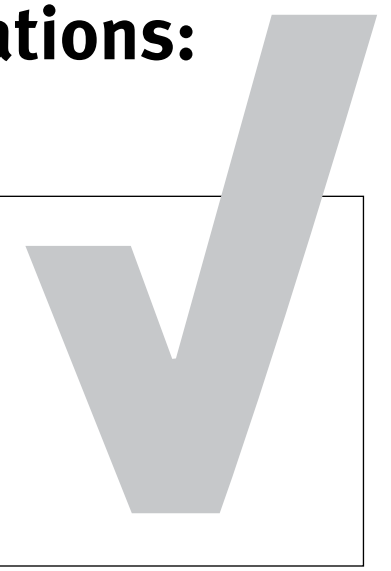
NATIONAL FORUM ON EARLY CHILDHOOD PROGRAM EVALUATION

Early Childhood Program Evaluations: A Decision-Maker's Guide



National Forum on Early Childhood Program Evaluation

A collaborative project involving Harvard University, Columbia University, Georgetown University, Johns Hopkins University, Northwestern University, University of Nebraska, and University of Wisconsin

Center on the Developing Child HARVARD UNIVERSITY

NATIONAL FORUM ON EARLY CHILDHOOD PROGRAM EVALUATION

Jack P. Shonkoff, M.D., Co-Chair
Julius B. Richmond FAMRI Professor of Child Health and Development; Director, Center on the Developing Child, Harvard University

Greg J. Duncan, Ph.D., Co-Chair
Edwina S. Tarry Professor of Human Development and Social Policy; Faculty Fellow, Institute for Policy Research, Northwestern University

Jeanne Brooks-Gunn, Ph.D.
Virginia and Leonard Marx Professor of Child Development and Education; Co-director, National Center for Children and Families, Columbia University

Bernard Guyer, M.D., M.P.H.
Zanvyl Kreiger Professor of Children's Health, Johns Hopkins Bloomberg School of Public Health

Katherine Magnuson, Ph.D.
Assistant Professor, School of Social Work, University of Wisconsin-Madison

Deborah Phillips, Ph.D.
Professor of Psychology and Associated Faculty, Public Policy Institute; Co-Director, Research Center on Children in the U.S., Georgetown University

Helen Raikes, Ph.D.
Professor, Family and Consumer Sciences, University of Nebraska-Lincoln

Hirokazu Yoshikawa, Ph.D.
Professor of Education, Harvard Graduate School of Education

The National Forum on Early Childhood Program Evaluation
This collaborative initiative fosters the analysis, synthesis, translation, and dissemination of findings from four decades of early childhood program evaluation studies to learn more about what interventions work best and for whom. Based at the Center on the Developing Child at Harvard University, the Forum involves researchers and data teams from Columbia University, Georgetown University, Harvard University, Johns Hopkins University, Northwestern University, the University of Nebraska, and the University of Wisconsin. Its work includes:

- building the nation's most comprehensive meta-analytic database on early childhood program evaluation, from the prenatal period to age 5 years;
- conducting rigorous analyses of the findings of well-designed studies of programs designed to improve outcomes for young children and/or provide effective support for their families;
- producing a variety of publications, including briefs for policymakers and civic leaders, peer-reviewed scientific papers, and web-based communications to assure both broad and targeted dissemination of high quality information.

For more information go to www.developingchild.harvard.edu/content/forum.html

PARTNERS

The FrameWorks Institute

The National Conference of State Legislatures

The National Governors Association Center for Best Practices

SPONSORS

The Buffett Early Childhood Fund

The McCormick Tribune Foundation

An Anonymous Donor

Suggested citation: National Forum on Early Childhood Program Evaluation (2007). *Early Childhood Program Evaluations: A Decision-Maker's Guide*. http://www.developingchild.harvard.edu

Early Childhood Program Evaluations

DESPITE INCREASING DEMANDS FOR EVIDENCE-BASED EARLY CHILDHOOD SERVICES, THE EVALUAtions of interventions such as Head Start or home-visiting programs frequently contribute more heat than light to the policy-making process. This dilemma is illustrated by the intense debate that often ensues among dueling experts who reach different conclusions from the same data about whether a program is effective or whether its impacts are large enough to warrant a significant investment of public and/or private funds.

