Overview and Rationale for the Systematic Screening and Assessment Method

Laura C. Leviton, Marjorie A. Gutman

Abstract

This chapter gives the rationale and conceptual defense for the Systematic Screening and Assessment (SSA) Method, a way to identify the most promising innovations in preparation for rigorous evaluation. The SSA Method starts by identifying the innovations that real-world practitioners have developed and then systematically assesses which innovations will offer the greatest payoff from further evaluation. Through sequential purchase of evaluative information, the SSA Method aims to avoid a variety of problems that evaluators encounter in studying innovations. Besides the identification of promising innovations for evaluation, the benefits include expert technical assistance and feedback to innovators, cross-site syntheses of trends and common practices, and cost efficiency compared to evaluating innovations that have not been selected or screened for promise. © Wiley Periodicals, Inc., and the American Evaluation Association.

The Systematic Screening and Assessment (SSA) Method identifies for evaluation the practice-based innovations that are most likely to be effective, thus increasing the potential for productive outcome evaluation. The SSA Method does this by improving the prior information about
these innovations’ plausibility for impact, feasibility, and readiness for evaluation. The most promising innovations then receive the highest priority for outcome evaluation resources. SSA is also useful for informing the field about prevalent practices and providing constructive feedback and technical assistance to innovators. However, its primary purpose is to prepare the way to evaluate the effectiveness of practice-based innovations and disseminate their models for adoption elsewhere.

We believe the SSA Method is a new and important addition to the methods available to evaluate innovations. However, there are so many unjustified claims about new evaluation practices that we need to be careful to justify exactly what is new about this one. The method is new in that it sequences two existing evaluation methods, namely evaluability assessment (Whooley, 1979, 2004) and connoisseurship or expert judgment (Rossi, Lipsey, & Freeman, 2003). By sequencing these methods, the SSA Method creates a systematic process for nominating, screening, and assessing promising innovations. As discussed later, SSA joins a family of emerging methods to translate practice into research by focusing evaluation attention on strategies and interventions that are currently being undertaken by practitioners (Glasgow, Green, et al., 2006).

The chapters in this issue of New Directions for Evaluation describe application of the Systematic Screening and Assessment Method to childhood obesity prevention in a 2-year initiative titled the Early Assessment of Programs and Policies to Prevent Childhood Obesity (the Early Assessment initiative). The Robert Wood Johnson Foundation (RWJF) funded the initiative and the Centers for Disease Control and Prevention (CDC) directed it. ICF Macro served as the coordinating center, trained 40 professionals in the conduct of evaluability assessments, and conducted 48 evaluability assessments.

Childhood obesity is not the only application of the SSA. The first year results of the Early Assessment project led to three other efforts to identify innovations worth evaluating, in nursing education (Rutgers University with funding by RWJF), programs to address intimate partner violence in immigrant populations (LTG Associates with funding by RWJF), and heart disease and stroke prevention (CDC and ICF Macro).

Overview of the SSA Method

Figure 1.1 summarizes the six-step Systematic Review and Assessment Method.

Step 1: Selection of priority areas for assessment, to focus a scan of innovations. Ideally, the priority areas for studying innovations should be selected in conjunction with key stakeholders and potential users. The Early Assessment initiative selected five such priority areas in consultation with federal stakeholders and experts: school district local wellness policies, day care and after-school settings, access to healthy food in poor communities,
comprehensive school-based physical activity programs, and features of the built environment that facilitate physical activity. They were chosen because relatively little is known about them, yet they are prevalent and have high potential leverage for changing policy and environmental conditions that affect childhood obesity.

Step 2: Scan of innovations in those priority areas. The SSA Method requires a way to identify a high volume of innovations, either through

Figure 1.1. The Systematic Review and Assessment Method

- Inputs
  - Guidance from foundation staff, collaborators, and advisors
  - Nominations, existing inventories, descriptions
  - Expert review panel(s)
  - Distributed network of practitioner/researchers
  - Expert review panel(s)
- Steps
  1. Choose priorities for the scan
  2. Scan environmental interventions
  3. Review and identify interventions that warrant evaluability assessment
  4. Evaluability assessments
  5. Review and rate interventions for promise/readiness for evaluation
  6. Use information
- Products
  - Focus for scan of environmental interventions
  - Brief description of each environmental intervention
  - List of interventions for evaluability assessment
  - Report on priority interventions
  - Ratings and reports
  - Position the promising interventions for rigorous evaluation
  - Constructive feedback for intervention refinement
  - Synthesis of findings
nomination or other alternatives (described below). In the Early Assessment initiative, we asked a wide range of public health professionals and policy makers for nominations in the priority areas. The CDC staff obtained 458 nominations over the 2-year period, of which 174 met inclusion criteria and 128 were summarized for step 3.

**Step 3: Initial review of innovations by an expert panel.** In SSA, the initial list of innovations is screened to further select those that have potential promise using explicit, mutually agreed criteria. Experts rated the plausibility of:

- Potential impact
- Reach to cover the target population
- Feasibility of adoption by similar organizations
- Generalizability
- Staff and organizational capacity to undergo and use evaluability assessment

The innovations that meet these criteria are then scheduled for evaluability assessment (EA). In the Early Assessment initiative, we created a panel of content experts and evaluation professionals who reviewed documentation on the selected, nominated programs and policies. The panel reviewed 128 of the innovations and recommended 53 of them for EA.

**Step 4: Evaluability assessment of the selected innovations.** For each innovation selected by expert judges in the SSA Method, an EA is conducted to elicit a program description that includes a logic model and theory of change, articulates the expectations of stakeholders, and helps us better understand the reality of program resources and activities (Leviton, Kettel Khan, Rog, Dawkins, & Cotton, in press; Wholey, 1979, 2004). On this basis we can further understand the innovations’ plausibility for effectiveness and feasibility of adoption in other settings. In the Early Assessment initiative, ICF Macro staff trained a distributed network of researchers and public health professionals in EA, which then conducted an EA on each innovation. Innovators generally welcomed these assessments: of the 53 selected for EA, all but 5 participated and 2 of these refusals were unrelated to their interest in participating.

