Contemporary Educational Research Quarterly
December, 2016, Vol. 24, No. 4, pp. 071-090
DOI : 10.6151/CERQ.2016.2404.03

《Scholarly Commentary》

PROGRAM EVALUATION POLICY, PRACTICE, AND THE USE OF RESULTS

Robert Boruch* Jessica L. Chao Selene S. Lee

ABSTRACT

This scholarly commentary addresses the basic questions that underlie program evaluation policy and practice in education, as well as the conditions that must be met for the evaluation evidence to be used. The evaluation questions concern evidence on the nature and severity of problems, the programs deployed to address the issues, the programs' relative effects and cost-effectiveness, and the accumulation of evidence. The basic conditions for the use of evidence include potential users' awareness of the evidence, their understanding of it, as well as their capacity and incentives for its use. Examples are drawn from studies conducted in the United States and other countries, focusing on evaluation methods that address the questions above.

Keywords: *evaluation policy and practice, use of evidence, randomized controlled trials, quasi-experiments, meta-analysis, education*

* Robert Boruch (corresponding author), University Trustee Chair Professor, University of Pennsylvania, Philadelphia, U.S.A.
E-mail: robertb@upenn.edu

Jessica L. Chao, Senior Research Assistant, University of Pennsylvania, Philadelphia, U.S.A.
E-mail: jec@gse.upenn.edu

Selene S. Lee, Research Assistant, Graduate School of Education, University of Pennsylvania, Philadelphia, U.S.A.
E-mail: selenel@gse.upenn.edu

Manuscript received: May 30, 2016; Modified: September 5, 2016; Accepted: October 3, 2016

Introduction

This commentary is divided into two major sections. The first concerns basic questions that underlie program evaluation policy and practice in different countries, as well as evaluative evidence that address the questions. The questions bear on the nature and severity of problems, the deployment of programs to address the problems, efforts to estimate the effects and cost-effectiveness of programs, and the accumulation of scientific evidence on what works and what does not. The second major section concerns the basic questions that help us understand how to enhance the use of evaluation evidence. Both parts include illustrations based on reports that are publicly-accessible. The emphasis is on education-related program evaluations.

Basic Themes in Program Evaluation

The word "evaluation" is common in the social, behavioral, health, and education sciences. The meanings of the word will vary across these sectors. In what follows, the focus is on program evaluation rather than other areas such as personnel evaluation, teacher evaluation, and policy analysis. In particular, the aim is to make plain the meanings by framing the simple themes that underlie program evaluation regardless of academic discipline, government agency, or country. The themes are put into interrogatory form in what follows:

1. What is the nature of the problem or issue to which attention is directed, and what is the evidence on the problem?
2. How and how well is the intervention deployed to address the problem, and what is the evidence?
3. Does the intervention work, which intervention works better, and what is the evidence?
4. How cost-effective are the interventions, and what is the evidence?
5. How might one accumulate dependable evidence from an assembly of evaluations?

In this context, "interventions" may include *practices* in education and related services. They may include *programs* designed to provide better or more specialized services to individuals, organizations, or geopolitical jurisdictions. At the broadest level, interventions may be construed as macro-level *policy*.

A rationale for this Socratic and interrogatory approach is that putting the questions plainly is important when dealing with differing languages and academic or government vernacular. Rossi, Lipsey, and Freeman (2004) pose the same questions slightly differently. The historical roots of evaluation work on each question are covered by articles in Alkin's (2012) edited volume.

The Nature of the Problem or Issue

In addressing the first class of questions on the nature and severity of the problem, evaluators may depend on probability sample surveys or administrative records of people or organizations in a target population at risk. They may depend on administrative ethnographic studies and focus groups or other street-level research in exploratory research.

Such resources are routinely used in health-oriented work, for example, to estimate the incidence and prevalence of injuries. In addition, administrative records on academic performance and international sample-based assessments are used by evaluators in education to understand the relative status of students locally and nationally.

In some of these studies, the correlations between children's academic achievement and their emotional, social, or economic well-being are often of interest. Understanding the levels of needs and their correlates is antecedent to developing interventions addressing those needs. Having such data must usually precede the invention of interventions that are thought to resolve a problem. Also, the data must often precede the evaluation of those interventions.

Cross-sectional studies. Such evaluations are a snapshot in time and characterize the empirical nature of problematic issues. Sznitman, Reisel, and Romer (2011), for instance, have studied the relation between the emotional well-being of adolescents and their educational achievement. The work is based on public data from dependable cross-sectional surveys which were conducted in 23 developed European countries and 39 states in the United States. The simple statistical correlation between emotional well-being and educational achievement is very high at the country level, even when child poverty and economic indicators are taken into account.

Longitudinal surveys. Longitudinal studies involve tracking people or entities over time. For instance, Zeng et al. (2012) used publicly accessible data from the U.S. Early Childhood Longitudinal Study to learn how nearly 10,000 kindergarten children's behavior varied as a function of the child's age.