Because the interpretation of program evaluation research is so often highly politicized, it is essential that policymakers and civic leaders have the independent knowledge needed to be able to evaluate the quality and relevance of the evidence provided in reports. This guide helps prepare decision-makers to be better consumers of evaluation information. It is organized around **five key questions** that address both the substance and the practical utility of rigorous evaluation research. The principles we discuss are relevant and applicable to the evaluation of programs for individuals of any age, but in our examples and discussion we focus specifically on early childhood.

1. **Is the evaluation design strong enough to produce trustworthy evidence?**
   Evaluations that randomly assign children to either receive program services or to a no-treatment comparison group provide the most compelling evidence of a program's likely effects. Other approaches can also yield strong evidence, provided they are done well.

2. **What program services were actually received by participating children and families and comparison groups?**
   Program designers often envision a model set of services, but children or families who are enrolled in "real" programs rarely have perfect attendance records and the quality of the services received rarely lives up to their designers' hopes. Thus, knowing the reality of program delivery "on the ground" is vital for interpreting evaluation results. At the same time, sometimes a comparison group is able to access services in their community that are similar to those provided as part of the intervention. If so, then differences between the services provided to the program and contrast groups may be smaller than would exist in a community where those services are not available.

3. **How much impact did the program have?**
   The difference between the outcomes for children and/or families who received services versus those of the comparison group are often expressed as "effect sizes." This section will explain what these mean and how to think about them.

4. **Do the program's benefits exceed its costs?**
   A key "bottom line" issue for any intervention is whether the benefits it generates exceed the full costs of running the program. This document will explain how costs and benefits are determined and what they mean for a program that is being considered for implementation.

5. **How similar are the programs, children, and families in the study to those in your constituency or community?**
   Program evaluations have been conducted in virtually every state and with children of diverse ethnicities and socioeconomic backgrounds. Knowing how the characteristics and experiences of comparison-group children compare to the characteristics and experiences of children in your own constituency or community is important for determining the relevance of any evaluation findings.

For guidelines and explanations that can help leaders use these key questions to determine the relevance of program evaluations for policy decisions, please continue.

1. Is the evaluation design strong enough to produce trustworthy evidence?

CHECK LIST #1

- ☑ **Value experimental designs (RCT) over non-experimental studies. Random assignment is the best way to ensure that differences in outcomes are the result of program effects rather than from something different about the children or families who received the services versus those who did not.**
- ☑ **Not all evaluations that use an RCT design are successful. Sometimes random assignment doesn't work. For example, problems arise when too many program or control group children cannot be located for a reliable "post-treatment" measurement of outcomes.**
- ☑ **Useful evaluation lessons can be drawn from rigorous non-random-assignment evaluation studies such as those employing regression discontinuity designs (RDD).**

EVALUATION STUDIES TAKE MANY FORMS, BUT the most useful studies answer the question that policymakers and parents most want to know the answer to—does a program or intervention "work?" What, for example, would have happened to children in Head Start if they had not been enrolled in the program? The presumption is that they would not have learned as much, but how much less? How can we be certain there really is a difference? How confidently can it be ascribed to Head Start?

It would be easy to determine how well a program works if we could somehow compare its effects on a group of children to what would have happened if those same children had not received the services. Since that's clearly impossible, all evaluations have to find some kind of comparison group to assess program impacts. And how close program and comparison-group children are to being the same *before the services are provided* is a major determinant of how valid the study findings will be. This is not easy, since children who attend programs are often different from those who do not. They may be healthier or sicker. Their parents may be better off or poorer. Parents of program children are often more motivated to seek out services than parents whose children do not attend. If comparison-group children differ in these or other ways from children who are enrolled in a program before the services are provided, then later differences are likely to reflect, in part, these initial differences and thus convey a false picture—either more or less favorable—of the program's impacts.