**Step 5: Second expert panel review of the assessed innovations.** In the SSA Method, the expert judges review the innovations a second time using the evaluability assessment reports to further reduce uncertainty about their promise, feasibility, and readiness for more rigorous evaluation. In the Early Assessment initiative, evaluability assessment reports were reviewed for 48 innovations, of which 33 were ready for more formal evaluation. The expert panel deemed 20 of these to be highly promising for substantial impact at a population level and identified 6 of them as having the highest priority for evaluation.

**Step 6: Use of information.** The SSA Method results in three distinct uses of information derived from the process. First, the most promising innovations
are positioned for rigorous evaluation. These are the expert panel's top priorities in terms of likely impact, reach, feasibility, generalizability, and potential information payoff. Funders may proceed immediately with full evaluation of the top priorities, or further planning and preparation may be required. The Early Assessment initiative identified 6 top priorities for evaluation: in 2009, evaluation was under way for 3 of these and another 2 were approved for funding. In addition, a variety of funders are considering evaluation of the remaining 15 innovations found to be both highly promising and ready for evaluation.

Second, the developers and managers of the innovations that undergo evaluability assessments receive tailored, constructive feedback about program operations on the basis of the evaluability assessments. In the Early Assessment initiative, feedback also included content expertise from CDC experts in physical activity and nutrition. As seen in Chapters 2 and 3, innovators found this feedback useful.

Third, after reviewing a high volume of innovations, we get a sense of the field in question. All the innovations that are screened and reviewed, whether promising or not, reveal a great deal about the types of projects being undertaken, their prevalence and popularity, and their facilitators, obstacles, and fatal flaws. Chapter 5 describes the utility of this information in the Early Assessment initiative, as do synthesis reports on the topics of focus, which will soon be posted by CDC at the following index: http://www.cdc.gov/nccdphp/dnpao (Cheung, Dawkins, Kettel Khan & Leviton, 2009; Pitt Barnes, Robin, Dawkins, Leviton, & Kettel Khan, 2009a, 2009b; Skelton, Dawkins, Leviton, & Kettel Khan, 2009; Wethington, Kirkconnell Hall, Dawkins, Leviton, & Kettel Khan, 2009).

**Why This Combination of Steps?**

The methods employed in SSA have all been used before, but alone or in isolated pairings. By combining them systematically, the SSA Method stretches limited evaluation resources by identifying innovations with promise for effectiveness. Combining these nomination, expert rating, and assessment activities in a seamless screening process, SSA aims to:

- Avoid wasting precious resources on a “no effect” conclusion (Whooley, 1979)
- Prevent the chilling effect that a premature negative evaluation can have on innovations to combat childhood obesity (Shadish, Cook, & Leviton, 1991; Weiss, 1987)
- Provide useful formative feedback to projects for a low cost and low response burden, optimizing their further development (Carman & Fredericks, 2008; Leviton et al., in press; Smith, 1989)
- Prepare the highest-priority innovations for rigorous evaluation of effectiveness (Cook, Leviton, & Shadish, 1985)
• Offer timely insights to the field and funding organizations on the range of innovations, their strengths, and limitations (Cheung et al., 2009; Pitt Barnes et al., 2009a, 2009b; Skelton et al., 2009; Wethington et al., 2009)

Once we have identified the innovations that are most promising and ready for evaluation, the logical next step (and the primary reason for the SSA Method) is to evaluate their effectiveness. RWJF and the CDC are taking this step for five of the six highest-priority innovations from the first year of the Early Assessment initiative. In addition, the rest of the innovations that were deemed promising and ready for evaluation are being considered for development and test by a variety of other funders and researchers interested in these issues.

Rationale for the SSA Method

We developed the SSA Method to address a dilemma in evaluation. Society expects evaluation to identify effective innovations, but society seems to spend a lot of money evaluating innovations that turn out not to be effective. Some innovations may have been well conceived but ultimately ineffective; others may be less well-conceived, or they may be implemented with inadequate consistency or intensity. Even with those that are deemed effective, front-line practitioners may find them not feasible to employ. This situation serves neither evaluation nor the society that supports it very well.

To address this problem, the SSA Method relies on five assumptions:

1. Practitioners and policy makers are developing many innovations to address a problem, such as childhood obesity.
2. At least some of these innovations may be effective.
3. Because practitioners developed the innovations, they are likely to be feasible for implementation in real-world situations.
4. Using the principle of sequential purchase of information (Wholey, 1979), we can screen a high volume of these innovations at relatively low cost.
5. Through the SSA Method, we can reduce uncertainty about which innovations are likely to be effective and feasible, and therefore appropriate for full evaluation.

Where these assumptions hold, we can capitalize on the variety of practice-based innovations at the same time that we preserve scarce evaluation resources to test the innovations that are most worth evaluating.

Evaluating Innovations: An Uncertain Endeavor. In the face of any new social problem, practitioners, organizations, and policy makers attempt a variety of innovations. In this sense, societal innovation is an adaptive mechanism; it produces variation, some of which may be effective in dealing with the social problem. When a new problem presents itself, one can
expect to see a lot of innovation, through both practice and research. Certainly this is the case in childhood obesity prevention, where neither research nor practice has yet developed sufficiently to offer good solutions (Koplan, Liverman, & Kraak, 2001). At the Robert Wood Johnson Foundation and at the Centers for Disease Control, evaluators have taken a front row seat to observe the many innovations and solutions that are put forth. But as funders we must ask, Which of these innovations should be further developed and tested for effectiveness?

In principle, evaluation serves a useful function as a selection mechanism to establish which variants are effective. Donald Campbell (1977) related this process to the philosophical concept of evolutionary epistemology, in which trial-and-error learning permits us to discover reliable causal relationships (Shadish et al., 1991). In the case of childhood obesity prevention, we would like to fund rigorous evaluation to assess the effectiveness, reach, and feasibility of the innovations that are put forth. But which of the many innovations should be evaluated? Resources do not permit an adequate test of each and every innovation. Like most evaluators and funders, we want to increase the probability that the innovations we evaluate turn out to be effective and have leverage to produce change in many places.