They focused on comparing young kindergartners to older ones. In particular, they focused on kindergarten children who were 7 months older than the young kindergartners. Their findings were that (a) the younger group exhibited appreciably more "internalizing problem behaviors" such as sadness, anxiety, low self-esteem, and loneliness, (b) elevated levels of these problems persisted through the fifth grade of their schooling, and (c) the rates of problem behavior for Black and Hispanic children were appreciably higher than for White children. The rates for Asian children were not appreciably different from White children.

Chen, Huang, Wang, and Chang (2012) repeatedly surveyed multiple sources – children and their peers, teachers, and others on over 1,000 Chinese children aged 9 to 12 years to understand the statistical relationships between the children's aggression, peer relations, and their adjustment over time. The implications of the work include the idea that effective strategies for assisting children with behavioral problems should consider both personal and social factors.

Focus groups. Lee, Fu, and Fleming (2006) convened six focus groups in Taiwan with women in prison who had been injecting heroin users so as to understand what the women might need to learn about their own risky behavior. The focus group results suggest that the women's misconceptions of risk were serious, that they distrusted assurances of confidentiality in HIV/AIDS testing, and that issues of stigma were important. Identifying these issues is important in understanding how one might then create educational and other programs to reduce the problems.

The foregoing examples address the first question underlying evaluations, leading to dependable evidence on the nature and magnitude of the problem. However, the descriptive information does not necessarily tell us directly how to solve the problem or reduce its severity. The next series of questions get at potential solutions and methods to estimate the effects of tested solutions.

The Deployment and Performance of Programs

The second family of questions described above often falls under the rubrics of *monitoring*, *implementation studies, process research,* and *formative evaluation.* At the World Bank, for instance, a relevant abbreviation is "M&E," where the "M" stands for monitoring and the "E" often stands for impact evaluation. Though the words or phrases differ across institutions, the general aim is to understand whether and how well a program is being delivered.

Typically, the evidence to answer these questions depends on performance indicators that permit one to judge the extent of the service, including outputs, such as the number of people served, and more importantly, indicators of the processes and quality of service. These kinds of studies may be done independent of any attempt to estimate the program's actual effects. In recent years, however, implementation indicators have often been combined with impact evidence in dose-response studies.

Administrative records. In education evaluation, administrative records are usually essential in understanding which teachers teach which classes, when they teach, which students attend classes, and so on. At times, the records are dependable, accessible, and informative. For instance, in the United States, there is enormous diversity among the states in the way that relevant records are generated, maintained, and made accessible for evaluations.

For instance, recent work based on such records showed that about 25% of science and mathematics teachers in public schools in Missouri's biggest cities in a given year no longer teach in the same class, school, or position in the following year (Bowdon & Boruch, 2014). Further, of the cohort of teachers in Ohio's public schools in the 2008-2009 academic year, only 47% were teaching in the same school and the same subject area five years later. The retention rate over five years in Ohio's five biggest cities is about 25% (Baker & Boruch, 2015). Evidence of this kind is being used to inform the design of multi-year interventions as well as to design experiments that aim to estimate the effects of the interventions.

Specialized surveys. In some cases, local administrative records may not be available or they may not be trustworthy. As a consequence, the evaluation evidence on the implementation of programs may be generated through the evaluators' independent observation of classroom behaviors or through surveys on the intervention's delivery. Qualitative studies are often used to generate hypotheses about the character of service from the points of view of service recipients or others.

Bruns, Filmer, and Patrinos (2011), for instance, summarized statistical evaluations that focused on accountability in low-income countries, a topic of major interest for the World Bank. In particular, they provided data on the time that teachers actually were present in class and the time that teachers spent on the tasks for which they were responsible. These data were compared with the official time that teachers were supposed to be present in class.

"Presence time," as one might expect, was appreciably lower than official time in some countries, such as Ghana. Time spent on teaching was also lower than the official time in all countries included in the study. The data informed subsequent discussions of various kinds of incentives for teaching, as well as the implications for accountability systems and the evaluation of any incentive's effectiveness.

Pan (2014) reported on the deployment of Taiwan's Professional Learning Communities (PLC) and drilled more deeply at the local levels. In particular, Pan mounted specialized surveys of teachers in 28 schools in the Taiwanese context. The work was undertaken so as to understand how the PLC practices are related to the observable dimensions of school capacity for change, teacher practices in the classrooms and their engagement in professional learning, and other factors. The vision of change, shared professional practices, and learning for change appear to be substantially important in explaining the PLC practices in this cross-sectional study.