The strongest evaluation designs compare children and parents who receive program services with a "virtually identical" comparison group of children and parents who do not receive those services. The ideal method for assessing program effects is an experimental study referred to as a randomized controlled trial (RCT). In an RCT, children who are eligible to participate in a program are entered into a "lottery" where they either win the chance to receive services or are assigned to a comparison (control) group. Parents or program administrators have no say in who is selected in this lottery. When done correctly, this process creates two groups of children who would be similar if not for the intervention. Any post-program differences in achievement, behavior, or other outcomes of interest between the two groups can thus be attributed to the program with a high degree of confidence.

It is possible for an RCT to be flawed and result in a comparison group that is not comparable to program participants. Examples of how this may occur include problems implementing the lottery process, too few children in the program and comparison groups, and too many children or families dropping out of the study after random assignment has occurred. For this reason, even an RCT study should demonstrate that the comparison group used was similar to the treatment group before the study began.

Although random assignment of children or parents to program and comparison groups is the "gold standard" for program evaluation, sometimes this is not possible. In some circumstances, a randomized controlled trial is neither practical nor ethical. For example, if access to services is a legal entitlement, denying program services to some children would be a violation of the law. In such cases, alternative ways of constructing "no treatment" groups are needed and it is essential that the children and families in the comparison group be as similar to the program group as possible.

The strengths of other evaluation methods are highly variable, with an approach called Regression Discontinuity Design (RDD) considered by experts to be the strongest alternative to random assignment. In this case, assignment to either the control or the intervention group is defined by a cut-off point along some measurable continuum (such as age). For example, some pre-K evaluations have taken advantage of strict birthday cut-off dates for program eligibility. Specifically, in some states, children who are 4 years old as of September 1 are eligible for enrollment in pre-K, while those who turn 4 after September 1 must wait a year to attend. In this case, the key comparison in an RDD is between children with birthdays that just make

or miss the cutoff. These children presumably differ only in the fact that the older children attend pre-K in the given year while the younger ones do not. Comparing kindergarten entry achievement scores for children who have completed a year in pre-K with the scores measured at the same time for children who just missed the birthday cutoff can be a strong assessment of program impacts.

Evaluations that select comparison groups in other ways should probably be assumed guilty of bias until proven otherwise. Countless studies have shown how difficult it is to create comparison groups that are similar, absent an RCT design or close approximation. Especially important indicators of treatment-comparison group comparability are assessments of test scores, behaviors and other outcomes of interest for both groups of children taken *just prior to the point of program entry*. Demonstrating that the program group and comparison group children or parents were initially similar on characteristics that the program was intending to affect is vital for trusting that differences emerging after the beginning of the program can be attributed to the program itself. Evaluations that do not compare and discuss pre-service characteristics of program and comparison-group children should be viewed with skepticism.

2. What program services were actually received by participating children and families and comparison groups?

AT THE HEART OF AN EVALUATION STUDY IS THE comparison of two groups of children—those who are enrolled in the program and a similar comparison group of children who are not. Sometimes, however, it is surprising to find out that the actual experiences of these two groups of children are very similar. This can occur either because the children enrolled in the program do not receive the services as intended or because many of the children in the comparison group seek out and receive similar services that are already available in the community. Good questions to ask when reviewing program outcomes include:

Were there problems with program delivery? No one wants to implement a poor quality program or a program that is so unappealing, inconvenient, or inaccessible to the target families that they do not make use of it. Although years of experience have shown what general program characteristics make services attractive to families, it is still essential to know the answers to the following questions. Was the intervention in the program evaluation actually delivered? What were the qualifications of those who delivered the service? Was it implemented in the way it was intended? What volume or "dosage" of program services did participating children and families actually receive?

Did participating families receive the services that were planned? The best intentions of program developers are often not reflected in the experiences of families with infants and young children "on the ground." This problem is most commonly caused by one of two reasons—either the program was not implemented as intended or families did not participate as expected. In fact, in some cases implementation or take-up problems can be so severe that the most reasonable conclusion would be that the intervention was not really tested. On the other hand, an intervention that is difficult to implement or that is not successful in engaging the children and families it seeks to serve is not likely to be effective, despite its theoretical appeal.