Unfortunately, evaluation as normally practiced does not make cost-effective use of evaluation resources to identify effective innovations. In evaluation practice of all varieties, the client (a funder, an organization, or a participatory coalition) usually presents the evaluator with a given innovation to be evaluated. Obviously, the client believes there is a reasonable likelihood that the innovation is effective, thinking this ideally on the basis of some prior information. The prior information can come from experience, literature review, program theory, management information, available data, or evaluation. All too often, however, prior information is very limited as to an innovation’s maturity, stability, and plausible effectiveness. Sadly, evaluation of such an innovation has a high likelihood of reaching a no-effect conclusion. Influential evaluation writers have spoken to why this might be the case (Lipsey, 1988; Reichert, 1994; Wholey, 1979; Wilson & Lipsey, 2001). Sometimes a no-effect conclusion can be valuable (Henry, 2003). However, in our experience many programs with no-effect conclusions could have been established as ineffective using much-less-costly methods than those employed in a conventional evaluation. Moreover, Weiss (1987) has noted that premature negative evaluation can have a chilling effect on innovation, leading to the assumption that nothing is effective, when in fact something might be effective to address the problem. In areas such as childhood obesity prevention, we cannot afford to discourage innovation—there is a serious need to reverse a disturbing national trend.

**More Cost-Effective Evaluation of Innovations.** One way to improve the likelihood of finding an effective innovation is to have a larger pool of potential innovations from which to choose for evaluation. In the SSA
Method, therefore, we begin by identifying as large an array of innovations as possible in a given area. However, testing all of them would be prohibitively expensive, and as seen below we have reason to believe that the yield from testing all such innovations would be small.

To ensure that the process yields more information in a more cost-effective manner, we borrowed from Wholey’s (1979) principle of sequential purchase of information. Wholey coined this term to describe a more orderly process for producing useful evaluative information. By employing less expensive methods first, Wholey used information to further refine stakeholder questions, assess whether a program could indeed be evaluated, and also assess whether stakeholders would in fact use the information. In other words, less expensive information was used to reduce uncertainty about whether more expensive evaluation information would be useful (Cronbach, 1982).

In the SSA Method, we employ sequential purchase of information in a somewhat different way than Wholey and his colleagues did: to select from among many innovations those that have a good prior probability of having impact. Only if the initial phases of SSA indicate the promise of an innovation are we justified in collecting additional information. Thus expert judgment is used to determine whether evaluability assessment is warranted. Evaluability assessment is used to gather information about the project’s promise and readiness for evaluation. A second expert judgment after the evaluability assessments is used to determine whether more expensive, rigorous evaluation of an innovation is warranted.

**Making the Case for Cost-Effectiveness.** To support our belief that the SSA Method is cost-effective, we invite the reader to participate in a simple thought experiment, or simulation. Examine the yield from the Early Assessment initiative as reported in Chapter 2, and then imagine the resources that would be required if formal evaluations of effectiveness were conducted on all of the nominated projects, or on any randomly selected project. Across the 2 years of the Early Assessment initiative, 458 policy and environmental innovations were nominated. Of these, 174 (38%) met inclusion criteria, meaning they were appropriate nominations (see Chapter 2). For various reasons, 128 of these appropriate nominations were presented to the expert panel for review. We budgeted for about 50 evaluability assessments, and the expert panel had few problems in identifying 53 that seemed the most promising to warrant inclusion. Of these, 48 innovations underwent evaluability assessment, 33 of these were ready for more formal evaluation and 20 of the innovations met our criteria for high promise and readiness for evaluation. Six of the innovations were judged to have the highest priority for formal evaluation. In other words, the yield from the initiative was that 4.4% of all nominations were deemed highly promising, between 11.5% and 15.6% of all appropriate nominations were viewed as promising, and 42% of the nominations that underwent evaluability assessment were both highly promising and ready for evaluation.
We should consider the yield from each of these steps in turn, starting with all 458 nominations. If only 4.4% of identifiable innovations in a new area of policy and programs are promising and therefore have even a reasonable likelihood of being effective, then this argues against a willy-nilly intensive evaluation of every innovation that is nominated. The cost of doing so would dwarf our capacity to implement these innovations, let alone study them. Evaluation of the innovations for other purposes—formative feedback, program improvement, monitoring—might be warranted if resources were available. But recall that the purpose of the SSA Method is to identify the most promising, highest-priority innovations for testing and potential spread.

Granted, 62% of the nominations in this initiative were not appropriate for one reason or another, and it might be feasible to narrow the focus for nominations to better select those that meet criteria. Nevertheless, someone decided that each and every one of the 458 innovations might be worth evaluating! Veteran evaluation practitioners will recognize a familiar pattern, because many of them have been asked to evaluate inappropriate innovations. Even in well-developed program areas, many programs undergo evaluation when they exist only on paper, or they require substantial development before summative evaluation would be appropriate (e.g., Lipsey, 1988; Reichert, 1994; Wholey, 1979, 2004).

Of course, a serious effort to study innovations would instead focus on the 174 appropriate nominations—provided they were recognized as such. An 11.5% yield of promising interventions, using appropriate nominations as the denominator looks much more optimistic than 4.4%. So does a 15.6% yield, considering that the expert panel reviewed only 128 of the appropriate nominations. We may also have missed a few promising projects in the transition from 174 to 128 innovations (see Chapter 2), so we might even inflate this percentage a bit. A CDC effort that preceded SSA identified 9 out of 41 projects as promising, so let us assume that between 15% and 25% of the appropriate nominations are likely to be promising. Yet even if 25% of innovations were promising and ready for evaluation, we would still argue that the SSA Method can save substantial evaluation resources and direct them to areas for the biggest payoff.

Here is our reasoning: of the $3 million RWJF/CDC initiative, $600,000 was reserved to conduct a formal evaluation of the highest-priority innovation identified in year one. Therefore the cost per innovation identified as highly promising and ready for evaluation, including all infrastructure, training, and technical assistance, was $2.4 million divided by 20, or $120,000. The SSA Method is cheap at that price; consider the many very expensive evaluations that have reached no-effect conclusions over the past 35 years! An adequate study of intervention effectiveness is at least $2 million, so if formal evaluation were conducted on all the eligible innovations, and one in four were both promising and ready for evaluation, we would have to spend at least $8 million in order to identify even one intervention
that had a good chance of being found effective. Even if evaluations could be done on the cheap, at around $150,000 each, then to find interventions with even a hope of being effective we would have to spend at least $600,000 to identify one promising intervention, and to find all of them we would spend 174 times $150,000, or more than $26 million.