The need to understand whether and how well particular interventions are deployed in different settings has led to the invention of peer-reviewed journals such as *Implementation Science* in the health sector (www.implementationscience.com). It has also led to the creation of specialized entities such as the Center for Implementing Technology in Education (CITEd). Academic journals such as *Educational Evaluation and Policy Analysis* and the *Journal of Research on Educational Effectiveness* often carry reports on the implementation of complex programs or projects. Patton's (2008) book is an informative resource on evaluations embedded in the development of programs.

Effects of Program Interventions

The third category of questions involves attempts to discern the relative effects of interventions. They invite attention to two broad categories of impact evaluation designs: randomized controlled trials and quasi-experiments.

Randomized trials. In randomized trials, individuals, organizations, or entire geopolitical jurisdictions are randomly assigned to one of several intervention programs. One of the interventions may be a control condition, i.e. the *status quo.* A major benefit of a randomized trial in education, medicine, or other sectors is that randomization ensures that there are no systematic differences in the groups at the outset of an impact evaluation. Put in other

words, there is no systematic difference between the groups so composed, and consequently, pre-existing group differences do not undermine or complicate causal inferences about the intervention's effects. Thus, the comparison of outcomes among the groups is fair. Well-run randomized trials generate a statistically unbiased effect of the interventions' relative effects and a legitimate statistical statement of one's confidence in the results.

An example of a randomized trial involves a study of the relative effects of single-sex schools and co-ed schools in Korea. Seoul, the capital city of South Korea, has a policy of using a lottery-based allocation system to assign students to a single-sex school or to a co-educational school. The lottery allocation is, in effect, a randomized trial. Park, Behrman, and Choi (2013) took advantage of this to estimate the relative effect of each kind of school on students' later scores on college entrance exams and on the percentage of students that went on to attend four-year colleges or junior colleges. The study found that single-sex schools in Seoul had a dependably positive effect on students' subsequent attendance at both kinds of institutions.

In Mexico, a randomized controlled trial was used to examine the effects of cash transfers meant to prevent school dropouts in Mexico. Mexico's Progresa program (now called Oportunidades) was preceded by statistical work and anthropological research on the nature and severity of the school dropout problem in poor rural villages. In the randomized trial, over 300 low-income villages were randomly assigned to conditional cash transfer support or to control conditions in order to examine whether the cash transfers were effective in reducing a chronically high rate of school dropout (Parker & Teruel, 2005). The cash transfers to mothers of children in the Mexican villages did indeed reduce problems of their children dropping out of school. Replications are under way in other countries including Zambia (American Institutes for Research, 2015).

Impact evaluations of the kinds just described, in which entire organizations or entities are randomly allocated to different interventions, have increased in frequency since the 1980s. They are called "cluster randomized trials" (CRTs) in the health and education sectors, "group randomized trials" in psychological research that focuses on families, and "place-randomized trials" in criminology. In education, for instance, students are naturally grouped into classrooms or schools, and these classrooms or schools are randomly assigned to intervention and control conditions, so as to understand whether the new intervention works any better than ordinary

practices. Rimm-Kaufmann et al. (2014) and Wijekumar et al. (2014) used the cluster randomized trial method in their research.

When randomized trials are not ethical or feasible, evidence on what intervention works may be generated through quasi-experiments or through approaches that depend on passive observational data, such as surveys. Sophisticated statistical models and econometric model-based approaches to estimating effects typically depend on more assumptions than a randomized trial does.

Quasi-experiments. A common quasi-experimental approach involves two groups that differ initially, one being assigned to the program under investigation and the second which is not afforded the program. The groups may differ appreciably at the outset of the evaluation. In this context, quasi-experimental and observational study approaches try to approximate the results of a randomized trial by constructing matched pairs of members from each group that differ initially, i.e. they try to construct sub-groups that are similar in all ways except for the treatment condition. The matching may depend on simple matching algorithms or they may be model-based. Model-based approaches include "propensity scores," "selection models," "structural models," and "instrumental variables." Propensity score matching, for instance, is a statistical technique that is intended to allow researchers to adjust for confounders by conditioning on a large number of observed covariates (characteristics of members of the groups). The aim is to mimic the results of random allocation in a randomized controlled trial.

The use of model-based approaches engenders important assumptions that are usually not needed in a randomized trial: (a) the right covariates have been identified, (b) they have been measured properly, and (c) they are incorporated into models whose functional form is adequate. Rosenbaum (2002) covers these methods in detail. His book contains numerous examples as well as the relevant mathematics. A recent paper by Jakubowski, Patrinos, Porta, and Wisniewski (2016) on the effects of ability tracking and vocational education in Polish schools on subsequent student performance is a detailed illustration of the methods and the challenges in using these methods.