Implementation refers to whether all of the components that were planned and/or described were actually put into place at all of the sites. Sometimes, especially in evaluations of services that are implemented in multiple locations, the program is well implemented in some places but not others. Poor implementation can arise for many reasons—a building is not completed on time, a director quits unexpectedly, or the enrollment of families takes much longer than anticipated. Not surprisingly, studies that measure variation in implementation often show that the most fully implemented sites have the strongest impacts. But at the same time, it is not

CHECK LIST #2

☑ **It is important to know whether the program was experienced as intended. What type and volume of program services was a typical participating child or family supposed to receive? Was this model implemented each year and in every site? To what extent did children or families fail to "take up" services offered to them or show up as often as planned?**

CONTINUES P.4

CHECK LIST #2, CONT.

- ☑ **Examine multiple characteristics of the program that was delivered (e.g., intensity, duration, skills and credentials of the service providers, and participation rates). If important services were not provided as intended, the program is not likely to be as effective as hoped. Remember that the evaluation assesses the program as delivered, not as designed.**
- ☑ **Look carefully for lessons about program improvement. Do the reports include a section on implications for other programs? Is there information about implementation or program design that can be translated into practical guidelines for further program refinement?**
- ☑ **Find out as much as you can about the experiences of the evaluation's control group. Often the "does a program work" question should be rephrased as "does the program work in comparison to the experience of those who didn't receive the same services?"**

realistic to expect that a program implemented in your own community would be lucky enough to avoid all of the problems encountered by the poor-implementation sites. Thus, impacts that are averaged across all locations are probably a better guide to what to expect than impacts attained by only the best sites.

In some circumstances, a program could be implemented exactly as intended, but the participation rates could still be low. This may be a sign that the program is not attractive or accessible to potential participants. An example would be a parent outreach service connected to an early education program that offers home visits in the afternoon, when most working parents cannot participate because of difficulty in adjusting their work schedules. Another example is a program whose services are not a good fit with the cultural norms of the particular population being targeted (e.g., home-based services for a cultural group that may have strong values concerning privacy of the home). In such circumstances, the failure to "take up" the home visitation piece does not necessarily mean that this program component could not be beneficial to families. It may simply mean that the program delivery needs to be designed to fit with the daily routines, values, and preferences of the specific group being served. Issues related to language for families who do not speak English are also very important in this context.

Participation (sometimes called program "take up") refers to the services that children and families actually receive. The measurement of participation has two dimensions—how many of the parents or children participated and, for those who were involved, how much service did they receive. The first dimension is measured by take-up fractions (i.e., the number of families who were engaged divided by the total number of possible participants). Every evaluation should include information about how many families never enrolled or dropped out of the program. The second dimension includes measures of program "dosage," such as numbers of visits, hours of service received, and weeks, months, or years of program participation. In addition to including information about these two dimensions of participation, studies are even more useful if they include data from systematically conducted interviews or focus groups that describe what parents and children actually experienced.

Do implementation or take-up problems point to more promising practices? No intervention is perfect. Changing behavior and shifting the course of children's development is challenging, and even promising programs can be strengthened. Increasingly, contemporary intervention programs are turning to "continuous improvement" or "action research" frameworks, guided by a knowledge base that assists service providers and policymakers in improving program effectiveness. To this end, a supplementary set of inquiries beyond the simple "did it work?" question can be very useful. This approach is particularly important for evaluations of programs that must be provided, such as public schools. Don't hesitate to contact evaluators directly and ask, "What do the data tell us about how the program can be improved?"

What type of services did the comparison children receive? Another important question about program receipt is the extent to which children in the comparison group were able to access similar services. Good evaluations detail exactly what services or programs were received by children and families in the comparison group. In some studies, children in the comparison group could not have participated in a similar program because it was not available to them. In other studies, however, children and families in the comparison group were able to seek out and access similar programs. Over time and across communities, there is considerable variation in the extent to which alternative programs and services are available to comparison group children. Sometimes the contrast of the program and comparison group service experiences is quite small, and thus the program may appear to be less effective.