The same argument applies to the 48 innovations that underwent EA; 28 of them were either not promising or not ready for evaluation, so outcome evaluation would certainly be premature, and probably the evaluation of implementation as well, given that program development could be guided by the EAs as a first step (Smith, 1989). Among the 20 highly promising innovations that were ready for evaluation, expert judges were very clear that evaluating 6 of them would produce the biggest payoffs for childhood obesity prevention.

We are still underestimating the resources that would be needed if we evaluated unselected innovations to detect the effective ones. Wilson and Lipsey (2001) have demonstrated how the study conditions in practice-based settings reduce the obtained effect sizes from evaluations. Effect sizes decrease because the study conditions are not so well controlled, introducing problems of measurement, design, analysis, and treatment intensity. For these reasons, the probability of finding an effective intervention would be lower, and the actual cost per effective innovation would be substantially more than the amounts estimated in our thought experiment. It takes powerful interventions, like the six priorities identified by our expert panel, to overcome these conditions. A $120,000 expenditure to identify a highly promising innovation starts to look thrifty indeed compared to the general payoff from evaluation of innovations.

The cost per highly promising innovation does not take into account the other benefits deriving from the SSA Method, such as technical assistance and feedback to the developers and managers of the innovations (see Chapter 2), the five syntheses of knowledge about innovations or the improved evaluation capacity of public health (Chapter 5). As they stated in surveys, the innovators found the formative feedback to be useful (see Chapters 2 and 5). For technical assistance and feedback alone, disregarding identification of promising projects, the cost was $50,000 per innovation that underwent EA—a reasonably cost-effective format to improve development of promising innovations in a new area for which solutions are badly needed.

**Comparison With Other Strategies to Test, Disseminate, and Replicate Innovations.** A focus on evaluating innovations is not new, of course. Many federal agencies have undertaken the strategy of research, development, and dissemination (R&D) of programs and practices, including the U.S. Department of Education (n.d.); Katzenmeyer & Haertel, 1986), the Substance Abuse and Mental Health Services Administration (2009), National Institute of Mental Health (2009; Davis & Salasin, 1975), the National Institutes of Health (2008), the Agency for Healthcare Research
OVERVIEW AND RATIONALE FOR THE SSA METHOD

and Quality (2002), National Institute of Justice (2008), and the CDC (for HIV prevention, 2008). Generally these R&D programs solicit innovations that are evaluated for effectiveness and then disseminate the effective ones to practitioners.

These are extremely important initiatives, and we do not pretend that the SSA Method could replace them. However, the methods for soliciting innovations to undergo a test could be more systematic and cost-effective. We are not the first to make this observation (e.g., Cronbach, 1982). Consider the total amount of time, effort, and resources that were required for testing the innovations in these federal initiatives and then culling the ineffective ones. The total cost is daunting indeed. How many thousands of evaluations had to be conducted to identify the programs that are endorsed and disseminated by federal agencies? Would they have shortened the process by increasing the prior probability of selecting the most promising interventions, and ensuring that they could be evaluated?

By adapting the evaluation principle of sequential purchase of information, the SSA Method focuses on the innovations that are most likely to be effective and amenable to evaluation. In this way, the funders of evaluation can increase the chances that subsequent, more rigorous evaluation focuses on those programs and practices with the best promise of achieving the desired effects.

One might argue that the peer review process of research grant funding also represents a way to improve the probability that only the promising innovations will receive an evaluative test. Certainly, federal programs select many innovations this way. However, peer reviewers rely almost exclusively on paper documents and assurances about pilot tests. On occasion, site visits are made, but rarely does the process include reality testing of the kind represented by evaluability assessment.

Assuring Feasibility for Real-World Implementation. Conventional federal research and development initiatives rely on researchers to develop innovations that practitioners and program managers are then expected to adopt. This is the evidence-based practice model. One challenge to these efforts is that researchers often do not have a good conception of the constraints on practice in real-world settings. Program managers and practitioners repeatedly express concern that they cannot implement these innovations: they do not have the resources and time, or they must adapt the innovations to fit the needs of the setting. In some cases, the innovations may not be acceptable to the populations they aim to affect. These complaints arise chronically in areas as disparate as medical care quality (Berwick, 2003), community mental health (Davis & Salasin, 1975), AIDS prevention (Leviton & Guinan, 2003), and criminal justice and education (Emshoff et al., 1987). Although real-world organizations can indeed implement well-constructed programs with fidelity (Emshoff et al., 1987; Griffin et al., 2009), this issue still poses a constant challenge to the R&D model of innovation.
To address the problem of real-world implementation, Glasgow, Green, et al. (2006) have called for an effort to “translate practice into research” to identify and develop practices, programs, and policies that have proven feasible and acceptable in the real world of health and social services. This effort, the practice-based evidence model, would parallel the federal initiatives that rely on researchers to develop and disseminate innovations.

The SSA Method translates practice into research, because it starts with the premise that practitioners and program managers have themselves developed innovations worth investigating. Some relevant efforts have preceded this work. The school improvement literature offers examples (Teddie & Stringfield, 2007), as does the CDC SWAT initiative to be described here (Dunet et al., 2008; Hersey et al., 2008). RWJF has sometimes sought real-world practices for a test; for example, the Finding Answers national program solicits interventions on cardiac care for minority populations that then undergo evaluation and dissemination of those found to be effective (Schlotthauer et al., 2008).

Three cautions are in order about the focus on real-world practices. First, the SSA Method assumes that such innovations exist and at least some of them are likely to be effective. In the case of childhood obesity, this assumption is reasonable. However, other new social programming areas may not have much practice-based innovation, or the innovations that exist may not be sufficiently powerful to achieve the objective. If this were the case, then the R&D focus as employed in federal initiatives might be more appropriate than SSA, because researchers could apply existing behavioral and social theory to develop innovations. Nevertheless, SSA might be helpful because of its explicit focus on the theory of change and real-world assessment.