A second broad class of quasi-experiment is the regression discontinuity design. In the simplest form of this design, the individual or entity's assignment to the program is based on a dependable prior measure of their need or merit for the program. For instance, everyone on one side of a clear cutoff point along the continuum is assigned to the treatment group, and

everyone on the other side is not. The sharp cutoff is called a "forcing rule," "threshold," or "assignment rule," depending on the context. This is in contrast to a more complex model in which the regression discontinuity design employs a decision rule that is probabilistic. Regression discontinuity is a particularly useful design in contexts in which eligibility for participation is often assigned using some cutoff point, e.g. clinical practice, public health, social welfare programs. A critical assumption underlying the simplest regression discontinuity is that the early (pretest) measures of need or merit are known to have a simple relationship to an outcome. A simple linear model for the program participants, for instance, is compared to the model for program non-participants to determine if they differ in the intercept, slope, or both. Differences are then causally attributable to the program, unless there are other complications. A recent example produced by Palmer, Mitra, Mont, and Groce (2015) involves attempts to estimate the effect of a new policy in Vietnam which was designed to enhance the use of care by families with young children. The simple prior measure of eligibility for the program was age (children under 6 years old), and the outcome variable was inpatient and outpatient visits to health care providers. The results suggest a positive effect of the policy on health service utilization. Standards for judging the quality in regression discontinuity design and execution are promulgated in a document on the What Works Clearinghouse website (see Schochet et al., 2010). Empirical comparisons of the results of randomized trials against the results of non-randomized impact evaluations suggest that results often do differ. Further, differences in neither the magnitude nor the direction are predictable. The discrepancies have been explored through reviews of intervention studies in health (Deeks et al., 2003), employment and training (Glazerman, Levy, & Myers, 2003), education and economic development (Rawlings, 2005), and other areas. Identifying specific domains in which the non-randomized intervention studies are dependable is crucial for education evaluation and for building better evidence-based policy.

Prevention researchers, among others, distinguish between efficacy trials and effectiveness trials (Flay et al., 2005). The U.S. Department of Education's Institute for Education Sciences, for example, also distinguishes between efficacy and effectiveness studies in its guidelines for grant applications (Institute for Education Sciences, 2014). "Efficacy" trials depend on experts who deploy an intervention in highly controlled local contexts that are well understood, with highly reliable measures of outcomes. The

"effectiveness" trials are mounted later, in environments that are real-world in that the interventions may not be delivered as they ought to be, the outcome measures may not be as reliably measured, and so on.

Cost-Effectiveness

Addressing the fourth class of questions, related to the cost-effectiveness of different interventions, depends on dependable evidence on the first three questions. Economists add value beyond this evidence, provided that dependable estimates of costs can be obtained.

Levin et al. (2012), for instance, developed an interesting analysis of the cost-effectiveness of interventions that improve high school completion rates in the United States. They focused on five such programs in which dependable evidence on the effects of the programs were accessible. Their choice of programs on which to focus depended on systemic reviews of evidence generated by the U.S. Department of Education's What Works Clearinghouse.

The results from Levin et al.'s (2012) analysis are tentative. They are limited by assumptions about the dependability of cost estimates, and are also limited to the programs mounted in the United States. Nonetheless, the results are provocative. Roughly speaking, the cost per extra high school completer (a prevented dropout) is 5 to 10 times higher than the average cost of educating students who are likely to complete high school.

The report is conscientious in warning readers that "cost data should be collected at the same time as impact data, using consistent methods of data collection…and that site level analyses are far more informative than overall program estimates that may mask a very wide range of results" (p. 1).

Few trustworthy studies of the effects of interventions also report on the intervention costs. Guidelines for conducting cost-effectiveness analyses of interventions have been developed for various substantive areas of study (see for instance, Yates, 1999, on prevention and treatment; Levin and McEwan, 2001, in education; and Rossi et al., 2004).

Accumulating Dependable Evidence

The fifth family of questions underlying program evaluation emphasizes the accumulation of evidence of an intervention's effects. The main idea is that a single evaluation is usually insufficient for informing debates about how to improve a major program or practice. Further, a presumption in science, and in evaluation policy, is that the replication of studies and the analysis of assemblies of these studies are crucial. See Valentine et al. (2011) in the

context of prevention science in school-based interventions. See Gueron and Rolston (2013) in the context of welfare experiments in the United States that attend to education matters, including school dropouts.

The effects of a particular program, of course, may vary across ethnic, racial, or economic groups, geopolitical jurisdictions, and so on. Recognizing the average levels of program effect and the variation across replicated studies is then important. In this context, phrases such as "meta-analysis" and "systematic reviews" are used to label the evaluative activity.

The best approaches emphasize quality of evidence. For example, Petrosino, Morgan, Fronius, Tanner-Smith, and Boruch (2012) reported on the effect sizes produced in studies of a large assembly of programs which were designed to reduce school dropout rates among children in low-income countries. Quantitative systematic reviews such as this have become more transparent, accessible, of high quality, and complete in their coverage, on account of organizations such as the international Campbell Collaboration (http://campbellcollaboration.org) and the What Works Clearinghouse in the United States (http://whatworks.ed.gov).