For example, a couple of decades ago, most children who were not assigned to participate in an early education program simply stayed home and were cared for by their mothers. The world has changed dramatically since that time, and most young children today—even infants—do not spend all of their time at home. In fact, child care and family support services are pervasive throughout the nation, although there is striking variability in their quality, accessibility, and affordability. These changes have important implications for drawing lessons from program evaluations that were conducted in the past or for guidance

in communities with different service configurations. Stated simply, program evaluations can only show how a specific program works *in comparison to the existing landscape* of other community-based services available to the control group, including child care, health care, and other early intervention programs, among others.

3. How much impact did the program have?

THE MEASUREMENT OF PROGRAM IMPACTS—THE differences between the treatment and comparison groups on a range of outcomes of interest—is a central feature of the evaluation process. Impacts can be expressed in a variety of ways, such as percentage differences or differences in the proportion of program and control-group children who fall into a specific category, such as assignment to special education classes.

Effect sizes. Increasingly, program evaluators express impacts as "effect sizes," which are a statistical means for comparing outcomes that may otherwise be difficult to compare. For example, the scales of the SAT test and the IQ test are completely different, so it's difficult to compare one program that raises SAT test scores by 20 points, and another that raises IQ scores by 5 points. "Effect sizes" provide the solution. By subtracting the outcomes of the control group from the outcomes of the treatment group, we get an effect (e.g., raising SAT scores by 20 points). By dividing that effect by the study's "standard deviation" (which indicates how widely dispersed the results are from the mean), we get an effect size—a fraction that indicates how large the effects are in comparison to the scale of results.

The SAT test, for example, is scaled with a standard deviation of 100, so a program that boosted SAT scores by 10 points would have an effect size of 0.1, or one-tenth of a standard deviation—which is considered very small. IQ tests are typically scaled with a standard deviation of 15, so a program that boosted IQ scores by 10 points would have an effect size of 0.66, or two-thirds of a standard deviation—which is much larger. Generally speaking, the larger the effect size, the better. Conventional guidelines consider effect sizes of at least 0.8 as "large"; 0.3 to 0.8 as "moderate"; and less than 0.3 as "small." Nevertheless, since inexpensive programs can hardly be expected to perform miracles, we will soon see that an even better measure of a program's worth is the value of its effects relative to its cost.

The best studies translate effect sizes into practical information. For example, effects on a standardized measure of achievement might be translated into how much of a fraction of a school year the program group exceeds the control group. Effect sizes on grade retention can be translated into percentages of children held back a grade.

Statistical significance. Impacts are usually accompanied by a statement regarding their statistical significance. This indicates how much confidence we have that the measured impact is real and not just something that appeared by chance. Impacts that are statistically significant at the 5 percent level—a common standard—mean that if we could somehow conduct 100 evaluation trials, we would expect to confirm those impacts in 95 of them. That is a good bet that the impacts are real.

As the number of children or families in the treatment and control groups increases, smaller effect sizes become more statistically significant, simply because a larger sample means a lower probability of a chance finding. Typically, evaluations involving fewer than 100 children require very large effect sizes to be judged statistically significant, while evaluations based on several thousand children are much more likely to calculate small effects as statistically significant. All other things being equal, bigger studies are better. Even in large studies, however, small effect sizes imply that the program is not likely to change outcomes very much, so policymakers should consider carefully the cost required to achieve small benefits.

Pattern of results. Good program evaluations present or summarize results for all of the outcomes they measure, not just the ones that

CHECK LIST #3

- ☑ **Program impacts are often expressed as "effect sizes," which provide a uniform way to compare influences on different kinds of outcomes and across evaluation studies.**
- ☑ **Statistical significance provides a valuable judgment of how likely an estimated impact is real and truly different from zero.**
- ☑ **Distrust evaluations that report only measures with statistically significant impacts. Every rigorous evaluation is likely to generate a mix of significant and non-significant findings. The overall pattern of effects is most important.**
- ☑ **It is important to understand whether the offer of services (ITT) or the receipt of services (TOT) is being evaluated and whether there are some groups of participants that may benefit from the program more than others.**

produced statistically significant impacts. It is unrealistic to expect that even highly effective programs will produce statistically significant impacts on all of the measured outcomes. And a quirk of the standard practice of applying tests of statistical significance is that even if a program were completely ineffective, for every 100 outcomes tested, you would still expect five of them to emerge as statistically significant simply by chance! "Cherry picking" small numbers of statistically significant results can be very deceptive. Generally speaking, it is the overall pattern of results that matters the most.