A second caution is that we assume innovations developed by practitioners are more likely to be implemented than those developed by researchers. This may not be the case for all innovations, however. Some researchers may develop innovations with practitioners in mind, and some practitioner-developed innovations may be so constrained by local circumstances as to be completely infeasible for replication elsewhere. In the case of childhood obesity, however, three of the top innovations identified through the SSA Method are already being adopted or considered for adoption by other policy makers and practitioners in the nation.

A third caution is perhaps most serious from an evaluation standpoint. SSA assumes that more rigorous evaluation will follow the selection of innovations through the six-step process. However, SSA may be misused; practitioners or decision makers could choose to disseminate the identified models without a proper evaluative test. This danger is exemplified by the many efforts to identify best practices in various program and policy sectors. By what standard do we regard these practices as “best”? The standards are sometimes good, sometimes not so good. At times, dissemination may
be warranted without an evaluation, as when a body of evidence gives compelling support to the theory behind the best practice, or when common sense makes it patently obvious that the best practice should be disseminated. This was the case for the CDC SWAT initiative (to be described), in which high-quality studies of worksite obesity prevention had already been conducted. However, in a new field, or when innovations depart from existing evidence-based practice, it is not sufficient to disseminate best practices. SSA is explicit in its aim to subject promising approaches to a rigorous test.

Development of the SSA Method

The idea for the SSA Method originated in discussions between the first author and Thomas Cook in the early 1980s (Cook, Leviton, & Shadish, 1985). However, the idea was never implemented, to our knowledge or Cook’s, perhaps because only a limited number of funders can implement it at the necessary scale. At RWJF, the timing was right to explore this idea as we tackled the question of how to prevent childhood obesity. The Institute of Medicine had recommended changing children’s environments to prevent obesity: the availability of healthy foods, limits on access to unhealthy foods, and reintroducing opportunities for physical activity into children’s lives (Koplan et al., 2001). To implement environmental changes often requires policy changes at the local, state, or federal level. Many localities and some states were beginning to innovate in creation of such policies; however, very little was known about their effectiveness. The SSA Method reemerged out of discussions of the need for more rapid identification of innovations that might be effective in preventing childhood obesity.

Three other initiatives informed development of SSA. They help to locate it within a family of methods that first identify promising programs and practices and then investigate them. First was the school improvement effort of the early 1980s, which identified schools that were performing better than would be predicted on the basis of their resources and the background of the children (Teddie & Stringfield, 2007). High-performing schools were selected either because of “outlier” status or consistently high performance year after year, controlling for these background characteristics. Investigators found consistent differences between these schools and lower-performing schools. However, independent outcome studies of any of these characteristics were extremely rare (Teddie & Stringfield, 2007). Unlike the outlier methods as applied to school improvement research, the SSA Method aimed specifically to plan for independent tests of the effectiveness of the promising exemplars. Nevertheless, the school improvement literature gave us confidence that we might identify promising practices using SSA.

A second initiative consisted of 27 EAs that RWJF funded from 2005 to 2007, on pilot projects on childhood obesity prevention. Three separate
evaluation teams conducted these evaluability assessments so that we could get an idea of the diversity of approaches and findings (OMG Center for Collaborative Learning, 2007; Rhodes, 2007; Rog & Gutman, 2007). These EAs gave us a sense of the unit cost for third-party evaluability assessments, at about $35,000 per innovation studied. The cost can be lower; Leviton, Collins, Laird, and Kratt (1998) describe 25 EAs that master’s of public health students conducted as service learning projects.

The third project was a CDC initiative, the Swift Worksite Assessment and Translation (SWAT) project. We need to give credit where credit is due: SWAT predated the Early Assessment initiative and employed many of the same principles. Indeed, it gave decision makers confidence to proceed with the Early Assessment initiative. The CDC SWAT project’s goal was to develop a rapid assessment method that could investigate innovations being implemented in the field one at a time against a standard set of criteria. At the time of the SWAT project’s initiation, much attention was being turned to obesity in various settings, including worksites, and CDC staff was being asked about the effectiveness of new interventions that had not been formally studied. In addition, some innovations spread rapidly, despite questionable appropriateness (such as public weigh-ins meant to shame employees into healthier behavior). Thus the need arose for a systematic way to quickly investigate an innovation and make a preliminary determination about its public health relevance and probable effectiveness in addressing obesity. In its developmental first round, the SWAT project identified 41 innovations of worksite obesity prevention and implemented 9 SWAT assessments in a “batchlike” approach in order to further refine SWAT (Dunet et al., 2008; Hersey et al., 2008). Even so, the intent was never to examine large groups of innovations together. The goal was strictly to assess programs one-at-a-time, in the same way as CDC’s traditional public health “outbreak investigations.”

During their developmental stages, the SSA and SWAT projects interacted extensively. At the recommendation of the first author (Leviton), the SWAT project employed evaluability assessment in site visits. The director of the SWAT project (Dunet) was directly involved in first-year implementation of the Early Assessment initiative and is a coauthor of Chapter 2.

There are several similarities between the SWAT approach and the SSA Method, but some notable differences as well. Like SSA, the SWAT project employed an initial search for innovations, using available publications, the Internet, and nominations. SWAT used evaluability assessment and expert judgment (although in a different sequence from SSA). Based on SWAT experience, the SSA Method included feedback and technical assistance to each innovation that received an EA. In a later section, we describe some notable differences and discuss some conditions under which SWAT might be preferable to the SSA Method. Certainly there were differences in use of findings.
To plan the Early Assessment initiative, RWJF invited expert advisors to review the proposed approach and offer suggestions. The advisors had expertise in evaluation and in the substantive areas of children’s diet, physical activity, and related issues. In 2007 RWJF awarded $3 million to the Foundation of the Centers for Disease Control and Prevention for the Early Assessment initiative. Of this award, $2.4 million was devoted to screening and reviewing innovations. The remaining $600,000 was reserved to conduct a formal evaluation of the top priority identified in year one.