The Use of Evaluation Evidence

Use of evaluation evidence may take many forms. In some cases, the use means that the evaluation results are cited in a legislative, parliamentary, or executive proceeding. In other cases, the evidence may be used to illuminate a discussion in such a proceeding, in a meeting among teachers, or among NGO staff members. Such use may be real or it may be symbolic in the sense of merely using evidence as window dressing.

The use of evaluation evidence may not be well documented. Indeed, research on the use of applied research in education in the United States, including evaluations, has been sparse. For Nutley, Walter, and Davies (2007), it is an irony that products of social and health research, including evaluations, have also not been well tracked as to their use.

Despite the foregoing, one can find good studies of use and of non-use of evaluation evidence, at times. Focus groups of teachers, for instance, reveal that teachers are not disinclined to ignore evidence when it is easily accessed, but their standards of quality differ often from those of the evaluator (Sheratt, Drill, & Miller, 2011). Case studies of actual use are given in Finnigan and Daly (2014), including evidence on how State Education Agencies (SEAs) in

the United States capitalize on U.S. federal agency resources relating to different kinds of evaluation questions. Penuel et al.'s (2016) national survey of research use among school and school district leaders in the United States is the most recent and ambitious such effort available. It focused on education "research" generally rather than evaluation construed as applied specifically. It is explicit, however, in its definitions of use and conscientious in the survey's design, execution, and analysis of results. The "…pieces of research that they (school leaders) named as useful were books, research or policy reports, or peer reviewed journal articles…focused on instructional practices and learning in the classroom…(rather than) selecting curriculum materials" (p. 3).

Case studies of the *failures* to depend on dependable evaluation evidence are no less important. The Scared Straight program in the United States aimed to dissuade young people at risk of committing crime from being delinquent. Dependable evidence summarized by Petrosino, Turpin-Petrosino, and Finckenauer (2000) shows that its effects are negligible or negative. Nevertheless, the results have not prevented the television sector from turning this into a profitable reality series. The Drug Abuse Resistance Education (D.A.R.E.) program in the United States is also a case in point. Some communities have abandoned the program, given the absence of any discernible effects on adolescent drug use, based on controlled trials. Other communities, however, appear to have continued the program, only because they believed the program would be good for relations between police and adolescents (Birkeland, Murphy-Graham, & Weiss, 2005).

There are also examples of continued political support for an education program whose value is unknown, yet popular. The Texas legislature's investment of $37,000,000 in the "Texas Fitness" program is a case in point. No dependable evidence for the program's effectiveness was used by the Texas legislature in continuing the program, and no evidence was used in its eventual termination (von Hippel, 2015).

What lessons might one draw from such examples and from other work on the topic of how the use of dependable evidence might be enhanced? The following covers the factors that drive the use of information in any form. Again, the topic is put into Socratic form.

A Question-Based Theory on the Use of Evaluation Evidence

An informative theory about the use of evidence can be based on simple questions such as the following:

1. Is the potential user aware of the evidence?
2. Does the potential user understand the evidence?
3. Does the potential user have the capacity to use the evidence?
4. Does the potential user have incentives to use the evidence and do these surpass the disincentives for use?

Such questions are implicit in experts' handling of the topic of use, e.g. Newcomer, Hatry, and Wholey (2015).

Certain aspects of the factors underlying each question may be controllable, while others may not be controllable. The controllable aspects may be evaluated empirically through evidence on the probability of positive answers to each question, or the influence may be actively explored by doing controlled trials. The topic is sufficiently important that it has demanded the attention of the U.S. National Academy of Sciences (2016), which has produced videos of its deliberations on the topic (see http://sites.nationalacademies.org/DBASSE/DBASSE_170287).

Awareness of Evidence

If potential users of dependable evidence do not know about the evidence, they will not be able to use it. This is a basic reason for the invention and circulation of academic journals, as well as for the growth of electronic circulation of reports on evaluations.

Enhancing the likelihood of the use of dependable evidence lies partly in assuring that potential users (stakeholders) are involved as advisors or collaborators in the evaluation itself. Virtually all government-sponsored evaluations in education in the United States, for instance, include provisions for a "technical advisory committee" that includes potential users of the evaluation results. Roholt and Baizerman (2014) provide an overview of the use of evaluation advisory groups in several countries, explain their structural differences, and illustrate how they may handle issues related to the use of evidence.

Understanding the Evaluation's Results

Publication of evaluation results in academic journals is insufficient to assure potential users' understanding of it. Many potential users of evaluations will be put off by the academic jargon and by the length of such reports. As a consequence, results of major evaluations are reported in several different ways: abstracts, executive summaries, full web-accessible reports, abbreviated reports for academic journals, and at times, in the trade press.