Relevance. In reading evaluation reports, it is always useful to ask how much measured program outcomes are relevant to the desired outcomes for your constituents or community. Of the outcomes measured, which do you care most about? Was the program more effective for those outcomes than for others? If you care about boosting children's school achievement, are most of the achievement impacts in the evaluation statistically significant? If one purpose of the intervention is to save money for school districts, did the program produce statistically significant impacts on school-related measures that have financial effects, such as grade failure and enrollment in special education? Use these kinds of questions to guide your assessment of the program's relevance to your goals for the health and development of children.

"Intent to treat" impacts. In evaluations of interventions in which substantial numbers of children or families fail to take up any of the offered services, there is an important technical detail that must be addressed. Should program effects be considered for only those who receive the services or for all families who are offered the program, regardless of whether they participate? This question is illustrated in programs designed to promote residential mobility among public housing residents, in which between one-quarter and one-half of the families that are offered financial assistance and mobility counseling fail to take advantage of the offer. Thus, an evaluation of child and family outcomes influenced by the mobility program faces a choice—should outcome differences between the program and comparison group be calculated across all families offered the chance to move, or only for those families that actually moved in conjunction with the program?

Effects assessed across all children or families offered program services, regardless of whether they actually used them, are called *"intent to treat"* (or ITT) impacts. They answer the vital policy question about the effects of the program on all families that are offered services. Suppose, however, that services are highly effective for those who participate, but only a small fraction of the targeted children or families actually use them. The intent to treat impact estimates will show that the overall impact on targeted families is small and will point to implementation or program take-up as a key problem in program design.

"Treatment on the treated" impacts. Under certain circumstances, it is also possible to isolate program impacts on the subset of families that actually use the services and compare them to families that did not use similar services. These are sometimes called "treatment on the treated" (or TOT) impacts, and amount to scaling up intent-to-treat estimates in proportion to program take-up. Treatment-on-the-treated estimates address important policy questions about program impacts on the children or families who actually use the services. If program take up is not a concern and you want to concentrate on how a program affects children or families who participate in it, then TOT estimates are most relevant. Finally, when comparing across studies it is important to compare like with like—ITT with ITT impacts or TOT with TOT impacts.

Subgroup effects. Some programs are more effective for some subsets of children or families over others. For example, an intensive program designed to help low birth-weight babies was found to be considerably more effective for children whose birth weights were close to normal than for children with very low birth weights, some of whom exhibited serious neurological problems. It is common for evaluations to report effects on various subgroups of participants. These findings may be useful for forecasting potential program impacts on the children, particularly if the measured impacts are largest among subgroups with characteristics similar to likely participants in your own community.

4. Do the program's benefits exceed its costs?

A CLEAR AND OBJECTIVE ANALYSIS OF THE COSTS and benefits of specific programs has become an increasingly important consideration for many policymakers as they face decisions about investments in young children. Stated simply, do the total benefits generated by the intervention exceed its costs? Just as business executives want to know how an investment would affect their company's bottom line, it is useful to ask not only whether government program expenditures have their intended effects, but also whether investing in early childhood programs generates financial "profits" for the children themselves, for taxpayers, and for society as a whole.

Costs and benefits. Although the details can be tricky, the basic idea behind a cost-benefit accounting is fairly straightforward. On the cost side, we want to know the value of all the time and money expenditures incurred by the program on behalf of the participants. Salaries typically dominate program costs, and services that provide one-on-one or small group sessions administered by a professional staff are more expensive than those that are delivered within large groups or by less well-trained personnel.