Three divisions of the National Center for Chronic Disease Prevention contributed staff to direct the effort: the Division of Nutrition, Physical Activity, and Obesity; the Division of Adolescent and School Health; and the Prevention Research Center Program of the Division of Adult and Community Health. ICF Macro coordinated the nomination, screening, and EA process and convened the expert panel meetings. The model has been presented twice at meetings of the American Evaluation Association, first as a planned initiative and later on the basis of the year one results. Because of feedback from prevention researchers and evaluators, and on the basis of 2 years’ experience, we believe the approach to be sound.

**What We Discovered About SSA From the Early Assessment Initiative**

The chapters that follow give a detailed description of the Early Assessment initiative on childhood obesity prevention: the methods and overall results, the conduct of evaluability assessments, and the effects of the process on our research agenda and the field of childhood obesity prevention. In this first chapter, however, we want to reflect on the initiative as our first full implementation of the SSA Method as well as the challenges that SSA posed from an evaluation standpoint.

**Selection of Priority Areas.** Our first challenge lay in selecting priority areas for assessment. CDC selected these priority areas once a year, in consultation with RWJF, expert advisors, potential users, and collaborating institutions and organizations such as NIH and the U.S. Department of Agriculture (USDA). Selection of topic areas reflects the first difference between the SSA Method and the SWAT approach; SWATs are intended to be one-at-a-time assessments. The initial pool was an artifact of the funding and timing, not an essential design feature of SWAT.

The Early Assessment topics reflected a variety of priorities, such as a need to assess lesser-known areas for potential investment and the particular information needs of collaborating organizations and potential users. Cronbach (1982) addressed this challenge: the need to select evaluation topics for their ability to reduce uncertainty about social programs and to maximize leverage for social change. Our selection of the priority areas in the Early Assessment initiative reflects some of the tradeoffs in doing so. We debated
whether to focus primarily on highly topical but prevalent types of innovations, such as school district local wellness policies and farmers’ market interventions. We argued that choosing prevalent policy or program types would tell us about opportunities and challenges in these emerging areas of prevention. Alternatively, we could focus on finding truly unusual innovations that might constitute a potential breakthrough for childhood obesity prevention. We concluded that a mix of both prevalent and unusual innovations would ensure the most useful information. The three innovations described in Chapter 4 give a sense of the range of innovations, and two of them are potential breakthroughs.

**The Nomination Process.** CDC staff called for nominations of innovations in each of the 2 years. The call went to public health professionals and researchers, supplemented by existing inventories of environmental interventions maintained by CDC, RWJF, and other foundations. The challenge encountered in the nomination process was to diversify the sources for nominations. Evaluation professionals have encountered this problem in the use of snowball sampling, where the goal is to diversify the “seeds” for the growing sample, thus covering many perspectives. Even though the call for nominations was fairly broad, it addressed health professionals primarily. With time, the network for possible nominations is becoming much broader, encompassing for example, city managers and planners, parks and recreation directors, and the Grocery Manufacturers Association. Future versions of the SSA Method might consider alternative strategies to identify innovations. Some possibilities would be to deliberately expand the professional networks that might provide nominations, or to issue a challenge akin to the X Prize (http://www.xprize.org/x-prizes/overview). Another strong possibility is to employ outlier strategies to identify populations and communities where innovations may have contributed to positive changes. This approach was employed usefully in school improvement research (Teddie & Stringfield, 2007). It would be eminently feasible given the many localities that now routinely collect information on children’s physical activity, dietary intake, and body mass index (Dietz, Story, & Leviton, 2009; National Center for Chronic Disease Prevention and Health Promotion, 2009).

In the area of nominations and inclusion criteria we find one of the strongest differences between the SSA Model and the SWAT initiative to which it bears such a strong resemblance. The SWAT project required that worksites had to be tracking some positive gains in obesity reduction and other health outcomes that were the focus of the SWAT project. By examining existing data collection systems at sites, the CDC team was also able to assess data access and the potential capacity for data collection essential for a rigorous evaluation. In the SSA Method, we assumed things were happening so quickly on so many fronts that we did not want to preclude potentially good ideas for lack of existing data. Moreover in the areas under study by the Early Assessment initiative, good data were unlikely to be available. Worksite data varied in quality, but at least the worksites were collecting participation data, body
mass index, or other health-related measures (Dunet et al., 2008; Hersey et al., 2008).

Initial Expert Panel Review. The panel members rated each innovation independently, and then through discussion they determined which innovations warranted evaluability assessment. Thus the expert panel served as gatekeepers for progress to the later steps. This is necessary because some innovators may push to get recognition and funding for their innovations, regardless of potential merit. We employed a number of criteria throughout the process of selecting innovations. These criteria were informed by the literature on evaluation and on prevention and were refined by a group of thought leaders early in the development of the SSA Method:

- **Potential impact.** The intervention appears to have potential for impact on the social or physical environment that enables healthy diet and physical activity, and ultimately on the behaviors themselves. Potential impact is assessed on the basis of the intervention's conceptual logic and other pertinent characteristics such as intensity and duration. Estimate of impact is based on “face value,” program documents, and brief expert input from funding organization staff and contractors, and other experts who know the intervention but are independent of it.

- **Innovativeness.** The intervention is new and different, or a significant variation on an existing promising intervention. Emphasis on innovativeness may be mitigated if the intervention represents a type or category of intervention that is prevalent in the field or of particular interest to CDC, RWJF, and collaborating organizations.

- **Reach to target population.** The likelihood or actual evidence that the intervention will achieve participation (and even retention and completion) by the target population. What proportion of the target population is likely to be affected by the intervention (Glasgow, Vogt, & Boles, 1999)?

- **Acceptability to stakeholders.** The potential or actual evidence that the intervention is acceptable and even attractive to pertinent collaborators, gatekeepers, and other necessary groups such as schools, businesses, government agencies, and grassroots groups.

- **Feasibility of implementation.** The likelihood that the intervention as designed can be implemented fully given the clarity of its goals, objectives, and strategies; complexity and leadership requirements; financial and other costs; and training and supervision requirements. If evidence exists regarding program implementation, then the extent to which the intervention “on paper” has been fully and faithfully implemented and the degree of difficulty in achieving implementation.

- **Feasibility of adoption.** The potential for similar sites or organizations to adopt the intervention.