For example, the Campbell Collaboration publishes its systematic reviews in several formats that vary in length and in level of detail, ranging from full reports to two-page Plain Languages Summaries. Similarly, the What Works Clearinghouse in the United States issues several types of publications, such as intervention reports (which summarize findings on an intervention), practice guides (which provide recommendations for educators), single study reviews, and quick reviews.

Some academics also establish good relationships with science writers who work for newspapers or magazines. Professional science writers and journalists usually write better for lay audiences than do academics.

Capacity to Use Evidence

The potential user of evidence who understands an evaluation report may or may not be positioned well to use its results. For instance, good evaluators are attentive to users' interests. Evaluators, however, usually have no authority or power to guarantee the use of evaluation results. Any influence must then be indirect.

On the other hand, government or private foundation staff members may be in a good position to encourage their bosses to use the information in certain ways. Also, a person in authority, a school principal, head of an education entity, or an elected official may be better positioned. They can issue directives, frame laws, and allocate budgets at times and in ways that are guided by evaluation results.

Not much research has been done on any government agencies' or non-governmental organizations' capacity to use evidence. However, a recent book by Haskins and Margolis (2014) does produce evidence on the use of evidence in a half-dozen federal sectors in the United States. In particular, the authors identify legislative initiatives that are indeed evidence-based, and they give details (in Chapter 1 and elsewhere) on the dollar amounts allocated to the programs.

Incentives and Disincentives to Use Dependable Evidence

A potential user may know about the evaluation, may understand its results, and may have the capacity to use the results. These facts, however, do not ensure that they will be willing to use the results, or that their incentives to use evidence will outweigh their disincentives. For instance, a single evaluation may be insufficiently persuasive or informative to take action on. A particular evaluation may yield results that are scientifically dependable, but

run against a moral or religious value held by the potential user or the user's constituency, family, etc. Beyond all this, it can take considerable time before the results can be used, especially when immediate problems confront the potential user of evidence (see, for instance, Gueron & Rolston, 2013, on 40 years of impact evaluations in the welfare sector and related education sectors in the United States). Consequently, the long-term problems addressed by the evaluation may have to take a lower priority in the local or national political arena.

To meet the challenges to the use of dependable evidence in education, a variety of approaches have been taken in the United States. These include, for instance, improving linkages between legislation on programs with evaluation evidence on the programs' effects. The legislative linkage is of course not new in the pharmaceutical sector: Government agency approval of evidence on effectiveness is required prior to marketing a drug. In education, recent legal structures involve linking federal dollars to programs with demonstrable ability to attenuate a problem. The evidence is often based on randomized controlled trials. The Coalition for Evidence Based Policy has been a leader in assisting the federal government to enact statutes and construct rules on this. See the website http://coalition4evidence for illustrations of governmental enhancement of the use of evidence in education-related teen pregnancy prevention programs, post-secondary education, and nurse-family home visiting programs with an emphasis on education.

At the sub-national level in the United States, Regional Education Laboratories (RELs) have undergone a transformation from entities that have had low standards of evidence to ones that have higher standards and are obligated to tailor the education research and evaluations to suit the needs of schools. In particular, the emphasis is on recognizing local needs, culture, political and bureaucratic preferences, so as to enhance the utility of the evidence they produce. See http://www.ies.ed.gov/ncee/edlabs for descriptions of laboratories in different parts of the country and for hyperlinks to their products and use. In some respects, the REL approach emulates a parallel effort in the health sector where translating research into practice has also presented challenges and has led to the creation of entities that focus on the use of evidence (Grimshaw, Eccles, Lavis, Hill, & Squires, 2012).

In recent years, award systems have been developed to recognize people who have contributed remarkably to the production and use of dependable evaluation evidence. Since 2010, for instance, the Campbell Collaboration has annually awarded the "Robert Boruch Award for Distinctive Contributions to

Research that Informs Public Policy" to individuals who made important contributions to the use of evidence in public policy. In 2013, this award was given to Grover Whitehurst, the first director of the U.S. Institute of Education Sciences. In criminology, the Center for Evidence-Based Crime Policy at George Mason University gives awards to police chiefs and others who have collaborated substantially to mount good evaluations on crime and justice policies, and to the eventual use of the results. Awards help to elevate the visibility of people who have contributed to high-quality evaluation work and help the public and professional communities to understand the importance of the work and use in policy or practice. Evaluation evidence on the incentives that these kinds of awards and others produce is sparse.

Concluding Remarks

The rate at which opinions, anecdotes, and bold claims are reported in the social media and popular press will always far exceed the production of dependable evaluation evidence. The peer review system in evaluation research, as in other scientific quarters, is imperfect. Nonetheless, they far exceed many other sources of information for dependability of evidence. The web is a blessing on account of its allowing good evaluators to access information. It will continue to be a resource, beyond the hyperlinks given in this paper, in searching for dependable evidence and being able to criticize reports that are seriously incorrect or serially mendacious in their intent.