On the benefit side, we want to know the value of the program for taxpayers and for the participants themselves. For example, if the program reduces grade repetition or assignment to special education classes, the value of savings to taxpayers can easily total thousands of dollars per child. Similarly, substantial long-term impacts on educational achievement can be translated into both higher labor market earnings for the participants and increasing tax payments and general economic productivity for society as a whole.

By the same token, behavior-oriented interventions can profit from reductions in criminal behavior, as crime generates large costs for adjudication and incarceration as well as for crime victims. Health-related effects can also be important, as reductions in obesity and smoking rates can be linked to savings in health expenditures.

Return on investment. Economists tell us that the most profitable investments are not necessarily generated by programs that produce the biggest "effect sizes" but rather by those that lead to the largest benefits relative to costs. According to such calculations, less intensive programs cost less and therefore do not need to generate the same volume of benefits as more intensive programs in order to produce a social profit. On balance, it is impossible to generalize about the relative profitability of programs based on costs, benefits, or effect sizes taken alone.

Some program evaluations include a detailed cost-benefit accounting in their analyses. If done well (you may wish to consult with someone with expertise in cost-benefit assessment to judge the quality of a specific study's accounting), the obvious question is whether a program's benefits exceeded its costs. Properly done, costs and benefits are calculated on a "present value" basis to reflect the fact that tying up public money in the short run to produce longer-run benefits entails a genuine "opportunity cost" to society. Benefits in excess of costs indicate that a program is a worthy expenditure of public funding *from a financial perspective*. An equivalent calculation can be made to determine whether the program produced a favorable "rate of return" on the investment.

If a cost-benefit accounting is not provided, it is vital to consider an order-of-magnitude estimate of the likely costs of recommended policy changes. Are costs per child or family likely to amount to $100, $1000, or $10,000? If services are required for several years to produce their effects, then per-year costs must be multiplied accordingly. If a program provides one-on-one or small-group services, it is likely to be more expensive to deliver. The level of professional training that is required of the service providers will also have a significant impact on cost.

Other measures of value. Notwithstanding the importance of cost-benefit analyses, it is important to remember that some investments may be justified because of their intrinsic value, independent of their financial return. For example, if the policy goal is reducing crime or high school drop-out rates, policymakers and the public may simply be interested in achieving the goal, regardless of what any cost-benefit analysis might show. In other cases, investments in children who are highly vulnerable (such

CHECKLIST #4

- ☑ **Cost-benefit accounting provides an important indication of a program's value to the public. Programs that generate the largest surplus of benefits relative to costs (or the most positive rates of return) generate greater value for public and private investments.**
- ☑ **Costly programs with large effects are not necessarily better financial investments than inexpensive programs with smaller impacts. Conversely, inexpensive programs with little to no effects may be a waste of money when a more expensive program will generate larger effects. The key calculation, from an economic perspective, is the size of the benefits generated by the program relative to program costs.**
- ☑ **The greatest economic returns from investments in early childhood typically are long-term. Thus, it's important to look at costs and benefits longitudinally, and to consider social and economic benefits as a legacy for tomorrow built from sound decision-making today.**
- ☑ **Financial payback is not the only measure of a program's worth. Some public investments are made as a matter of social responsibility. In such cases, costs are viewed in terms of efficiency.**

as those who have been abused or seriously neglected) may be justified solely because of their humanitarian significance, independent of the long-term financial gains that may be realized from better health and developmental outcomes. In such cases, cost-effectiveness studies that tell us how to deliver services in the most efficient manner will be more useful than cost-benefit studies that assess their economic payback.

5. How similar are the programs, children and families in the study to those in your constituency or community?

CHECKLIST #5

- ☑ **Look for specific information about the program. Can you form a clear picture of the services offered and how they differ from what is currently available in your community? Does this match the way in which your own community would provide these services?**
- ☑ **Consider the constituency or population for whom you might provide a particular program. How well does the study sample approximate this population?**
- ☑ **If it does not overlap substantially with your own constituency, examine the study carefully to determine which aspects of the program, if any, might need to be adapted to fit your community's needs.**

LET'S SAY YOU ARE A BUSINESSMAN IN CLEVELAND, Ohio, wanting to know whether a successful program that was evaluated in Hawaii in 1990 would work as well for your community today. Your first question should be: What kinds of children or families would receive services if the program were implemented in Cleveland? Would it be targeted toward children from low-income families? Children of immigrant parents from particular groups? Children with disabilities? The more precisely you can characterize the intended recipients of the services and how the services differ from what is currently available in your community, the easier it will be to determine the relevance of the findings of a given evaluation study. The more closely the use of services by children in the study's comparison group matches those of children in your own community, the more relevant the study findings will be.