- **Intervention sustainability.** The likelihood that the intervention can continue over time without special resources or extraordinary leadership.
• **Generalizability.** The degree to which the intervention has been or has potential to be adapted for other populations and settings.

• **Staff and organizational capacity.** Sponsoring organization and staff have the capacity to participate fully in an evaluability assessment, learn from it, and further develop the intervention.

A challenge at this stage concerned the assumptions that individual experts brought to the table. It was imperative that the experts be present and engaged in discussion of their assumptions as they applied these criteria. Not even the most accomplished experts have perfect knowledge of their substantive field, especially when it is evolving as rapidly as childhood obesity prevention. Together, however, they had enough knowledge to rule in and rule out what was likely to be effective, on the basis of the criteria. In fact, the experts were fascinated by these real-world practices.

In Chapter 5, we offer an example of how experts’ assumptions changed because of their in-person interaction; initially they dismissed the idea that farmers’ markets would have any power to affect diet. In the expert panel discussions, both at this stage and after the evaluability assessments, the panel had the opportunity for reality testing about new developments making it more plausible that farmers’ markets would assist the goal of obesity prevention.

In the use of expert judgment, we find a second difference between the SSA Method and SWAT. SSA requires a large number of nominations precisely so that expert judges can begin to identify the high-priority innovations that plausibly have the largest effects, reach the largest portion of the affected population, and decrease uncertainty about leverage for change (Cronbach, 1982). Unlike SSA, the SWAT project did not use an initial expert review to screen nominations, moving directly to staff analysis and requests for site visits. SWAT employed outside expert judgment only after site visit reports and data were available. Nevertheless, Hersey et al. (2008) concluded that the nine worksite programs were exemplary. The SWAT procedures were probably warranted because the field already knew a lot about worksite obesity prevention and SWAT could focus its efforts without the initial expert judgment.

**Evaluability Assessments.** The innovations identified by the expert panel were then invited to undergo evaluability assessment. Joseph Wholey originated evaluability assessments, a method that documents “the objectives, expectations, and information needs of program managers and policy makers; explores program reality; assesses the likelihood that program activities will reach measurable progress toward program objectives; and assesses the extent to which evaluation information is likely to be used by program management” (1979, p. xiii). Evaluability assessment depends on documenting the intervention’s design, developing a logic model of the intervention being assessed, consulting stakeholders who have an interest in the intervention, and documenting (often in briefest outline) implementation of the intervention. At each step in the process, the logic model is revised to reflect the reality of the intervention. A report is produced on each intervention,
which assesses the plausibility of the intervention producing the outcomes and also assesses the degree to which the organization, the available data collection, and other features of the innovation make it ready for evaluation.

Later chapters present much more operational detail on how evaluability assessments were conducted in the RWJF/CDC initiative. In this chapter, we merely want to contrast the usual practice of conducting evaluability assessments with how they were conducted in the context of the SSA. We see two major differences.

1. *Standalone versus multiple assessments.* In the literature on evaluability assessments, the presumption is that a single policy or program is assessed. In the SSA Method, 48 assessments of innovations sharing common themes were undertaken. Several published evaluability assessments do include multiple sites or states, but they do so in order to assess state or local implementation and stakeholder viewpoints about a single program, not to explore innovative approaches to solve a common problem (e.g., Mulkern, 2005).

2. *Decision rule for continuing to full evaluation.* Although the information from evaluability assessments generally has been found to be useful, often it does not result in further evaluation (Rog, 1985). Two likely reasons are that program logic models may not be consistent with resources and activities, and that the logic models themselves may not be plausible to achieve program outcomes. For the SSA Method, it is to be expected that only a modest number of interventions would merit subsequent evaluation; in fact, the approach takes advantage of this tendency. SSA depends on the evaluability assessments to winnow through interventions that may have such problems and identify those diamonds in the rough that merit further evaluation.

A challenge at this stage was to make clear to all stakeholders that evaluability assessment is not the same as evaluation. EAs are a preevaluation activity, designed to maximize the chances that any subsequent evaluation will result in useful information. Another challenge was that resources did not permit an extensive process of information gathering and revision of logic models. A single site visit was supplemented by many phone calls—yet in reality, the process had to be curtailed in comparison to other evaluability assessments we have read about and conducted ourselves.

**Expert Panel Review of Evaluability Assessments.** The expert panel reconvened to review the reports from the evaluability assessments and rate each intervention, for both its promise to prevent childhood obesity and its readiness for evaluation. Ratings focused on the degree of promise and readiness, not a yes-or-no decision about promise. This gave a more nuanced picture of each intervention and encourage further development, improvements, and constructive feedback to the developers and managers of the interventions.
Again the expert panels served as gatekeepers, but this time their role was to assure that the nine criteria listed in step 3 (potential effectiveness, innovativeness, reach, and so on) were applied in these ratings. Using these criteria, with the expert panel as gatekeepers, ensures that the project can resist any outside pressures to pronounce certain interventions as promising before a formal evaluation is available. In several federal settings, expert panels such as these have been effective in ensuring that only the evidence-based programs are endorsed by the government for dissemination (Centers for Disease Control and Prevention, 2008; Katzenmeyer & Haertel, 1986).

The expert panel deliberations were interesting at this phase because the evaluability assessments constituted a further reality test of the panel’s assumptions about the way that environment and policy would affect children’s diet and physical activity. Indeed, as Chapter 5 describes further, the panel had to be restrained from developing entire research programs around some of the innovations presented. The process was helpful not only for identifying the innovations worth testing but also for the range of ideas about constructive evaluation questions and the designs to study these questions.

In some cases the panel was able to review several examples of a similar innovation, as in the case of school wellness policies for low-income districts and farmers’ markets for inner-city communities. The panel proposed a cluster evaluation for the farmers’ markets, for which CDC has now authorized funding. Further, the discussions offered the possibility of developing new innovation “types.” By this we mean several exemplars gave rise to an abstract model that included several essential program components. It will be instructive in the future to see whether the idea of an innovation type assists the process of evaluation and later translation back to practice.