The digital landscape will change, of course. It is up to us, our colleagues, and our students in the evaluation communities to learn how to locate and use dependable resources that concern the production and use of evaluations that can inform policy and practice. It is a fine opportunity and a fine challenge.

Acknowledgements

Research on the topic has been sponsored by the National Science Foundation Grant 1337237. This paper was the basis for a keynote address at the National Taiwan Normal University's Center for Educational Research and Evaluation. The particular venue was the Center's "International Conference on Educational Evaluation: Accountability, Policy Learning and Capacity Building" convened in Taipei in 2015.

References

Alkin, M. (Ed.). (2012). *Evaluation roots: A wider perspective of theorists' views and influences*. Thousand Oaks, CA: Sage.

American Institutes for Research. (2015). *The economics of improving lives: Cash transfers in Zambia*. Retrieved from http://www.air.org/resource/economics-improving-lives-cash-transfers-zambia

Baker, J., & Boruch, R. (2015). *Ambient positional instability among Ohio math and science teachers: 2008 to 2014*. Retrieved from University of Pennsylvania, Scholarly Commons website: http://repository.upenn.edu/gse_pubs/269

Birkeland, S., Murphy-Graham, E., & Weiss, C. (2005). Good reasons for ignoring good evaluation: The case of the drug abuse resistance education (D.A.R.E.) program. *Evaluation and Program Planning, 28*(3), 247-256.

Bowdon, J., & Boruch, R. (2014). *Teacher churn in Missouri's five biggest cities, 2005-2014: A briefing*. Retrieved from University of Pennsylvania, Scholarly Commons website: http://repository.upenn.edu/gse_pubs/264

Bruns, B., Filmer, D., & Patrinos, H. A. (2011). *Making schools work: New evidence on accountability reforms*. Washington, DC: The World Bank.

Chen, X., Huang, X., Wang, L., & Chang, L. (2012). Aggression, peer relationships, and depression in Chinese children: A multiwave longitudinal study. *Journal of Child Psychology and Psychiatry, 53*(12), 1233-1241. doi: 10.1111/j.1469-7610.2012.02576.x.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., & Song, F., … European Carotid Surgery Trial Collaborative Group. (2003). Evaluating non-randomized intervention studies. *Health Technology Assessment, 7*(27), 1-173. doi: http://dx.doi.org/10.3310/hta7270

Finnigan, K. S., & Daly, A. J. (Eds.). (2014). *Using research evidence in education: From the schoolhouse door to Capitol Hill*. New York, NY: Springer.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., … Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness, and dissemination. *Prevention Science, 6*(3), 151-175. doi: 10.1007/s11121-005-5553-y

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The ANNALS of the American Academy of Political and Social Sciences, 589*(1), 63-93. doi: 10.1177/0002716203254879

Grimshaw, J. M., Eccles, M. P., Lavis, J. N., Hill, S. J., & Squires, J. E. (2012). Knowledge translation of research findings. *Implementation Science, 7*(50), 1-17. doi: 1011861/1748-5908-7-50

Gueron, J. M., & Rolston, H. (2013). *Fighting for reliable evidence*. New York, NY: Russell Sage Foundation.

Haskins, R., & Margolis, G. (2014). *Show me the Evidence: Obama's fight for rigor and results in social policy*. Washington, DC: Brookings Institution Press.

Institute for Education Sciences. (2014). *Request for applications: Education research grants, CFDA number 84.305A*. Washington, DC: U.S. Department of Education.

Jakubowski, M., Patrinos, H. A., Porta, E. E., & Wisniewski, J. (2016). The effects of delaying tracking in secondary school: Evidence from the 1999 educational reform in Poland. *Education Economics*. Advance online publication. doi: 10.1080/09645292.2016.1149548

Lee, T. S., Fu, L. A., & Fleming, P. (2006). Using focus groups to investigate the educational needs of female injecting heroin users in Taiwan in relation to HIV/AIDS prevention. *Health Education Research, 21*(1), 55-65. doi: 10.1093/her/cyh041

Levin, H. M., Belfield, C., Hollands, F., Bowden, A. B., Cheng, H., Shand, R., … Hanisch-Cerda, B. (2012). *Cost-effectiveness analysis of interventions that improve high school completion*. New York, NY: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (Eds.). (2015). *Handbook of practical program evaluation* (4th ed.). Hoboken, NJ: Jossey-Bass.

Nutley, S. M., Walter, I., & Davies, H. T. O. (2007). *Using evidence: How research can improve public services*. Edinburgh, England: Policy Press.

Palmer, M., Mitra, S., Mont, D., & Groce, N. (2015). The impact of health insurance for children under age 6 in Vietnam: A regression discontinuity approach. *Social Science & Medicine, 145*(C), 217-226. doi: 10.1016/j.socscimed.2014.08.012

Pan, H. W. (2014). School practices of leading learning in Taiwan. *Leading and Managing, 20*(2), 27-42.