Next, compare the characteristics of the Cleveland target population with those of the children or families in the Hawaiian program evaluation. On how many dimensions (e.g., poverty status, inner-city location, languages used at home and in other settings, parent education levels, cultural beliefs, and parenting practices) are they similar? On what dimensions are they different? If the study was conducted years ago, the circumstances for children with identical characteristics today may differ in important ways. Both the nature and the extent of the diversity of your target group of families in Cleveland is important to consider.

Finally, carefully examine the description of the program. Is it tailored to the particular group in that study in a specific way (e.g., in its language, materials, cultural values, staffing, or approach)? Is it difficult to imagine how the program might be "refitted" for your community? Does it require specially trained and qualified staff who may be too scarce or costly in your community? Some programs might be easier to adapt than others. For example, an intervention that provides a high-quality preschool experience might be easier to reproduce than a child literacy intervention that is based on folk tales among a particular cultural group.

There is much to be learned from rigorous evaluations of early childhood interventions. Applying those lessons to one's own community, however, requires a careful eye toward understanding which aspects of the interventions are most likely to be replicable given your current situation, target population, and goals.

Putting it All Together

TO ASSESS THE OVERALL VALUE OF A PROGRAM for your constituents or community, there are several overarching guidelines that can help determine how to use the evidence of previous evaluations.

Consider whether the evaluation is strong enough to provide trustworthy evidence. This is the first and probably most important question to be answered. If the study fails to meet scientific standards of strong evidence, it is difficult to assess its program or policy implications.

Consider how closely the program that was evaluated matches your goals. Of all the information provided by the evaluation, which elements are most useful and relevant for your constituency and goals? For example, if reducing the achievement gap is your constituency's primary objective, you might heavily weight Question 3 (How much impact did the program have?), with particular emphasis on whether the outcomes differ for different income or racial/ethnic groups in your community.

Consider how successful programs can be modified to best meet the needs of your particular community or constituency. Although fidelity to the specific methods used in an effective program is critical to achieve similar outcomes in another setting, it is important to note that some programs may require adjustments that make sense for different circumstances. This could be warranted because the particular service systems, organizations, or cultural groups in your community are different from those in the original study. Perhaps the delivery system (e.g., child care providers, preschool, or health care system) should be changed. Perhaps the setting in which services are delivered should be modified. In many circumstances, credible information about costs as well as about how the program was implemented will provide important guidance for determining whether a program is feasible for your constituents. If local factors require changes in a program whose effectiveness has been documented previously, it is essential that the modified program be evaluated to assure that it is achieving the desired results.

Consider getting expert assistance to answer your continuing questions. Mastering the complexities and nuances of evaluation research is beyond the limits (or interest) of most policymakers, civic leaders, and the general public. Thus, developing trustworthy consultants in the areas of programs that most interest you may be well worth your time. Researchers are often happy to respond to questions about their own study findings. Getting to know local experts in your community can also be quite helpful for digesting the massive amount of information provided in the full body of program evaluation studies. Trusting relationships with such consultants could be particularly useful in "translating" study findings for local application.

ALSO FROM THE FORUM

A Science-Based Framework for Early Childhood Policy: Using Evidence to Improve Outcomes in Learning, Behavior, and Health for Vulnerable Children (2007)
http://www.developingchild.harvard.edu/content/publications.html

NATIONAL FORUM ON EARLY CHILDHOOD PROGRAM EVALUATION

Center on the Developing Child HARVARD UNIVERSITY

50 Church Street, 4th Floor, Cambridge, MA 02138
617.496.0578
www.developingchild.harvard.edu