**Use of Information.** Three distinct uses of information have resulted from the SSA Method as applied to childhood obesity prevention. First and foremost, we have identified several outstanding innovations that are worthy of support for more rigorous evaluation. Other innovations were deemed promising and ready for evaluation. For this secondary group, the goal is to assist them and their research partners in applying for development and evaluation support, from foundations or the federal government.

The intended evaluation projects that follow from the SSA Method are fundamentally different from those intended by the SWAT initiative. SWAT assumes that the innovating organizations will carry out their own evaluations with technical assistance to increase their internal evaluation capacity. The SWAT focus on internal evaluation is consistent with CDC’s public health partnership with state and local health departments. The worksite programs studied by SWAT might decide to undertake internal evaluations, and the results might offer valuable information to the field. They might, but we note that researchers have been largely responsible for the body of evidence on worksite obesity prevention (Katz et al., 2005).

In contrast, the SSA Method assumes that the next step is rigorous evaluation of the highest-priority innovations, not self-evaluation of all promising
innovations. For the policy and environmental innovations studied through the SSA Method in the Early Assessment initiative, we assumed that research expertise, not internal evaluation, was absolutely essential to conducting follow-on evaluation. The challenges of measuring environmental changes to prevent children’s overweight were simply too new and formidable for non-researchers to address by themselves (Ohri Vachaspati & Leviton, in press; Story et al., 2009). Follow-on evaluation resources must go to organizations with substantial expertise in measurement, sampling, design, high-quality data collection, and analysis.

A second major use of the findings from the Early Assessment initiative and the SSA Method was the constructive feedback given to 48 intervention developers and technical assistance so they could improve their efforts. The evaluability assessments and the second expert review are excellent vehicles for program development. Yet the program development that results from evaluability assessment does not receive much attention in the literature (Leviton et al., in press; Smith, 1989). The intervention developers and managers genuinely appreciated CDC technical assistance and found it useful (see Chapter 5). To achieve this, CDC staff oversaw the feedback process and gave access to content experts to further advise the innovation developers on program improvement. This type of use also demonstrates differences between SWAT and the SSA Method. SWAT extends technical assistance for program development and building the innovators’ own evaluation capacity, while the SSA Method offers it for program development only.

The third use of information was for cross-site synthesis. The syntheses were developed from several sources: the available literature on these topics, the evaluability assessments (for emerging themes), and the discussions of the expert panel. A surprise in this process was that all the innovations were informative, whether they passed the test for promise, or not! Indeed, it was helpful to be able to discuss fatal flaws in the logic or activities, because this helped to articulate some of the critically important features of similar innovations. Indeed, one of the syntheses is being used as the basis of published CDC guidance on development of comprehensive school physical activity programs (Pitt Barnes et al., 2009a). Even though none of the innovations on this topic met the highest standard of promise and readiness for evaluation, the EAs furnished many insights for CDC guidance on how to improve program operation.

Cross-site use of findings was similar in some ways to the experience of the SWAT initiative. The data that worksites collected in the SWAT initiative were helpful to decrease uncertainty about effective practices in any worksite obesity prevention program. In the same way, synthesis of cross-site learning from the Early Assessment initiative informed decision makers about program elements that would plausibly increase the effectiveness of policy and environmental changes to prevent childhood obesity.

In another sense, however, the cross-site syntheses of the Early Assessment initiative differed radically from the SWAT synthesis for worksite obesity
prevention. The SWAT approach is intended as a middle ground between rigorous study of outcomes and the connoisseurship model of evaluation that uses expert judgment only (Dunet et al., 2008). For worksite programs this middle ground is defensible, but it is not defensible for most of the innovations identified by the SSA Method in the Early Assessment initiative. Good studies have already been conducted on the effectiveness of obesity prevention and treatment in worksites (Katz et al., 2005). This permits elaboration of components or practices that could optimize the effects of worksite programs and allows SWAT to take a defensible middle ground in recommending those practices for wider adoption. Unlike the worksite setting, tests of effectiveness have not been conducted for the policy and environmental changes examined by the SSA Method in the Early Assessment initiative. Therefore we need to be exceedingly cautious in fully testing their effectiveness before recommending adoption by others.

Conclusions

We conclude that the Systematic Screening and Assessment Method offers a cost-effective strategy to ensure more productive and useful evaluations. The Early Assessment initiative represents proof of concept for the step-by-step screening process, which clearly sharpened the focus for evaluation at every step. We acknowledge that no evaluability assessments were conducted for the nominations that were eliminated by the expert panel, so we cannot compare the potential yield from those that were screened out. As described in Chapters 2 and 3, however, it is implausible that they would have been deemed promising or ready for evaluation—barring any new information that the evaluability assessments might have revealed. The same expert panelists who dismissed them in the first review would have dismissed them again in the second review, using the very same criteria.

One might argue that truly powerful innovations might have been identified for evaluation without SSA. The New York City day care regulations and the Philadelphia Fresh Food Financing Initiative are described in Chapter 4. Both of these innovations were fairly well known to the field of childhood obesity prevention, or could easily have become so. Yet for every innovation that passed the screening and assessment stages, there were several that were well known and did not pass. In any case, the evaluability assessment process assisted in formulating evaluation plans, and indeed, programs of evaluation inquiry about the innovations.

Some of our expert panelists were dismayed at how few innovations seemed truly potent to achieve impact, even in the group of 20 that were finally identified. Does a school policy really contribute to a reversal of the childhood obesity epidemic? Does a new supermarket in a “food desert” do this for poor populations, who are most at risk of obesity? Readers will need to judge this for themselves. Certainly, two of the innovations described as case studies in Chapter 4 are very plausible and seem powerful enough to
achieve such impact. Also, it is clear that no single policy or environmental change will produce the impact that is needed; these are cumulative changes that only together may prevent childhood obesity (Leviton, 2008).

References


LAURA C. LEVITON is the coauthor of *Foundations of Program Evaluation* and is currently Special Advisor for Evaluation at the Robert Wood Johnson Foundation, where for the past 10 years she has overseen more than 80 evaluations at the national, state, and local levels.

MARJORIE A. GUTMAN is the Principal of Gutman Research Associates, with 25 years’ experience in program evaluation and development to address public health and behavioral health challenges.