Park, H., Behrman, J. R., & Choi, J. (2013). Causal effects of single-sex schools on college entrance exams and college attendance: Random assignment in Seoul high schools. *Demography, 50*(2), 447-469. doi: 10.1007/s13524-012-0157-1

Parker, S. W., & Teruel, G. M. (2005). Randomization and social program evaluation: The case of Progresa. *The ANNALS of the American Academy of Political and Social Science, 599*(1), 199-219. doi: 10.1177/0002716205274515.

Patton, M. Q. (2008) *Utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Penuel, W. R., Briggs, D. C., Davidson, K. L., Herlihy, C., Sherer, D., Hill, H. C., … Allen, A.-R. (2016). *Findings from a national survey of research use among school and district leaders* (Technical Report No. 1). Retrieved from http://ncrpp.org/ assets/documents/NCRPP_Technical-Report-1.pdf.

Petrosino, A., Morgan, C., Fronius, T., Tanner-Smith, E., & Boruch, R. (2012). *Interventions in developing nations for improving primary and secondary school enrollment of children: A systematic review*. Retrieved from the Campbell Collaboration Library of Systematic Reviews website: http://www.campbellcollaboration.org/lib/project/123/

Petrosino, A., Turpin-Petrosino, C., & Finckenauer, J. O. (2000). Well-meaning programs can have harmful effects! Lessons from experiments of programs such as Scared Straight. *Crime and Delinquency, 46*(3), 354-379.

Rawlings, L. (2005). Operational reflections on evaluating development programs. In G. K. Pitman, O. N. Feinstein, & G. N. Ingram (Eds.), *Evaluating development effectiveness* (pp. 193-204). Piscataway, NJ: Transaction.

Rimm-Kaufman, S. E., Larsen, R. A. A., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., … DeCoster, J. (2014). Efficacy of the responsive classroom approach: Results from a 3-year, longitudinal randomized controlled trial. *American Educational Research Journal, 51*(3), 567-603. doi: 10.3102/0002831214523821

Roholt, R., & Baizerman, M. (2014). "Making things better" by using evaluation advisory groups. In S. Kalliola (Ed.), *Evaluation as a tool for research, learning and making things better* (pp. 69-86). Newcastle upon Tyne, England: Cambridge Scholars.

Rosenbaum, P. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter J., & Smith, J. (2010). *Standards for regression discontinuity designs.* Retrieved from What Works Clearinghouse website: https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rd.pdf

Sherratt, E., Drill, K., & Miller, S. R. (2011). *Is the supply in demand? Exploring how, when, and why teachers use research*. Washington, DC: American Institutes for Research. Retrieved from http://www.air.org/sites/default/files/downloads/report/Exploring%20How%20Teachers%20Use%20Research_Jan%2011.pdf.

Sznitman, S. R., Reisel, L., & Romer, D. (2011). The neglected role of adolescent emotional well-being in national educational achievement: Bridging the gap between education and mental health policies. *Journal of Adolescent Health, 48*(2), 135-142. doi:10.1016/j.jadhealth.2010.06.013

U.S. National Academy of Sciences. (2016). *Seminar on clear and credible information from social and behavioral science to improve societal outcomes: Current approaches and future directions.* Retrieve from http://sites.nationalacademies.org/DBASSE/DBASSE_170287

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., … Schinke, S. P. (2011). Replication in prevention science. *Prevention Science, 12*(2), 103-117. doi: 10.1007/s11121-011-0217-6

von Hippel, P. T. (2015, August 27). Texas needs more evidence-based policy decisions. *The Dallas Morning News*. Retrieved from http://www.dallasnews.com

Wijekumar, K., Meyer, B. J., Lei, P., Lin, Y., Johnson, L. A., Spielvogel, J. A., … Cook, M. (2014). Multisite randomized controlled trial examining intelligent tutoring of structure strategy for fifth-grade readers. *Journal of Research on Educational Effectiveness, 7*(4), 331-357. doi: 10.1080/19345747.2013.853333

Yates, B. T. (1999). *Measuring and improving cost, cost-effectiveness, and cost-benefit for substance abuse treatment programs* (NIH Publication No. 99-4518). U.S. Department of Health and Human Services, National Institutes of Health, National Institute on Drug Abuse. Retrieved from http://archives. drugabuse.gov/pdf/Costs.pdf

Zeng, G., Fu, P., May, H., Lopez, B., Suarez-Morales, L., Voelkle, M. C., … Boruch, R. F. (2012). America's youngest kindergarteners' elevated levels of internalizing problems at school entry and beyond: Evidence from the Early Childhood Longitudinal Study. *School Mental Health, 4*(3), 129-142. doi: 10.1007/s12310- 012-9077-x