, Lauren H., and Allison Metz. 2015. Opportunities and challenges in evidence-based social policy. *Social Policy Report* 28 (4).
- Unterman, Rebecca. 2014. *Headed to college: The effects of New York City's small high schools of choice on postsecondary enrollment*. New York, NY: MDRC.
- Weiss, Michael J., Howard Bloom, and Thomas Brock. 2013. *A conceptual framework for studying the sources of variation in program effects*. New York, NY: MDRC.

Weisz, John R., Lauren S. Krumholz, Lauren Santucci, Kristen Thomassin, and Mei Ng. 2015. Shrinking the gap between research and practice: tailoring and testing youth psychotherapies in clinical care contexts. *American Review of Clinical Psychology* 11: 139-163.